

Khakimova M.F.

TEXTBOOK
INTRODUCTORY ECONOMETRICS

*Textbook recommended by the Ministry of education and science
of the Republic of Tajikistan No 17/41 from 24.12.2019 for Higher
education institutions with economic specializations*

2020

UDC 330.43(075.8)
LBC (ББК) 65в6я73
Kh – 16

Reviewers:

Jurabaev G.J., Doctor of Sciences, Full professor of the public and local management department of the Institute of tourism, entrepreneurship and services of the Ministry of industry and new technologies of the Republic of Tajikistan;

Tabarov O.S., Doctor of Science, Associate professor, head of economic cybernetics department of the Tajik national university;

Najmiddinov A.A., Doctor of Science, Associate professor, head of mathematical and informational modelling in economics department of the Tajik state university of finance and economics.

Khakimova M.F.

Kh -16 Introductory Econometrics. Textbook – Dushanbe, 2020. – 432 pages.
Введение в эконометрику. Учебник – Душанбе, 2020. – 432 стр.

The textbook contains a systematic representation of the fundamentals of econometrics and based on lectures that author gave over the several years at the Tajik national university and the Tajik state university of finance and economics. Textbook contains detailed information regarding linear (least squares method, hypothesis testing, heteroskedasticity, autocorrelation, model specification) and nonlinear regression (time series, dummy dependent variable techniques, simultaneous equations, two stage-least square method) models.

The textbook is intended for students, masters and lecturers of economic specialities.

CONTENTS

Introduction	7
CHAPTER 1. AN OVERVIEW OF REGRESSION ANALYSIS	9
1.1 What is econometrics?	9
1.2 What is regression analysis?	14
1.3 The nature and sources of data	17
1.4 Features of interconnection.....	28
1.5 Causality and the notion of ceteris paribus in econometric analysis.....	30
1.6 Econometric packages for modelling economic and financial data.....	34
SUMMARY	36
CHAPTER 2. THE SIMPLE REGRESSION MODEL	37
2.1. Definition of the simple regression model	37
2.2. Deriving ordinary least square estimates	46
2.3 Some further terminology	54
2.4 Precision and standard errors.....	59
2.5 Regression through the origin.....	65
SUMMARY	66
APPENDIX 2:	67
2A. Minimizing the sum of squared residuals	67
2B. Mathematical derivations of CLRM results.....	69
CHAPTER 3. MULTIPLE REGRESSION ANALYSIS: ESTIMATION	75
3.1 Motivation for multiple regression	76
3.2 Mechanics and interpretation of ordinary least squares	82
3.3 Comparison of simple and multiple regression estimates	97
SUMMARY	99
APPENDIX 3:	99

3A.1 Mathematical derivations of CLRM results.....	99
3A.2 Derivation of the first order conditions.....	102
3B. A brief introduction to factor models and principal components analysis.....	103
CHAPTER 4. LEARNING TO USE REGRESSION ANALYSIS	109
4.1 Steps in applied regression analysis.....	109
4.2 Using regression analysis to pick restaurant locations.....	117
4.3 Dummy variables	123
SUMMARY	128
CHAPTER 5. THE CLASSICAL MODEL	129
5.1 The classical assumptions	129
5.2 The sampling distribution of $\hat{\beta}$	138
5.3 The Gauss–Markov theorem and the properties of OLS estimators	145
5.4 Standard econometric notation.....	147
SUMMARY	148
CHAPTER 6. AN INTRODUCTION TO STATISTICAL INFERENCE.....	149
6.1 What is hypothesis testing?	150
6.2 The t-test.....	156
6.3 Hypothesis testing in Eviews	169
6.4 Examples of <i>t</i> -tests	176
6.4 Confidence interval	185
6.5 The <i>F</i> -test	189
6.5 Sample Eviews output for multiple hypothesis tests.....	198
SUMMARY	205
CHAPTER 7. SPECIFICATION: CHOOSING THE INDEPENDENT VARIABLES	206
7.1 Omitted variables	207
7.2 Irrelevant variables.....	215

7.3 An illustration of the misuse of specification criteria.....	218
7.4.Specification searches.....	220
SUMMARY	229
APPENDIX: 7	230
7A. Additional specification criteria.....	230
CHAPTER 8. SPECIFICATION: CHOOSING A FUNCTIONAL FORM	235
8.1 The use and interpretation of the constant term.....	236
8.2 Alternative functional forms.....	238
8.3 Lagged independent variables	249
8.4 Slope dummy variables.....	251
8.5 Exercises with incorrect functional forms	255
SUMMARY	258
CHAPTER 9. MULTICOLLINEARITY	259
9.1 Perfect versus imperfect multicollinearity	260
9.2 The consequences of multicollinearity	264
9.3 The detection of multicollinearity	271
9.4 Remedies for multicollinearity	276
9.5 An example of why multicollinearity often is best left unadjusted.....	280
SUMMARY	282
CHAPTER 10. SERIAL CORRELATION	284
10.1 Time series.....	284
10.2 Pure versus impure serial correlation	286
10.3 The consequences of serial correlation.....	293
10.4 The detection of serial correlation	295
SUMMARY	310
CHAPTER 11. HETEROSCEDASTICITY	312
11.1 Pure versus impure heteroskedasticity	313
11.2 The consequences of heteroskedasticity	318

11.3 Testing for heteroskedasticity	321
11.4 Remedies for heteroskedasticity	328
11.5 A more complete example	333
SUMMARY	339
CHAPTER 12. TIME SERIES MODELS	341
12.1 Distributed lag models	342
12.2 Dynamic models.....	344
12.3 Serial correlation and dynamic models	349
12.4 Granger causality	353
12.5 Spurious correlation and nonstationarity	355
SUMMARY	366
CHAPTER 13. DUMMY DEPENDENT VARIABLE	
TECHNIQUE.....	368
13.1 The linear probability model.....	368
13.2 The binomial logit model	375
13.3 Other dummy dependent variable techniques.....	383
SUMMARY	386
CHAPTER 14. SIMULTANEOUS EQUATIONS	387
14.1 Structural and reduced-form equations	388
14.2 The bias of ordinary least squares.....	395
14.3 Two-stage least squares (2SLS)	399
14.4 The identification problem.....	402
SUMMARY	408
CHAPTER 15. FORECASTING	410
15.1 What is forecasting?	411
15.2 More complex forecasting problems.....	416
15.3 ARIMA models.....	425
SUMMARY	429
REFERENCE	431

INTRODUCTION

The development of the economy, complication of economic processes and the increasing requirements for management decisions in the field of macro- and microeconomics required a more precise and objective analysis of actual processes based on the use of modern mathematical and statistical methods. On the other hand, the problem of violation of the classical statistical methods' backgrounds in solving real economic exercises led to necessity of develop and improvement the classical methods of mathematical statistics and clarifying the statements of corresponding exercises. As a result of these processes, formation of a new field of knowledge called **Econometrics** connected with development and application of methods for the quantitative assessment of economic phenomena and processes as well as their interrelationships has been realized.

The main method of econometric research is economic and mathematical modeling. A properly constructed model should give understanding of quantity characteristics of the research objects' changes depending on changes in the external environment. For example, how will affect the gross domestic product by increase or decrease of level of investment, what additional resources will be needed for the planned increase in output, etc.

The practical significance of econometrics implies application of its methods which makes it possible to identify actual relationships between phenomena, to give a well-grounded forecast of the phenomenon development under given conditions, to check and numerically evaluate the economic consequences of policy decisions made.

The construction of econometric models has to be carried out under conditions where backgrounds of classical statistical methods are violated and taking into account the presence of such phenomena as:

- multicollinearity of explanatory variables;
- closeness of the mechanism of relationship between variables in an isolated regression;
- the effect of heteroscedasticity, i.e. the absence of residuals' normal distribution for the regression function;

- residual autocorrelation;
- false correlation.

Developing methods to solve these difficulties presents the theoretical part of econometrics. Along with the logically correct formal application of the existing mathematical and statistical tools, the economically adequate formulation of the problem and the subsequent economic interpretation of the results obtained are important components of the success of econometric research.

A great impact to the development of econometric methods and their wide application in practice was given by the informational technologies development and especially the appearance of personal and portable computers. The development of software packages that implement methods for constructing and running econometric models has led to accomplishing of econometric analysis by the widest range of analysts, economists and managers. Up-to-date, the main efforts of the researcher are reduced to the preparation of qualitative data, to the correct formulation of the problem and the economically justified interpretation of the research results. At the same time, the researcher is required to have a knowledge of the areas applicability of the methods used as well as complexity of the process of transferring the theoretical results to real life.

This textbook reflects the content of a one-semester course of lectures to students of economic and financial specialties and corresponds to the State educational standard for the discipline "Econometrics". The textbook consists of fifteen chapters of theoretical parts. Main advantage of this textbook is the fact that all chapters has practical part with explanations of theory. This allow students to get deeply understanding of methodology of econometric research. The appendix provides frequently used statistical aspects. The presentation of the material is focused on the reader who has knowledge of courses in higher mathematics and mathematical statistics, as well as economic statistics. The textbook will also be useful for those who wants to be acquainted with the main tasks, models and methods of econometrics.

Theoretical part of the textbook is added by tutorial "Econometric analysis using Eviews: a practical tutorial" which consider theoretical part of this textbook by practical examples.

CHAPTER 1. AN OVERVIEW OF REGRESSION ANALYSIS

1.1 What is econometrics?

Econometrics as a science has been developed in the first half of the 20th century as a result of the active use of mathematical and statistical methods for solving the exercises of the economic theory. The term ***econometrics*** was introduced into the scientific literature in 1930th by the Norwegian statistician Ragnar Frisch. He first defined econometrics as a scientific discipline based on a synthesis of economic theory, statistics and mathematics.

Obviously, econometrics means different things to different people. To beginning students, it may seem as if econometrics is an overly complex obstacle to an otherwise useful education. To skeptical observers, econometric results should be trusted only when the steps that produced those results are completely known. To professionals in the field, econometrics is a fascinating set of techniques that allows the measurement and analysis of economic phenomena and the prediction of future economic trends.

Economic phenomena and processes are interrelated and interdependent. The consequence of this is a fact that values of the corresponding economic indicators change over time, taking into account these relationships. For example, it is known that aggregate demand depends on the price level, consumption depends on disposable income, investment - on the interest rate, and so on. The researcher is faced with the task of identifying such relationships, their quantitative assessment and study possibility of using the identified relationships in economic analysis and forecasting. Hence, econometrics involved development of appropriate tools and its application to solve specific practical economic exercises.

Econometrics has both a formal definition and a larger context. Although you can easily memorize the formal definition, you'll get the complete picture only by understanding the many uses of and alternative approaches to econometrics. That said, we need a formal definition. ***Econometrics*** - literally, "economic

measurement” - is the quantitative measurement and analysis of actual economic and business phenomena. It attempts to quantify economic reality and bridge the gap between the abstract world of economic theory and the real world of human activity. To many students, these worlds may seem far apart. On the one hand, economists theorize equilibrium prices based on carefully conceived marginal costs and marginal revenues; on the other, many firms seem to operate as though they have never heard of such concepts. Econometrics allows us to examine data and to quantify the actions of firms, consumers, and governments. Such measurements have a number of different uses, and an examination of these uses is the first step to understanding econometrics.

Generally, econometrics as a science has three major uses:

1. describing economic reality
2. testing hypotheses about economic theory and policy
3. forecasting future economic activity

The simplest use of econometrics is description. We can use econometrics to quantify economic activity and measure marginal effects because econometrics allows us to estimate numbers and put them in equations that previously contained only abstract symbols. For example, consumer demand for a particular product often can be thought of as a relationship between the quantity demanded (Q) and the product's price (P), the price of a substitute (P_s), and disposable income (Y_d). For most goods, the relationship between consumption and disposable income is expected to be positive, because an increase in disposable income will be associated with an increase in the consumption of the product. Econometrics actually allows us to estimate that relationship based upon past consumption, income, and prices. In other words, a general and purely theoretical functional relationship like:

$$Q = \beta_0 + \beta_1 P + \beta_2 P_s + \beta_3 Y_d \quad (1.1)$$

can become explicit:

$$Q = 27.7 - 0.11P + 0.33P_s + 0.23Y_d \quad (1.2)$$

Equation (1.2) are called estimated equation. This technique gives a much more specific and descriptive picture of the function.¹ Let's compare Equations 1.1 and 1.2. Instead of expecting consumption merely to "increase" if there is an increase in disposable income, Equation 1.2 allows us to expect an increase of a specific amount (0.23 units for each unit of increased disposable income). The number 0.23 is called an estimated regression coefficient, and it is the ability to estimate these coefficients that makes econometrics valuable.

The second use of econometrics is hypothesis testing, the evaluation of alternative theories with quantitative evidence. Much of economics involves building theoretical models and testing them against evidence, and hypothesis testing is vital to that scientific approach. For example, you could test the hypothesis that the product in Equation 1.1 is what economists call a normal good (one for which the quantity demanded increases when disposable income increases). You could do this by applying various statistical tests to the estimated coefficient (0.23) of disposable income (Y_d) in Equation 1.2. At first glance, the evidence would seem to support this hypothesis, because the coefficient's sign is positive, but the "statistical significance" of that estimate would have to be investigated before such a conclusion could be justified. Even though the estimated coefficient is positive, as expected, it may not be sufficiently different from zero to convince us that the true coefficient is indeed positive.

The third and most difficult use of econometrics is to forecast or predict what is likely to happen next quarter, next year, or further into the future, based on what has happened in the past. For

¹ It's of course naïve to build a model of sales (demand) without taking supply into consideration. Unfortunately, it's very difficult to learn how to estimate a system of simultaneous equations until you've learned how to estimate a single equation. As a result, we will postpone our discussion of the econometrics of simultaneous equations until Chapter 14. Until then, you should be aware that we sometimes will encounter right-hand-side variables that are not truly "independent" from a theoretical point of view.

example, economists use econometric models to make forecasts of variables like sales, profits, Gross Domestic Product (GDP), and the inflation rate. The accuracy of such forecasts depends in large measure on the degree to which the past is a good guide to the future. Business leaders and politicians tend to be especially interested in this use of econometrics because they need to make decisions about the future, and the penalty for being wrong (bankruptcy for the entrepreneur and political defeat for the candidate) is high. To the extent that econometrics can shed light on the impact of their policies, business and government leaders will be better equipped to make decisions. For example, if the president of a company that sold the product modeled in Equation 1.1 wanted to decide whether to increase prices, forecasts of sales with and without the price increase could be calculated and compared to help make such a decision.

According from practical purposes of econometrics mentioned above, initial task of any econometric research is development of an mathematical model, adequate to the studied real economic phenomena and processes. The process of building econometric models begins with qualitative research of the problem through methods of economic theory, formulation objectives of the study, definition of the factors affecting the indicator being studied, and finally formulation assumptions about the features of the proposed dependence. On this basis, the studied dependencies are expressed in the form of mathematical equations and relations.

It should be noted, that due to the impossibility of simultaneous taking into account a large number of factors affecting the indicator under study, the assumed dependencies between the variables will not be fulfilled accurately, but with a certain error. In addition, economic phenomena are characterized by internal uncertainty associated with the purposeful activities of economic entities.

Aforesaid postulate application of statistical methods by which selection of significant factors carry out, the presence and degree of relationship closeness between the studied indicators determine, a quantitative assessment of the parameters of the assumed

dependencies is given and the degree of their relevance to reality examine. The main tool of mathematical statistics used for building econometric models, are the methods of *correlation* and *regression analysis*.

The tools commonly used in financial applications are fundamentally the same as those used in economic applications, although the emphasis and the sets of exercises that are likely to be encountered when analyzing the two sets of data are somewhat different. Financial data often differ from macroeconomic data in terms of their frequency, accuracy, seasonality and other properties.

In economics, a serious problem is often a lack of data at hand for testing the theory or hypothesis of interest - this is often called a 'small samples problem'. It might be, for example, that data are required on government budget deficits, or population figures, which are measured only on an annual basis. If the methods used to measure these quantities changed a quarter of a century ago, then only at most twenty-five of these annual observations are usefully available.

Two other exercises that are often encountered in conducting applied econometric work in the arena of economics are those of measurement error and data revisions. These difficulties are simply that the data may be estimated, or measured with error, and will often be subject to several vintages of subsequent revisions. For example, a researcher may estimate an economic model of the effect on national output of investment in computer technology using a set of published data, only to find that the data for the last two years have been revised substantially in the next, updated publication.

These issues are rarely of concern in finance. Financial data come in many shapes and forms, but in general the prices and other entities that are recorded are those at which trades actually took place, or which were quoted on the screens of information providers. There exists, of course, the possibility for typos and possibility for the data measurement method to change (for example, owing to stock index re-balancing or re-basing). But in general the measurement error and revisions exercises are far less serious in the financial context.

Similarly, some sets of financial data are observed at much higher frequencies than macroeconomic data. Asset prices or yields are often available at daily, hourly, or minute-by-minute frequencies. Thus the number of observations available for analysis can potentially be very large - perhaps thousands or even millions, making financial data the envy of macroeconometricians. The implication is that more powerful techniques can often be applied to financial than economic data, and that researchers may also have more confidence in the results.

Furthermore, the analysis of financial data also brings with it a number of new exercises. While the difficulties associated with handling and processing such a large amount of data are not usually an issue given recent and continuing advances in computer power, financial data often have a number of additional characteristics. For example, financial data are often considered very ‘noisy’, which means that it is more difficult to separate underlying trends or patterns from random and uninteresting features. Financial data are also almost always not normally distributed in spite of the fact that most techniques in econometrics assume that they are. High frequency data often contain additional ‘patterns’ which are the result of the way that the market works, or the way that prices are recorded. These features need to be considered in the model-building process, even if they are not directly of interest to the researcher.

1.2 What is regression analysis?

Regression analysis is almost certainly the most important tool at the econometrician’s disposal. But what is regression analysis? In very general terms, regression is concerned with describing and evaluating the *relationship between a given variable and one or more other variables*.

Econometricians use regression analysis to make quantitative estimates of economic relationships that previously have been completely theoretical in nature. After all, anybody can claim that the quantity of iPhones demanded will increase if the price of those

phones decreases (holding everything else constant), but not many people can put specific numbers into an equation and estimate by how many iPhones the quantity demanded will increase for each dollar that price decreases. To predict the direction of the change, you need a knowledge of economic theory and the general characteristics of the product in question. To predict the amount of the change, though, you need a sample of data, and you need a way to estimate the relationship. The most frequently used method to estimate such a relationship in econometrics is regression analysis.

Regression analysis is a statistical technique that attempts to “explain” movements in one variable, the dependent variable (usually denoted as y), as a function of movements in a set of other variables, called the independent (or explanatory, denoted as x) variables, through the quantification of one or more equations. For example, in Equation 1.1:

$$Q = \beta_0 + \beta_1 P + \beta_2 P_s + \beta_3 Y_d \quad (1.1)$$

Q is the dependent variable and P , P_s , and Y_d are the independent variables. Regression analysis is a natural tool for economists because most (though not all) economic propositions can be stated in such equations. For example, the quantity demanded (dependent variable) is a function of price, the prices of substitutes, and income (independent variables).

All readers will be aware of the notion and definition of *correlation*. The correlation between two variables measures the *degree of linear association* between them. If it is stated that y and x are correlated, it means that y and x are being treated in a completely symmetrical way. Thus, it is not implied that changes in x cause changes in y , or indeed that changes in y cause changes in x . Rather, it is simply stated that there is evidence for a linear relationship between the two variables, and that movements in the two are on average related to an extent given by the correlation coefficient. In regression, the dependent variable (y) and the independent variable(s) (x) are treated very differently. The y variable is assumed to be random or ‘stochastic’ in some way, i.e. to

have a *probability distribution*. The x variables are, however, assumed to have fixed (‘non stochastic’) values in repeated samples. Regression as a tool is more flexible and more powerful than correlation.

EXAMPLE 1.1

Job training and worker productivity

Consider the problem where 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

$$wage = f(educ, exper, training) \quad (1.3)$$

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 (1.3) captures the essence of the problem.

As long as dependent variable is a random variable, it can be measured on four different scales: ratio scale, interval scale, ordinal scale, and nominal scale.

Ratio scale: A ratio scale variable has three properties: (1) ratio of two variables, (2) distance between two variables, and (3) ordering of variables. On a ratio scale if, say, Y takes two values, Y_1 and Y_2 , the ratio Y_1/Y_2 and the distance $(Y_2 - Y_1)$ are meaningful quantities, as are comparisons or ordering such as $Y_2 \leq Y_1$ or $Y_1 \geq Y_2$. Most economic variables belong to this category. Thus we can talk about whether GDP is greater this year than the last year, or whether the ratio of GDP this year to the GDP last year is greater than or less than one.

Interval scale: Interval scale variables do not satisfy the first property of ratio scale variables. For example, the distance between two time periods, say, 2007 and 2000 ($2007 - 2000$) is meaningful, but not the ratio $2007/2000$.

Ordinal scale: Variables on this scale satisfy the ordering property of the ratio scale, but not the other two properties. For examples, grading systems, such as A, B, C, or income classification, such as low income, middle income, and high income, are ordinal scale variables, but quantities such as grade A divided by grade B are not meaningful.

Nominal scale: Variables in this category do not have any of the features of the ratio scale variables. Variables such as gender, marital status, and religion are nominal scale variables. Such variables are often called dummy or categorical variables. They are often “quantified” as 1 or 0, 1 indicating the presence of an attribute and 0 indicating its absence. Thus, we can “quantify” gender as male = 1 and female = 0, or vice versa.

1.3 The nature and sources of data

To conduct regression analysis, we need data. There are generally three types of data that are available for analysis: (1) cross-sectional, (2) time series, (3) pooled cross sectional, and (4) panel data.

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.

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.

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 exercises 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 below contains, in abbreviated form, a cross-sectional data set on 526 working individuals for the year 2015. 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).

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

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 2000 to 2016) and variables determined in part by government policy in 2000 (government consumption as a percentage of GDP and adult secondary education rates). Such a data set might be represented as in Table 1.2.

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

<i>obsno</i>	<i>Country</i>	<i>Gpcrgdp</i>	<i>govcons2000</i>	<i>second2000</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 2000 to 2016. The fact that *govcons2000* (government consumption as a percentage of GDP) and *second2000* (percent of adult population with a secondary education) correspond to the year 2000, while *gpcrgdp* is the average growth over the period from 2000 to 2016, does not lead to any special exercises 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.

Like time series data, cross-section data have their particular exercises, particularly the problem of **heterogeneity**. For example, if you collect data on wages in several firms in a given industry at the same point in time, heterogeneity arises because the data may contain small, medium, and large size firms with their individual

characteristics. We show in Chapter 11 how the size or scale effect of heterogeneous units can be taken into account.

Cross-sectional data usually denoted by the subscript i (e.g. Y_i , X_i).

Time series

Time series data, as the name suggests, are data that have been collected over a period of time on one or more variables. Time series data have associated with them a particular frequency of observation or collection of data points. The frequency is simply a measure of the *interval over*, or the *regularity with which*, the data are collected or recorded. 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 Tajikistan is reported monthly. 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 or annual series. Examples of time series data include stock prices, money supply, consumer price index, gross domestic product, annual homicide rates, and automobile sales Figures. Table 1.3 below shows some examples of time series data.

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.

Table 1.3 Time-series data

<i>Series</i>	<i>Frequency</i>
Industrial production	Monthly, or quarterly
Government budget deficit	Annually
Money supply	Weekly
The value of a stock	As transactions occur

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. In the case of dealing with time series data, researcher denote the observation subscript by t (e.g. Y_t , X_t).

Table 1.4 contains a time series data set. The earliest year in the data set is the first 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), *unemp* is the unemployment rate, and *gnp* is the gross national product.

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 1.4 An example of time series data

Obsno	Year	avgmin	Avgcov	Unemp	Gnp
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

Since successive observations in time series data may be correlated, they pose special exercises for regressions involving time series data, particularly, the problem of autocorrelation. In Chapter 10 we will illustrate this problem with appropriate examples.

Time series data pose another problem, namely, that they may not be stationary. Loosely speaking, a time series data set is stationary if its mean and variance do not vary systematically over time. In Chapter 12 we examine the nature of stationary and nonstationary time series and show the special estimation exercises created by the latter.

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 Tajikistan, one in 2007 and one in 2017. In 2007, a random sample of households is surveyed for variables such as income, savings, family size, and so on. In 2017, 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 Tajikistan). 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 indirect tax revenues taken in 2010 and 2014, when there was reform in tax

system in 2013. Suppose we have data on 250 entrepreneurs for 2010 and on 270 entrepreneurs for 2014. One way to store such a data set is given in Table 1.5. below.

Observations 1 through 250 correspond to the entrepreneurs, paid indirect taxes in 2010, and observations 251 through 520 correspond to the entrepreneurs, paid value added taxes in 2014. 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.

Table 1.5. Pooled cross sections: two years of value added taxes

<i>obsno</i>	<i>Year</i>	<i>Vattax</i>	<i>customs</i>	<i>Othtax</i>
1	2010	85500	1600	42
2	2010	67300	1440	36
3	2010	134000	2000	38
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
250	2010	243600	2600	41
251	2014	65000	1250	16
252	2014	182400	2200	20
253	2014	97500	1540	15
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
520	2014	57200	1100	16

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, longitudinal or micro-panel 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 Central Asia on migration flows, tax rates, wage rates, government expenditures, etc., for the years 2005, 2010, and 2015. 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 examples) are followed over a given time period. The data in Table 1.4 are not considered a panel data set because the indirect taxes are likely to be different in 2010 and 2014; 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 1.6. 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.

Table 1.6. A Two-year panel data set on city crime statistics

<i>obsno</i>	<i>City</i>	<i>year</i>	<i>murders</i>	<i>population</i>	<i>Unem</i>	<i>Police</i>
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

A second useful 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. Since 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 two pooled cross sections, where the same cities happen to show up in the same year. 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 multiple observations on the same units allows us to control certain unobserved characteristics of individuals, firms, and so on. As we will see, the use of more than one observation can facilitate causal inference in situations where inferring causality would be very difficult if only a single cross section were available. A second advantage of panel data is that it often allows us to study the importance of lags in behavior or the result of decision making. This information can be significant since many economic policies can be expected to have an impact only after some time has passed.

Continuous and discrete data

As well as classifying data as being of the time series or cross-sectional type, we could also distinguish it as being either continuous or discrete, exactly as their labels would suggest. Continuous data can take on any value and are not confined to take

specific numbers; their values are limited only by precision. For example, the rental yield on a property could be 6.2%, 6.24% or 6.238%, and so on. On the other hand, discrete data can only take on certain values, which are usually integers¹ (whole numbers), and are often defined to be count numbers. For instance, the number of people in a particular underground carriage or the number of shares traded during a day. In these cases, having 86.3 passengers in the carriage or 58571/2 shares traded would not make sense.

Sources of data

The success of any regression analysis depends on the availability of data. Data may be collected by a governmental agency (e.g. the Agency on Statistics), an international agency (e.g. the International Monetary Fund (IMF) or the World Bank), a private organization (e.g. some analytical organization), or individuals or private corporations.

These days the most potent source of data is the Internet. All one has to do is “Google” a topic and it is amazing how many sources one finds.

The quality of data

The fact that we can find data in several places does not mean it is good data. One must check carefully the quality of the agency that collects the data, for very often the data contain errors of measurement, errors of omission or errors of rounding and so on.

Sometime the data are available only at a highly aggregated level, which may not tell us much about the individual entities included in the aggregate. The researchers should always keep in mind that the results of research are only as good as the quality of the data.

Unfortunately, an individual researcher does not have the luxury of collecting data a new and has to depend on secondary sources. But every effort should be made to obtain reliable data.

1.4 Features of interconnection

The main task of econometrics, and certainly for evaluating public policy, is to study and quantify objectively existing relationships and dependencies between economic phenomena (such as output and labour productivity). Greatest interest of researcher devoted to study the causal relationships between the phenomena, which allows to identify the factors that have the main influence on the variation of the studied phenomena and processes.

A causal relationship is a connection between phenomena in which a change in one of them, called a cause, leads to a change in the other, called a consequence. Therefore, the cause always precedes the effect.

The causal relationships in socio-economic phenomena have the following features. First, the cause X and the effect Y do not interact directly, but through intermediate factors, which, as a rule, are omitted in the analysis. Formally, this can be expressed using the scheme $X \rightarrow X' \rightarrow X'' \rightarrow Y$, where X' and X'' represent following intermediate factors.

Secondly, socio-economic phenomena develop and form as a result of the simultaneous impact of a large number of factors. Therefore, one of the main exercises in the investigation of these phenomena becomes the task of identifying the main, sufficient causes and as a consequence abstraction from secondary ones.

The causal relationship between variables can be represented by *functional relationship* or *stochastic dependence*. Functional relationship is a relationship where a certain value of the independent variable corresponds to one and only one value of the dependent variable. The functional relationship expressed in all cases of observation and for each specific unit of the studied population. Such relationships are studied mainly in the natural sciences.

A dependence between random variables that is expressed by a change of the conditional distribution of any of the variables when the other variables are altered called as stochastic dependence. There are many forms of stochastic dependence: if the random variables are not mutually independent, they have to a greater or

lesser degree the property of stochastic dependence. One of the most common types of stochastic dependence is correlation dependence.

According to the analytical expression there are *linear* and *nonlinear relationships*.

A *linear* is relationship in which the change in the dependent variable is directly proportional to the change in the independent variables. Otherwise, the relationship is called *nonlinear*. Analytically, a linear stochastic connection between phenomena can be represented by a straight line equation on a plane, or by a hyper plane equation in n -dimensional space (if there are n factor variables).

It should be noted that simply finding an association between two or more variables might be suggestive, but unless causality can be established, it is rarely compelling. The causal relationship often is so subtle that it fools even the most prominent economists. For example, in the late nineteenth century, English economist Stanley Jevons hypothesized that sunspots caused an increase in economic activity. To test this theory, he collected data on national output (the dependent variable) and sunspot activity (the independent variable) and showed that a significant positive relationship existed. This result led him, and some others, to jump to the conclusion that sunspots did indeed cause output to rise. Such a conclusion was unjustified and called as false or spurious correlation (see box.1.1). In this case for defining causality, the knowledge of economic theory and logical reasoning have to be used.

Box 1.1

Don't be deceived by the words "dependent" and "independent," however. Although many economic relationships are causal by their very nature, a regression result, no matter how statistically significant, cannot prove causality. All regression analysis can do is test whether a significant quantitative relationship exists. Judgments as to causality must also include a healthy dose of economic theory and common sense. For example, the fact that the bell on the door of a flower shop rings just before a customer enters and purchases some flowers by no means implies that the bell causes

purchases! If events A and B are related statistically, it may be that A causes B, that B causes A, that some omitted factor causes both, or that a chance correlation exists between the two.

1.5 Causality and the notion of *ceteris paribus* in econometric analysis

The notion of *ceteris paribus* - which means “other (relevant) factors being equal” - plays an important role in causal analysis. You probably remember from economic theory that most economic questions are *ceteris paribus* by nature. For example, in analyzing consumer demand, we are interested in knowing the effect of changing the price of a good on its quantity demanded, while holding all other factors - such as income, prices of other goods, and individual tastes - fixed. If other factors are not held fixed, then we cannot know the causal effect of a price change on quantity demanded. Holding other factors fixed is critical for policy analysis as well.

In the job training example (Example 1.1), we might be interested in the effect of another week of job training on wages, with all other components being equal (in particular, education and experience). If we succeed in holding all other relevant factors fixed and then find a link between job training and wages, we can conclude that job training has a causal effect on worker productivity. While this may seem pretty simple, even at this early stage it should be clear that, except in very special cases, it will not be possible to literally hold all else equal. The key question in most empirical studies is: Have enough other factors been held fixed to make a case for causality? Rarely is an econometric study evaluated without raising this issue. In most serious applications, the number of factors that can affect the variable of interest - such as criminal activity or wages - is immense, and the isolation of any particular variable may seem like a hopeless effort. However, we will eventually see that, when carefully applied, econometric methods can simulate a *ceteris paribus* experiment.

At this point, we cannot yet explain how econometric methods can be used to estimate *ceteris paribus* effects, so we will consider some exercises that can arise in trying to infer causality in economics. We do not use any equations in this discussion. For each

example, the problem of inferring causality disappears if an appropriate experiment can be carried out. Thus, it is useful to describe how such an experiment might be structured, and to observe that, in most cases, obtaining experimental data is impractical. It is also helpful to think about why the available data fails to have the important features of an experimental data set. We rely for now on your intuitive understanding of terms such as *random*, *independence*, and *correlation*, all of which should be familiar from an introductory probability and statistics course. We begin with an example that illustrates some of these important issues.

EXAMPLE 1.2

Effects of fertilizer on crop yield

Let's consider the effects of new fertilizers on crop yields. Suppose the crop under consideration is soybeans. Since fertilizer amount is only one factor affecting yields some others include rainfall, quality of land, and presence of parasites—this issue must be posed as a *ceteris paribus* question. One way to determine the causal effect of fertilizer amount on soybean yield is to conduct an experiment, which might include the following steps. Choose several one-acre plots of land. Apply different amounts of fertilizer to each plot and subsequently measure the yields; this gives us a cross-sectional data set. Then, use statistical methods (to be introduced in Chapter 2) to measure the association between yields and fertilizer amounts.

As described earlier, this may not seem like a very good experiment, because we have said nothing about choosing plots of land that are identical in all respects except for the amount of fertilizer. In fact, choosing plots of land with this feature is not feasible: some of the factors, such as land quality, cannot even be fully observed. How do we know the results of this experiment can be used to measure the *ceteris paribus* effect of fertilizer? The answer depends on the specifics of how fertilizer amounts are chosen. If the levels of fertilizer are assigned to plots independently of other plot features that affect yield—that is, other characteristics of plots are completely ignored when deciding on fertilizer amounts - then we are in business. We will justify this statement in Chapter 2.

The next example is more representative of the difficulties that arise when inferring causality in applied economics.

EXAMPLE 1.3

Measuring the return to education

Labor economists and policy makers have long been interested in the “return to education.” Somewhat informally, the question is posed as follows: If a person is chosen from the population and given another year of education, by how much will his or her wage increase? As with the previous examples, this is a *ceteris paribus* question, which implies that all other factors are held fixed while another year of education is given to the person.

We can imagine a social planner designing an experiment to get at this issue, much as the agricultural researcher can design an experiment to estimate fertilizer effects. One approach is to emulate the fertilizer experiment in Example 1.2: Choose a group of people, randomly give each person an amount of education (some people have an eighth grade education, some are given a high school education, etc.), and then measure their wages (assuming that each then works in a job). The people here are like the plots in the fertilizer example, where education plays the role of fertilizer and wage rate plays the role of soybean yield. As with Example 1.2, if levels of education are assigned independently of other characteristics that affect productivity (such as experience and innate ability), then an analysis that ignores these other factors will yield useful results. Again, it will take some effort in Chapter 2 to justify this claim; for now we state it without support.

Unlike the fertilizer-yield example, the experiment described in Example 1.3 is infeasible. The moral issues, not to mention the economic costs, associated with randomly determining education levels for a group of individuals are obvious. As a logistical matter, we could not give someone only an eighth grade education if he or she already has a college degree. Even though experimental data cannot be obtained for measuring the return to education, we can certainly collect non-experimental data on education levels and wages for a large group by sampling randomly from the population of working people. Such data are available from a variety of surveys used in labor economics, but these data sets have a feature that makes it difficult to estimate the *ceteris paribus* return to education.

People *choose* their own levels of education, and therefore education levels are probably not determined independently of all other factors affecting wage. This problem is a feature shared by most non-experimental data sets. One factor that affects wage is experience in the work force. Since pursuing more education generally requires postponing entering the work force, those with more education usually have less experience. Thus, in a non-experimental data set on wages and education, education is likely to be negatively associated with a key variable that also affects wage. It is also believed that people with more innate ability often choose higher levels of education. Since higher ability leads to higher wages, we again have a correlation between education and a critical factor that affects wage. The omitted factors of experience and ability in the wage example have analogs in the fertilizer example. Experience is generally easy to measure and therefore is similar to a variable such as rainfall. Ability, on the other hand, is nebulous and difficult to quantify; it is similar to land quality in the fertilizer example. As we will see throughout this text, accounting for other observed factors, such as experience, when estimating the *ceteris paribus* effect of another variable, such as education, is relatively straightforward. We will also find that accounting for inherently unobservable factors, such as ability, is much more problematical. It is fair to say that many of the advances in econometric methods have tried to deal with unobserved factors in econometric models.

One final parallel can be drawn between Examples 1.2 and 1.3. Suppose that in the fertilizer example, the fertilizer amounts were not entirely determined at random. Instead, the assistant who chose the fertilizer levels thought it would be better to put more fertilizer on the higher quality plots of land. (Agricultural researchers should have a rough idea about which plots of land are better quality, even though they may not be able to fully quantify the differences.) This situation is completely analogous to the level of schooling being related to unobserved ability in Example 1.3. Because better land leads to higher yields, and more fertilizer was used on the better plots, any observed relationship between yield and fertilizer might be spurious.

The first three examples we have discussed have dealt with cross-sectional data at various levels of aggregation (for example, at the individual or city levels). The same hurdles arise when inferring causality in time series exercises.

EXAMPLE 1.4

The effect of the minimum wage on unemployment

An important, and perhaps contentious, policy issue concerns the effect of the minimum wage on unemployment rates for various groups of workers. While this problem can be studied in a variety of data settings (cross-sectional, time series, or panel data), time series data are often used to look at aggregate effects. An example of a time series data set on unemployment rates and minimum wages was given in Table 1.3.

1.6 Econometric packages for modelling economic and financial data

As the name suggests, this section contains descriptions of various computer packages that may be employed to estimate econometric models. The number of available packages is large, and over time, all packages have improved in breadth of available techniques, and have also converged in terms of what is available in each package. Some readers may already be familiar with the use of one or more packages, and if this is the case, this section may be skipped. For those who do not know how to use any econometrics software, or have not yet found a package which suits their requirements, then read on.

What packages are available?

Although this list is by no means exhaustive, a set of widely used packages is given in Table 1.1. The programs can usefully be categorized according to whether they are fully interactive, (menu-driven), command-driven (so that the user has to write mini-programs), or somewhere in between. Many driven packages, which are usually based on a standard Microsoft Windows graphical user interface, are almost certainly the easiest for novices to get started with, for they require little knowledge of the structure of the

package, and the menus can usually be negotiated simply. Eviews is a package that falls into this category.

On the other hand, some such packages are often the least flexible, since the menus of available options are fixed by the developers, and hence if one wishes to build something slightly more complex or just features for modelling financial time series. Also, many packages developed for time series analysis, such as TSP ('Time Series Processor'), can also now be used for cross-sectional or panel data. Of course, this choice may be made for you if your institution offers or supports only one or two of the above possibilities. Otherwise, sensible questions to ask yourself are:

- is the package suitable for your *intended applications* - for example, does the software have the capability for the models that you want to estimate? Can it handle sufficiently large databases?
- is the package *user-friendly*?
- is it *fast*?
- how much does it *cost*?
- is it *accurate*?
- is the package *discussed* or *supported* in a standard textbook, as Eviews is in this book?
- does the package have *readable and comprehensive textbooks*? is help available online?
- does the package come with *free technical support* so that you can e-mail the developers with queries?

A great deal of useful information can be obtained most easily from the web pages of the software developers. Additionally, many journals (including the *Journal of Applied Econometrics*, the *Economic Journal*, the *International Journal of Forecasting* and the *American Statistician*) publish software reviews that seek to evaluate and compare the packages' usefulness for a given purpose. Three reviews that this author has been involved with, that are relevant for chapter 8 of this text in particular, are Brooks (1997) and Brooks, Burke and Persaud (2001, 2003).

The Eviews package will be employed in this text because it is simple to use, menu-driven, and will be sufficient to estimate most

of the models required for this book. The following section gives an introduction to this software and outlines the key features and how basic tasks are executed.

SUMMARY

In this introductory Chapter, we have discussed the purpose and scope of econometric analysis. Econometrics is used in all applied economic fields to test economic theories, inform government and private policy makers, and to predict economic time series. Sometimes an econometric model is derived from a formal economic model, but in other cases econometric models are based on informal economic reasoning and intuition. The goal of any econometric analysis is to estimate the parameters in the model and to test hypotheses about these parameters; the values and signs of the parameters determine the validity of an economic theory and the effects of certain policies.

Cross-sectional, time series, pooled cross-sectional, and panel data are the most common types of data structures that are used in applied econometrics. Data sets involving a time dimension, such as time series and panel data, require special treatment because of the correlation across time of most economic time series. Other issues, such as trends and seasonality, arise in the analysis of time series data but not cross sectional data.

We discussed the notions of *ceteris paribus* and causal inference. In most cases, hypotheses in the social sciences are *ceteris paribus* in nature: all other relevant factors must be fixed when studying the relationship between two variables. Because of the non-experimental nature of most data collected in the social sciences, uncovering causal relationships is very challenging.

CHAPTER 2. THE SIMPLE REGRESSION MODEL

2.1. Definition of the simple regression model

For simplicity, suppose for now that it is believed that y depends on only one x variable. Again, this is of course a severely restricted case, but the case of more explanatory variables will be considered in the next chapter. Three examples of the kind of relationship that may be of interest include:

- How asset returns vary with their level of market risk;
- Measuring the long-term relationship between stock prices and dividends;
- Constructing an optimal hedge ratio.

Suppose that a researcher has some idea that there should be a relationship between two variables y and x , and that financial theory suggests that an increase in x will lead to an increase in y . A sensible first stage to testing whether there is indeed an association between the variables would be to form a scatter plot of them. Suppose that the outcome of this plot is Figure 2.1.

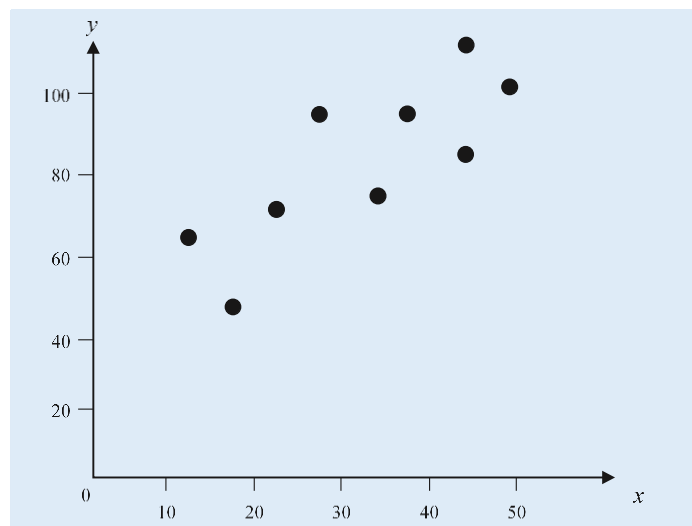


Figure 2.1 Scatter plot of y and x .

In this case, it appears that there is an approximate positive linear relationship between x and y which means that increases in x

are usually accompanied by increases in y , and that the relationship between them can be described approximately by a straight line. It would be possible to draw by hand onto the graph a line that appears to fit the data. The intercept and slope of the line fitted by eye could then be measured from the graph. However, in practice such a method is likely to be laborious and inaccurate.

It would therefore be of interest to determine to what extent this relationship can be described by an equation that can be estimated using a defined procedure. It is possible to use the general equation for a straight line to get the line that best ‘fits’ the data.

$$y = a + bx \quad (2.1)$$

Graphically, equation (2.1) can be represented as following:

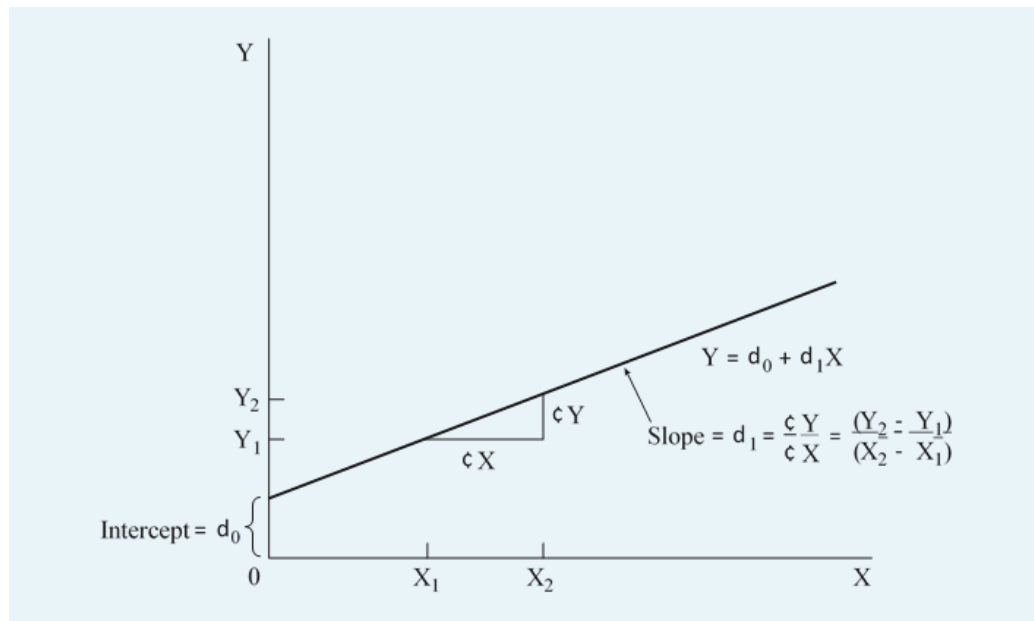


Figure 2.2 Graphical representation of the coefficients of the regression line.

The graph of the equation $y = a + \beta x$ is linear with a constant slope equal to $\beta = \Delta y / \Delta x$.

Equation (2.1) states that y , the dependent variable, is a single-equation linear function of x , the independent variable. The model is a single-equation model because it's the only equation specified.

The researcher would then be seeking to find the values of the parameters or coefficients, α and β , which would place the line as close as possible to all of the data points taken together. However, this equation ($y = \alpha + \beta x$) is an exact one. Assuming that this equation is appropriate, if the values of α and β had been calculated, then given a value of x , it would be possible to determine with certainty what the value of y would be. Imagine a model which says with complete certainty what the value of one variable will be given any value of the other!

Besides the variation in the dependent variable (y) that is caused by the independent variable (x), there is almost always variation that comes from other sources as well. This additional variation comes in part from omitted explanatory variables (e.g., x_2 and x_3). However, even if these extra variables are added to the equation, there still is going to be some variation in y that simply cannot be explained by the model.² This variation probably comes from sources such as omitted influences, measurement error, incorrect functional form, or purely random and totally unpredictable occurrences. By random we mean something that has its value determined entirely by chance.

Econometricians admit the existence of such inherent unexplained variation ("error") by explicitly including a stochastic (or random) error term in their regression models. A **stochastic error term** is a term that is added to a regression equation to introduce all of the variation in y that cannot be explained by the included x s. As mentioned in chapter 1, this relationship called as

² The exception would be the extremely rare case where the data can be explained by some sort of physical law and are measured perfectly. Here, continued variation would point to an omitted independent variable. A similar kind of problem is often encountered in astronomy, where planets can be discovered by noting that the orbits of known planets exhibit variations that can be caused only by the gravitational pull of another heavenly body. Absent these kinds of physical laws, researchers in economics and business would be foolhardy to believe that *all* variation in y can be explained by a regression model because there are always elements of error in any attempt to measure a behavioral relationship.

stochastic dependence. It is, in effect, a symbol of the econometrician's ignorance or inability to model all the movements of the dependent variable. The error term (sometimes called a disturbance term) usually is referred to with the symbol (u), although other symbols (like ε or v) sometimes are used.

The addition of a stochastic error term (u) to Equation (2.1) results in a typical regression equation:

$$y = a + bx + u \quad (2.2)$$

Equation (2.2) also addresses the issue of the functional relationship between y and x . If the other factors in u are held fixed, so that the change in u is zero, $\Delta u = 0$, then x has a *linear* effect on y :

$$\Delta y = \beta \Delta x \text{ if } \Delta u = 0 \quad (2.2a)$$

Thus, the change in y is simply β multiplied by the change in x . This means that β is the **slope parameter** in the relationship between y and x holding the other factors in u fixed; it is of primary interest in applied economics. The **intercept parameter** a also has its uses, although it is rarely central to an analysis.

Let's consider equation (2.2) in details. General equation (2.2) can be thought of as having two components, the *deterministic* component and the *stochastic*, or random, component. The expression $a + \beta_1 x$ is called the *deterministic* component of the regression equation because it indicates the value of that is determined by a given value of x , which is assumed to be non-stochastic. This deterministic component can also be thought of as the **expected value** of y given x , the mean value of the y_s associated with a particular value of x . For example, if the average height of all 13-year-old girls is 5 feet, then 5 feet is the expected value of a girl's height given that she is 13. The deterministic part of the equation may be written

$$E(y/x) = a + \beta x \quad (2.3)$$

which states that the expected value of y given x , denoted as $E(y/x)$, is a linear function of the independent variable (or variables if there are more than one).

Unfortunately, the value of y observed in the real world is unlikely to be exactly equal to the deterministic expected value $E(y/x)$. After all, not all 13-year-old girls are 5 feet tall. As a result, the stochastic element (u) must be added to the equation:

$$y = E(y/x) + \varepsilon = a + \beta x + u \quad (2.4)$$

Box 2.1 Reasons for stochastic error term

The stochastic error term must be present in a regression equation because there are at least sources of variation in Y other than the variation in the included X s:

1. Many minor influences on Y are omitted from the equation (for example, because data are unavailable).
2. It is virtually impossible to avoid some sort of measurement error in the dependent variable.
3. The underlying theoretical equation might have a different functional form (or shape) than the one chosen for the regression. For example, the underlying equation might be nonlinear.
4. All attempts to generalize human behavior must contain at least some amount of unpredictable or purely random variation.

More detailed explanation of stochastic error presented in Box 2.1. To get a better feeling for these components of the stochastic error term, let's think about a consumption function (aggregate consumption as a function of aggregate disposable income). First, consumption in a particular year may have been less than it would have been because of uncertainty over the future course of the economy. Since this uncertainty is hard to measure, there might be

no variable measuring consumer uncertainty in the equation. In such a case, the impact of the omitted variable (consumer uncertainty) would likely end up in the stochastic error term. Second, the observed amount of consumption may have been different from the actual level of consumption in a particular year due to an error (such as a sampling error) in the measurement of consumption in the System of National Accounts. Third, the underlying consumption function may be nonlinear, but a linear consumption function might be estimated. (To see how this incorrect functional form would cause errors, see Figure 2.3.) Fourth, the consumption function attempts to portray the behavior of people, and there is always an element of unpredictability in human behavior. At any given time, some random event might increase or decrease aggregate consumption in a way that might never be repeated and couldn't be anticipated.

One source of stochastic error is the use of an incorrect functional form. For example, if a linear functional form is used when the underlying relationship is nonlinear, systematic errors (the *us*) will occur. These nonlinearities are just one component of the stochastic error term. The others are omitted variables, measurement error, and purely random variation.

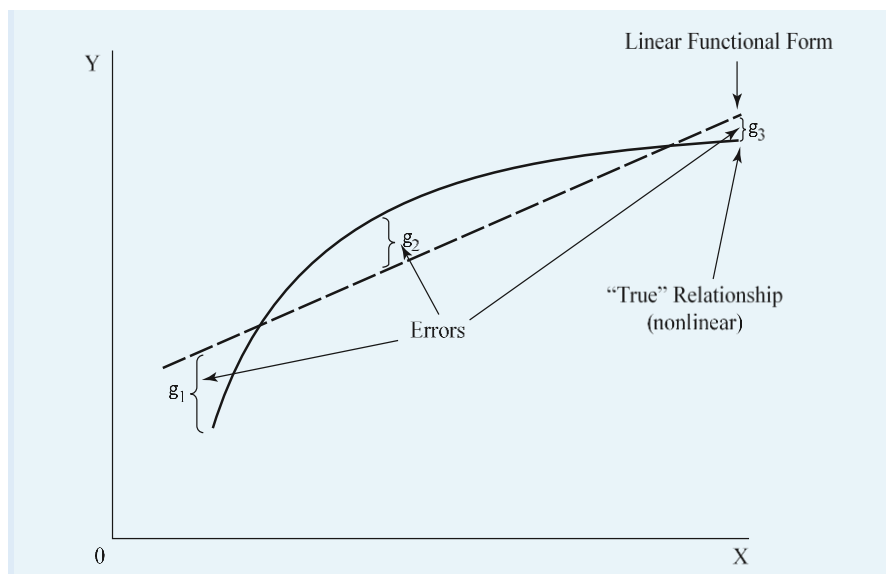


Figure 2.3 Error caused by using a linear functional form to model a nonlinear relationship.

These possibilities explain the existence of a difference between the observed values of y and the values expected from the deterministic component of the equation, $E(y/x)$. These sources of error will be covered in more detail in the following chapters, but for now it is enough to recognize that in econometric research there will always be some stochastic or random element, and, for this reason, an error term must be added to all regression equations.

EXAMPLE 2.1
Tomatoes yield and fertilizer

Suppose that tomatoes yield is determined by the model

$$yield = a + \beta \times fertilizer + u \quad (2.5)$$

so that $y = yield$ and $x = fertilizer$. The agricultural researcher is interested in the effect of fertilizer on yield, holding other factors fixed. This effect is given by β . The error term u contains factors such as land quality, rainfall, and so on. The coefficient β measures the effect of fertilizer on yield, holding other factors fixed: $\Delta yield = \beta \times \Delta fertilizer$.

EXAMPLE 2.2
A simple wage equation

A model relating a person's wage to observed education and other unobserved factors is

$$wage = a + \beta \times educ + u \quad (2.6)$$

If $wage$ is measured in dollars per hour and $educ$ is years of education, then β measures the change in hourly wage given another year of education, holding all other factors fixed. Some of those factors include labor force experience, innate ability, tenure with current employer, work ethics, and innumerable other things.

The linearity of (2.2) implies that a one-unit change in x has the *same* effect on y , regardless of the initial value of x . This is

unrealistic for many economic applications. For example, in the wage-education example, we might want to allow for *increasing* returns: the next year of education has a *larger* effect on wages than did the previous year. The most difficult issue to address is whether model (2.2) really allows us to draw *ceteris paribus* conclusions about how x affects y . We just saw in equation (2.2a) that β *does* measure the effect of x on y , holding all other factors (in u) fixed. Is this the end of the causality issue? Unfortunately, no. How can we hope to learn in general about the *ceteris paribus* effect of x on y , holding other factors fixed, when we are ignoring all those other factors? As we will see in Section 2.5, we are only able to get reliable estimators of a and β from a random sample of data when we make an assumption restricting how the unobservable u is related to the explanatory variable x . Without such a restriction, we will not be able to estimate the *ceteris paribus* effect, β . Because u and x are random variables, we need a concept grounded in probability. Before we state the key functional form assumption about how x and u are related, there is one assumption about u that we can always make. As long as the intercept a is included in the equation, nothing is lost by assuming that the average value of u in the population is zero.

Mathematically,

$$E(u) = 0 \tag{2.7}$$

Importantly, assume (2.7) says nothing about the relationship between u and x but simply makes a statement about the distribution of the unobservables in the population. Using the previous examples for illustration, we can see that assumption (2.7) is not very restrictive. In Example 2.1, we lose nothing by normalizing the unobserved factors affecting tomatoes yield, such as land quality, to have an average of zero in the population of all cultivated plots. The same is true of the unobserved factors in Example 2.2. Without loss of generality, we can assume that things such as average ability are zero in the population of all working people. If you are not convinced, you can work through Problem 2.2 to see that we can

always redefine the intercept in equation (2.2) to make (2.7) true. We now turn to the crucial assumption regarding how u and x are related. A natural measure of the association between two random variables is the *correlation coefficient*. If u and x are *uncorrelated*, then, as random variables, they are not *linearly* related. Assuming that u and x are uncorrelated goes a long way toward defining the sense in which u and x should be unrelated in equation (2.2). But it does not go far enough, because correlation measures only linear dependence between u and x . Correlation has a somewhat counterintuitive feature: it is possible for u to be uncorrelated with x while being correlated with functions of x , such as x^2 . (See Section B for further discussion.) This possibility is not acceptable for most regression purposes, as it causes exercises for interpreting the model and for deriving statistical properties. A better assumption involves the *expected value of u given x* . Because u and x are random variables, we can define the conditional distribution of u given any value of x . In particular, for any x , we can obtain the expected (or average) value of u for that slice of the population described by the value of x . The crucial assumption is that the average value of u does *not* depend on the value of x . We can write this as

$$E(u | x) = E(u) = 0 \tag{2.8}$$

where the second equality follows from (2.7). The first equality in equation (2.8) is the new assumption, called the **zero conditional mean assumption**. It says that, for any given value of x , the average of the unobservables is the same and therefore must equal the average value of u in the entire population.

Let us see what (2.8) entails in the wage example. To simplify the discussion, assume that u is the same as innate ability. Then (2.8) requires that the average level of ability is the same regardless of years of education. For example, if $E(abil8)$ denotes the average ability for the group of all people with eight years of education, and $E(abil16)$ denotes the average ability among people in the population with 16 years of education, then (2.8) implies that these must be the same. In fact, the average ability level must be the same

for *all* education levels. If, for example, we think that average ability increases with years of education, then (2.8) is false. (This would happen if, on average, people with more ability choose to become more educated.) As we cannot observe innate ability, we have no way of knowing whether or not average ability is the same for all education levels. But this is an issue that we must address before applying simple regression analysis.

In the fertilizer example, if fertilizer amounts are chosen independently of other features of the plots, then (2.8) will hold: the average land quality will not depend on the amount of fertilizer. However, if more fertilizer is put on the higher quality plots of land, then the expected value of u changes with the level of fertilizer, and (2.8) fails.

2.2. Deriving ordinary least square estimates

Now that we have discussed the basic components of the simple regression model, we will address the important issue of how to estimate the parameters a and β in equation (2.2). To do this, we need a sample from the population. Let $\{(x_i, y_i): i=1, \dots, n\}$ denote a random sample of size n from the population. Since these data come from (2.2), we can write

$$y_i = a + bx_i + u_i \quad (2.9)$$

for each i . Here, u_i is the error term for observation i since it contains all factors affecting y_i other than x_i .

As an example, x_i might be the annual income and y_i the annual savings for family i during a particular year. If we have collected data on 15 families, then $n=15$. A scatter plot of such a data set is given in Figure 2.4, along with the (necessarily fictitious) population regression function.

We must decide how to use these data to obtain estimates of the intercept and slope in the population regression of savings on income. There are several ways to motivate the following

estimation procedure. We will use (2.7) and an important implication of assumption (2.8): in the population, u has a zero mean and is uncorrelated with x . Therefore, we see that u has zero expected value and that the *covariance* between x and u is zero:

$$E(u) = 0 \quad (2.10)$$

$$\text{cov}(x, u) = E(xu) = 0 \quad (2.11)$$

In terms of the observable variables x and y and the unknown parameters a and β , equations (2.10) and (2.11) can be written as

$$E(y - a - \beta x) = 0 \quad (2.12)$$

and

$$E[x(y - a - \beta x)] = 0 \quad (2.13)$$

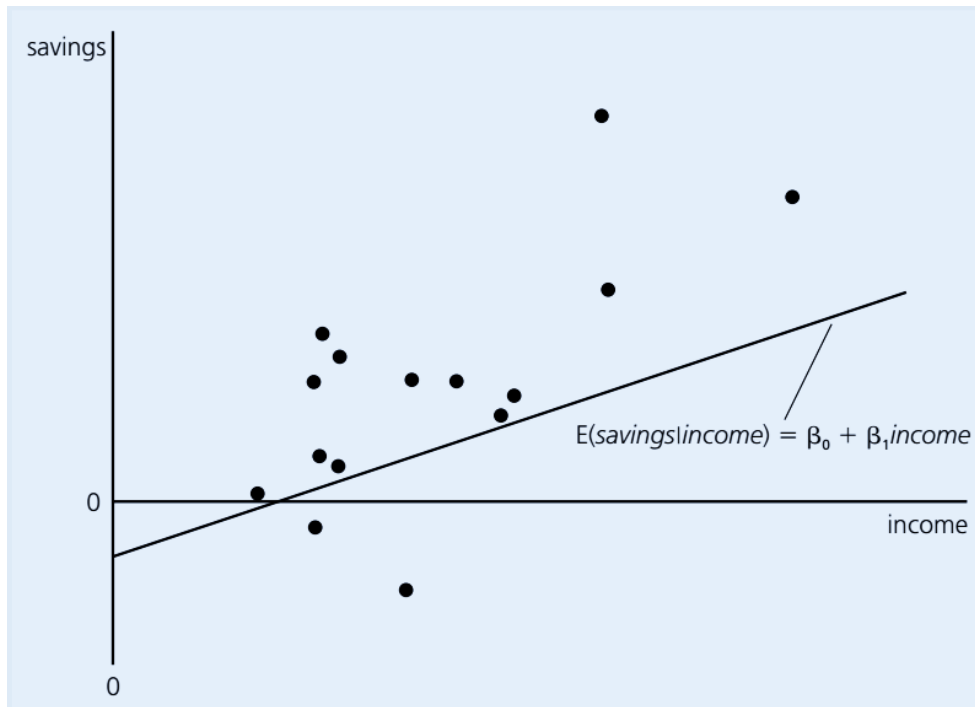


Figure 2.4 Scatter plot of savings and income for 15 families and the population regression $E(\text{savings}|\text{income})=a+\beta\text{income}$.

respectively. Equations (2.12) and (2.13) imply two restrictions on the joint probability distribution of (x,y) in the population. Since there are two unknown parameters to estimate, we might hope that equations (2.12) and (2.13) can be used to obtain good estimators of a and β . In fact, they can be. Given a sample of data, we choose estimates $\hat{a}$ and $\hat{\beta}$ to solve the *sample* counterparts of (2.12) and (2.13):

$$n^{-1} \sum_{i=1}^n (y_i - \hat{a} - \hat{\beta}x_i) = 0 \quad (2.14)$$

$$n^{-1} \sum_{i=1}^n x_i (y_i - \hat{a} - \hat{\beta}x_i) = 0 \quad (2.15)$$

This is an example of the *method of moments* approach to estimation. (See Section C.4 for a discussion of different estimation approaches). Another method, which is generally used is ordinary least squared method (OLS).

So how in our example with savings and income are the appropriate values of a and β determined? The estimators a and β are chosen so that the (vertical) distances from the data points to the fitted lines are minimized (so that the line fits the data as closely as possible). The parameters are thus chosen to minimize collectively the (vertical) distances from the data points to the fitted line. This could be done by ‘eye-balling’ the data and, for each set of variables y and x , one could form a scatter plot and draw on a line that looks as if it fits the data well by hand, as in Figure 2.4.

Note that the vertical distances are usually minimized rather than the horizontal distances or those taken perpendicular to the line. This arises as a result of the assumption that x is fixed in repeated samples, so that the problem becomes one of determining the appropriate model for y given (or conditional upon) the observed values of x .

This ‘eye-balling’ procedure may be acceptable if only indicative results are required, but of course this method, as well as being tedious, is likely to be imprecise. The most common method used to fit a line to the data is known as **ordinary least squares** (OLS). This approach forms the workhorse of econometric model

estimation, and will be discussed in detail in this and subsequent chapters.

Suppose now, for ease of exposition, that the sample of data contains only five observations. The method of OLS entails taking each vertical distance from the point to the line, squaring it and then minimizing the total sum of the areas of squares (hence ‘least squares’), as shown in Figure 2.5. This can be viewed as equivalent to minimizing the sum of the areas of the squares drawn from the points to the line.

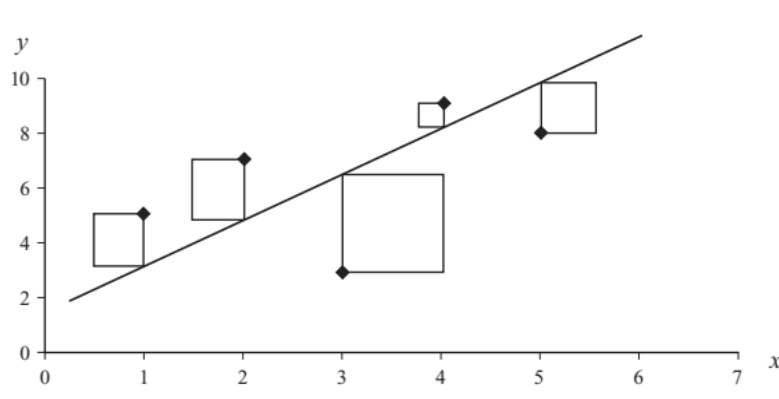


Figure 2.5 Method of OLS fitting a line to the data by minimizing the sum of squared residuals.

Tightening up the notation, let y_t denote the actual data point for observation t and let $\hat{y}_t$ denote the fitted value from the regression line – in other words, for the given value of x of this observation t , $\hat{y}_t$ is the value for y which the model would have predicted. Note that a hat (^) over a variable or parameter is used to denote a value estimated by a model. Finally, let $\hat{u}_t$ denote the residual, which is the difference between the actual value of y and the value fitted by the model for this data point - i.e. $(y_t - \hat{y}_t)$. This is shown for just one observation t in Figure 2.6.

What is done is to minimize the sum of the $\hat{u}_t^2$. The reason that the sum of the squared distances is minimized rather than, for example, finding the sum of $\hat{u}_t$ that is as close to zero as possible, is that in the latter case some points will lie above the line while others lie below it. Then, when the sum to be made as close to zero as

possible is formed, the points above the line would count as positive values, while those below would count as negatives. So these distances will in large part cancel each other out, which would mean that one could fit virtually any line to the data, so long as the sum of the distances of the points above the line and the sum of the distances of the points below the line were the same. In that case, there would not be a unique solution for the estimated coefficients. In fact, any fitted line that goes through the mean of the observations (i.e. $\bar{x}$, $\bar{y}$) would set the sum of the $\hat{u}_t$ to zero.

However, taking the squared distances ensures that all deviations that enter the calculation are positive and therefore do not cancel out. So minimizing the sum of squared distances is given by

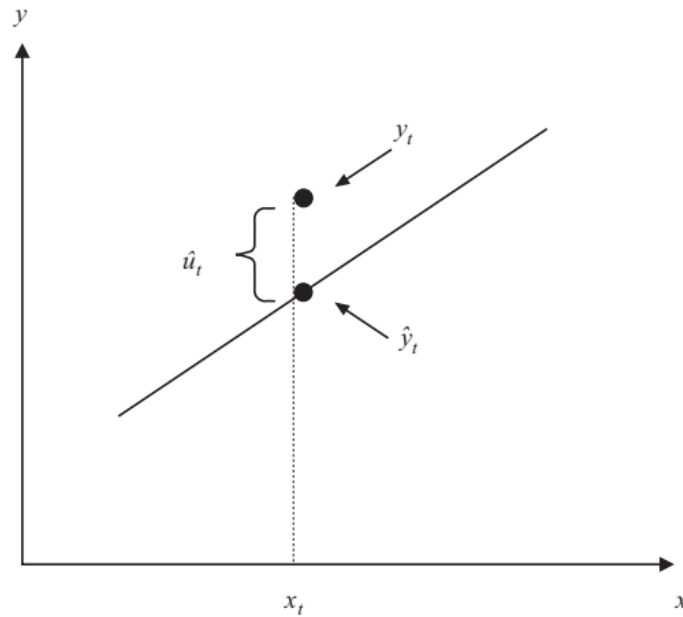


Figure 2.6 Plot of a single observation, together with the line of best fit, the residual and the fitted value.

minimizing $\hat{u}_1^2 + \hat{u}_2^2 + \hat{u}_3^2 + \hat{u}_4^2 + \hat{u}_5^2$ or minimizing $\left(\sum_{t=1}^5 \hat{u}_t^2 \right)$.

This sum is known as the residual sum of squares (RSS). But what is $\hat{u}_t$? Again, it is the difference between the actual point and the

line, $y_t - \hat{y}_t$. So minimizing $\sum_t \hat{u}_t^2$ is equivalent to minimizing $\sum_t (y_t - \hat{y}_t)^2$.

Letting $\hat{a}$ and $\hat{\beta}$ denote the values of a and β selected by minimizing the RSS, respectively, the equation for the fitted line is given by $\hat{y}_t = \hat{a} + \hat{\beta}x_t$. Now let L denote the RSS, which is also known as a loss function. Take the summation over all of the observations, i.e. from $t = 1$ to T , where T is the number of observations

$$L = \sum_{t=1}^T (y_t - \hat{y}_t)^2 = \sum_{t=1}^T (y_t - \hat{a} - \hat{\beta}x_t)^2 \quad (2.16)$$

L is minimized with respect to (w.r.t.) $\hat{a}$ and $\hat{\beta}$, to find the values of a and β which minimize the residual sum of squares to give the line that is closest to the data. So L is differentiated w.r.t. $\hat{a}$ and $\hat{\beta}$, setting the first derivatives to zero. A derivation of the ordinary least squares (OLS) estimator is given in the appendix to this chapter. The coefficient estimators for the slope and the intercept are given by

$$\hat{\beta} = \frac{\sum x_t y_t - T \bar{x} \bar{y}}{\sum x_t^2 - T \bar{x}^2} \quad (2.17)$$

$$\hat{a} = \bar{y} - \hat{\beta} \bar{x} \quad (2.18)$$

Equations (2.9) and (2.10) state that, given only the sets of observations x_t and y_t , it is always possible to calculate the values of the two parameters, $\hat{a}$ and $\hat{\beta}$, that best fit the set of data. Equation (2.11) is the easiest formula to use to calculate the slope estimate, but the formula can also be written, more intuitively, as

$$\hat{\beta} = \frac{\sum (x_t - \bar{x})(y_t - \bar{y})}{\sum (x_t - \bar{x})^2} \quad (2.19)$$

which is equivalent to the sample covariance between x and y divided by the sample variance of x . To reiterate, this method of finding the optimum is known as OLS. It is also worth noting that it is obvious from the equation for $\hat{a}$ that the regression line will go through the mean of the observations - i.e. that the point $(\bar{x}, \bar{y})$ lies on the regression line. An immediate implication is that if x and y are positively correlated in the sample, then $\hat{\beta}$ is positive; if x and y are negatively correlated, then $\hat{\beta}$ is negative.

We now present several examples of simple regression obtained by using real data. In other words, we find the intercept and slope estimates with equations (2.9) and (2.10).

EXAMPLE 2.3

Weight guesser

Let's look at a fairly simple example of regression analysis. Suppose you've accepted a summer job as a weight guesser at the Rudaki park. Customers pay two somoni each, which you get to keep if you guess their weight within 3 kg. If you miss by more than 3 kg, then you have to return the two somoni and give the customer a small prize that you buy from Rudaki park for three somoni each. Luckily, the friendly managers of Rudaki park have arranged a number of marks on the wall behind the customer so that you are capable of measuring the customer's height accurately.

Unfortunately, there is a one and half meter wall, so you can tell little about the person except for height and (usually) gender. On your first day on the job, you do so poorly that you work all day and somehow manage to lose two dollars, so on the second day you decide to collect data to run a regression to estimate the relationship between weight and height. Since most of the participants are male, you decide to limit your sample to males. You hypothesize the following theoretical relationship:

$$y_i = a + \beta x_i + u_i \quad (2.9)$$

where y_i - the weight (in pounds) of the i th customer

x_i - the height (in inches above 5 feet) of the i th customer

u_i - the value of the stochastic error term for the i th customer

In this case, the sign of the theoretical relationship between height and weight is believed to be positive (signified by the positive sign above β_1 in the general theoretical equation), but you must quantify that relationship in

order to estimate weights when given heights. To do this, you need to collect a data set, and you need to apply regression analysis to the data. The next day you collect the data summarized in Table 2.1 and run your regression on the computer, obtaining the following estimates:

$$\hat{a} = 60.4$$

2.1 Data for and results of the weight-guessing equation.

Observation i (1)	Height Above 5' X _i (2)	Weight Y _i (3)	Predicted Weight $\hat{Y}_i$ (4)	Residual e _i (5)	\$ Gain or Loss (6)
1	5.0	140.0	135.3	4.7	+2.00
2	9.0	157.0	160.8	-3.8	+2.00
3	13.0	205.0	186.3	18.7	-3.00
4	12.0	198.0	179.9	18.1	-3.00
5	10.0	162.0	167.2	-5.2	+2.00
6	11.0	174.0	173.6	0.4	+2.00
7	8.0	150.0	154.4	-4.4	+2.00
8	9.0	165.0	160.8	4.2	+2.00
9	10.0	170.0	167.2	2.8	+2.00
10	12.0	180.0	179.9	0.1	+2.00
11	11.0	170.0	173.6	-3.6	+2.00
12	9.0	162.0	160.8	1.2	+2.00
13	10.0	165.0	167.2	-2.2	+2.00
14	12.0	180.0	179.9	0.1	+2.00
15	8.0	160.0	154.4	5.6	+2.00
16	9.0	155.0	160.8	-5.8	+2.00
17	10.0	165.0	167.2	-2.2	+2.00
18	15.0	190.0	199.1	-9.1	+2.00
19	13.0	185.0	186.3	-1.3	+2.00
20	11.0	155.0	173.6	-18.6	-3.00
					TOTAL = \$25.00

This means that the equation

$$\text{estimated weight} = 60.40 + 0.638 \times \text{height} \quad (2.20)$$

is worth trying as an alternative to just guessing the weights of your customers. Such an equation estimates weight with a constant base of 60.40 kg and adds 0.638 kg for every centimeter of height over 1,5 meters. Note that the sign of $\hat{\beta}$ is positive, as you expected.

How well does the equation work? To answer this question, you need to calculate the residuals ($y_i - \hat{y}_i$) from equation (2.20) to see how many were greater than ten. As can be seen in the last column in Table 2.1, if you had

applied the equation to these 20 people, you wouldn't exactly have gotten rich, but at least you would have earned \$25.00 instead of losing \$2.00. Equation (2.13) would probably help a beginning weight guesser, but it could be improved by adding other variables or by collecting a larger sample.

Now, let's come back to our example with wage and education and estimate regression equation for this case.

EXAMPLE 2.4

Wage and education

For the population of people in the work force in 2016, let $y = \text{wage}$, where *wage* is measured in dollars per hour. Thus, for a particular person, if $\text{wage} = 6.75$, the hourly wage is \$6.75. Let $x = \text{educ}$ denote years of schooling; for example, $\text{educ} = 11$ corresponds to a complete high school education. Since the average wage in the sample is \$5.90, the consumer price index indicates that this amount is equivalent to \$16.64 in 2017 dollars.

Using the data, where $n = 526$ individuals, we obtain the following OLS regression line (or sample regression function):

$$\text{wage} = -0.90 + 0.54 \text{educ} \quad (2.21)$$

We must interpret this equation with caution. The intercept of 0.90 literally means that a person with no education has a predicted hourly wage of 90 cents an hour. This, of course, is silly. It turns out that no one in the sample has less than eight years of education, which helps to explain the crazy prediction for a zero education value. For a person with eight years of education, the predicted wage is $\text{wage} = -0.90 + 0.54 \times (8) = 3.42$, or \$3.42 per hour (in 2016 dollars). The slope estimate in (2.14) implies that one more year of education increases hourly wage by 54 cents an hour. Therefore, four more years of education increase the predicted wage by $4 \times 0.54 = 2.16$ or \$2.16 per hour. These are fairly large effects. Because of the linear nature of (2.14), another year of education increases the wage by the same amount, regardless of the initial level of education.

EXAMPLE 2.5

CEO salary and return on equity

For the population of chief executive officers (CEO), let y be annual salary (*salary*) in thousands of dollars. Thus, $y = 856.3$ indicates an annual salary of \$856,300, and $y = 1452.6$ indicates a salary of \$1,452,600. Let x be

the average return equity (*roe*) for the CEO's firm for the previous three years. (Return on equity is defined in terms of net income as a percentage of common equity.) For example, if *roe* 10, then average return on equity is 10%. To study the relationship between this measure of firm performance and CEO compensation, we postulate the simple model

$$salary = a + \beta roe + u$$

The slope parameter β measures the change in annual salary, in thousands of dollars, when return on equity increases by one percentage point. Because a higher *roe* is good for the company, we think $\beta \geq 0$. The data set contains information on 209 CEOs for the year 2016. In this sample, the average annual salary is \$1281120, with the smallest and largest being \$223000 and \$14822000, respectively. The average return on equity for the years 2014, 2015, and 2016 is 17.18%, with the smallest and largest values being 0.5 and 56.3%, respectively. Using the data, the OLS regression line relating *salary* to *roe* is

$$\hat{salary} = 963.191 + 18.501 roe \quad (2.22)$$

where the intercept and slope estimates have been rounded to three decimal places; we use “*salary hat*” to indicate that this is an estimated equation. How do we interpret the equation? First, if the return on equity is zero, *roe*=0, then the predicted *salary* is the intercept, 963.191, which equals \$963191 since *salary* is measured in thousands. Next, we can write the predicted change in salary as a function of the change in *roe*: $\Delta \hat{salary} = 18.501(\Delta roe)$. This means that if the return on equity increases by one percentage point, $\Delta roe = 1$, then *salary* is predicted to change by about 18.5, or \$18500. Because (2.13) is a linear equation, this is the estimated change regardless of the initial salary.

We can easily use (2.15) to compare predicted salaries at different values of *roe*. Suppose *roe*=30. Then $\hat{salary} = 963.191 + 18.501 \times 30 = 1518.221$, which is just over \$1.5 million. However, this does *not* mean that a particular CEO whose firm had an *roe*=30 earns \$1518221. There are many other factors that affect salary. This is just our prediction from the OLS regression line (2.15). The estimated line is graphed in Figure 2.7, along with the population regression function $E(salary|roe)$. We will never know the PRF, so we cannot tell how close the SRF is to the PRF. Another sample of data will give a different regression line, which may or may not be closer to the population regression line.

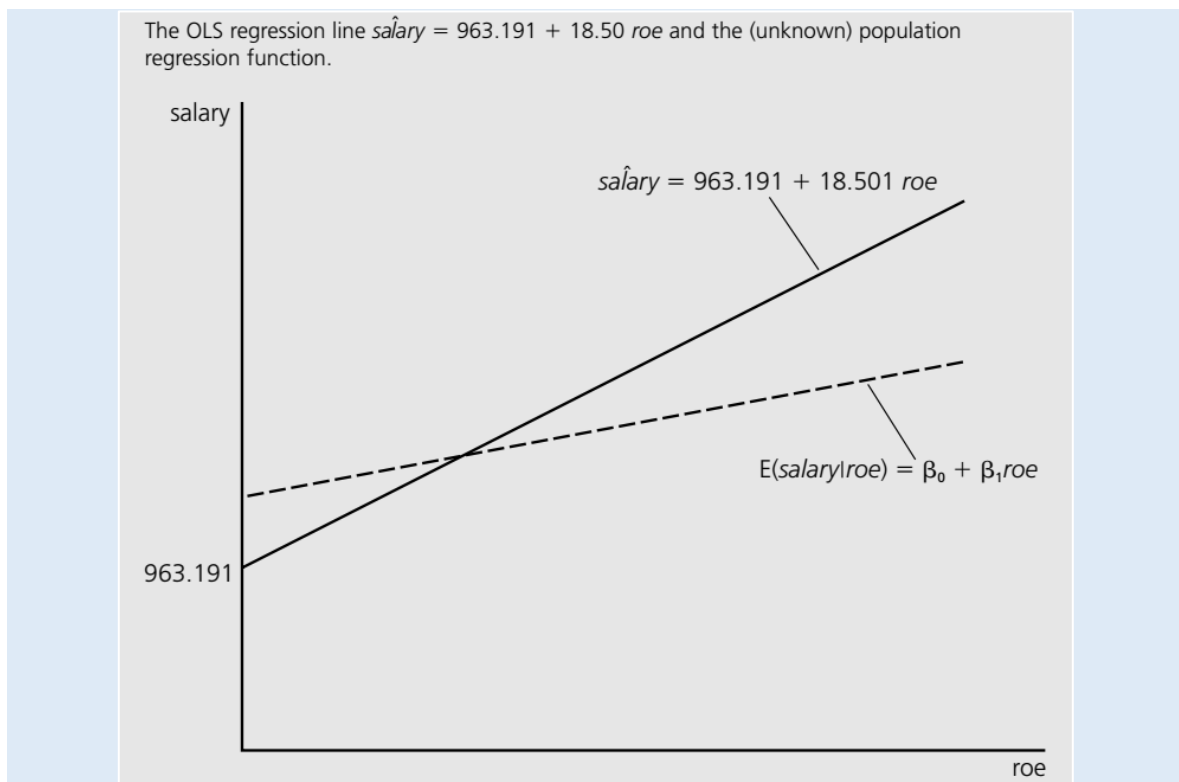


Figure 2.7 Regression line of salary.

2.3 Some further terminology

The population and the sample

The population is the *total collection of all objects or people to be studied*. For example, in the context of determining the relationship between risk and return for UK equities, the population of interest would be all time series observations on all stocks traded on the London Stock Exchange (LSE). The population may be either finite or infinite, while a sample is a selection of *just some items from the population*. In general, either all of the observations for the entire population will not be available, or they may be so many in number that it is infeasible to work with them, in which case a sample of data is taken for analysis. The sample is usually *random*, and it should be *representative* of the population of interest. A random sample is a sample in which each individual item in the population is equally likely to be drawn. The size of the

sample is the number of observations that are available, or that it is decided to use, in estimating the regression equation.

The data generating process, the population regression function and the sample regression function

The population regression function (PRF) is a description of the model that is thought to be generating the actual data and it represents the *true relationship between the variables*. The population regression function is also known as the data generating process (DGP). The PRF embodies the true values of α and β , and is expressed as

$$y_t = \alpha + \beta x_t + u_t \quad (2.23)$$

Note that there is a disturbance term in this equation, so that even if one had at one's disposal the entire population of observations on x and y , it would still in general not be possible to obtain a perfect fit of the line to the data. In some textbooks, a distinction is drawn between the PRF (the underlying true relationship between y and x) and the DGP (the process describing the way that the actual observations on y come about), although in this textbook, the two terms will be used synonymously. The sample regression function, SRF, is the relationship that has been estimated using the sample observations, and is often written as

$$\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t \quad (2.24)$$

Notice that there is no error or residual term in (2.17); all this equation states is that given a particular value of x , multiplying it by $\hat{\beta}$ and adding $\hat{\alpha}$ will give the model fitted or expected value for y , denoted $\hat{y}$. It is also possible to write

$$\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t + \hat{u}_t \quad (2.25)$$

Equation (2.18) splits the observed value of y into two components: the fitted value from the model, and a residual term. The SRF is used to infer likely values of the PRF. That is, the

estimates $\hat{a}$ and $\hat{\beta}$ are constructed, for the sample of data at hand, but what is really of interest is the true relationship between x and y - in other words, the PRF is what is really wanted, but all that is ever available is the SRF! However, what can be said is how likely it is, given the Figures calculated for $\hat{a}$ and $\hat{\beta}$, that the corresponding population parameters take on certain values.

Linearity and possible forms for the regression function

In order to use OLS, a model that is *linear* is required. This means that, in the simple bivariate case, the relationship between x and y must be capable of being expressed diagrammatically using a straight line. More specifically, the model must be linear in the parameters (α and β), but it does not necessarily have to be linear in the variables (y and x). By ‘linear in the parameters’, it is meant that the parameters are not multiplied together, divided, squared, or cubed, etc. Models that are not linear in the variables can often be made to take a linear form by applying a suitable transformation or manipulation. For example, consider the following exponential regression model

$$y_t = a x_t^\beta e^{u_t} \quad (2.26)$$

Taking logarithms of both sides, applying the laws of logs and rearranging the right-hand side (RHS)

$$\ln y_t = \ln(a) + \beta \ln x_t + u_t \quad (2.27)$$

where a and β are parameters to be estimated. Now let $a = \ln(a)$, $y_t = \ln(y_t)$, $x_t = \ln(x_t)$

$$y_t = a + \beta x_t + u_t \quad (2.28)$$

This is known as an *exponential regression model* since y varies according to some exponent (power) function of x . In fact, when a regression equation is expressed in ‘double logarithmic form’, which means that both the dependent and the independent variables are natural logarithms, the coefficient estimates are

interpreted as elasticities (strictly, they are unit changes on a logarithmic scale). Thus a coefficient estimate of 1.2 for $\hat{\beta}$ in (2.20) or (2.21) is interpreted as stating that ‘a rise in x of 1% will lead on average, everything else being equal, to a rise in y of 1.2%. Conversely, for y and x in levels rather than logarithmic form (e.g. (2.16)), the coefficients denote unit changes as described above. Similarly, if theory suggests that x should be inversely related to y according to a model of the form

$$y_t = a + \frac{\beta}{x_t} + u_t \quad (2.29)$$

the regression can be estimated using OLS by setting $z_t = 1/x_t$ and regressing y on a constant and z . Clearly, then, a surprisingly varied array of models can be estimated using OLS by making suitable transformations to the variables. On the other hand, some models are *intrinsically non-linear*, e.g.

$$y_t = a + \beta x_t^\gamma + u_t \quad (2.30)$$

2.4 Precision and standard errors

Any set of regression estimates $\hat{\alpha}$ and $\hat{\beta}$ are specific to the sample used in their estimation. In other words, if a different sample of data was selected from within the population, the data points (the x_t and y_t) will be different, leading to different values of the OLS estimates.

Recall that the OLS estimators ($\hat{\alpha}$ and $\hat{\beta}$) are given by (2.9) and (2.10). It would be desirable to have an idea of how ‘good’ these estimates of α and β are in the sense of having some measure of the reliability or precision of the estimators ($\hat{\alpha}$ and $\hat{\beta}$). It is thus useful to know whether one can have confidence in the estimates, and whether they are likely to vary much from one sample to another sample within the given population. An idea of the sampling variability and hence of the precision of the estimates can be calculated using only the sample of data available. This estimate

is given by its standard error. Given assumptions 1 - 4 above, valid estimators of the standard errors can be shown to be given by

$$SE(\hat{a}) = s \sqrt{\frac{\sum x_t^2}{T \sum (x_t - \bar{x})^2}} = s \sqrt{\frac{\sum x_t^2}{T((\sum x_t^2) - T \bar{x}^2)}} \quad (2.31)$$

$$SE(\hat{\beta}) = s \sqrt{\frac{1}{\sum (x_t - \bar{x})^2}} = s \sqrt{\frac{1}{\sum x_t^2 - T \bar{x}^2}} \quad (2.32)$$

where s is the estimated standard deviation of the residuals (see below). These formulae are derived in the appendix to this chapter. It is worth noting that the standard errors give only a general indication of the likely accuracy of the regression parameters. They do not show how accurate a particular set of coefficient estimates is. If the standard errors are small, it shows that the coefficients are likely to be precise on average, not how precise they are for this particular sample. Thus standard errors give a measure of the *degree of uncertainty* in the estimated values for the coefficients. It can be seen that they are a function of the actual observations on the explanatory variable, x , the sample size, T , and another term, s . The last of these is an estimate of the variance of the disturbance term. The actual variance of the disturbance term is usually denoted by σ^2 . How can an estimate of σ^2 be obtained?

Estimating the variance of the error term (σ^2)

From elementary statistics, the variance of a random variable u_t is given by

$$\text{var}(u_t) = E[(u_t) - E(u_t)]^2 \quad (2.33)$$

Assumption 1 of the CLRM was that the expected or average value of the errors is zero. Under this assumption, (2.32) above reduces to

$$\text{var}(u_t) = E[u_t^2] \quad (2.34)$$

So what is required is an estimate of the average value of u_t^2 , which could be calculated as

$$s^2 = \frac{1}{T} \sum u_t^2 \quad (2.35)$$

Unfortunately (2.34) is not workable since u_t is a series of population disturbances, which is not observable. Thus the sample counterpart to u_t , which is $\hat{u}_t$, is used

$$s^2 = \frac{1}{T} \sum \hat{u}_t^2 \quad (2.36)$$

But this estimator is a biased estimator of σ^2 . An unbiased estimator, s^2 , would be given by the following equation instead of the previous one

$$s^2 = \frac{\sum \hat{u}_t^2}{T - 2} \quad (2.37)$$

where $\sum \hat{u}_t^2$ is the residual sum of squares, so that the quantity of relevance for the standard error formulae is the square root of (2.36)

$$s = \sqrt{\frac{\sum \hat{u}_t^2}{T - 2}} \quad (2.38)$$

s is also known as the *standard error of the regression* or the standard error of the estimate. It is sometimes used as a broad measure of the fit of the regression equation. Everything else being equal, the smaller this quantity is, the closer is the fit of the line to the actual data.

Some comments on the standard error estimators

Some general intuition is now given as to why the formulae for the standard errors given by (2.31) and (2.32) contain the terms that they do and in the form that they do. The clearest presentation of this see in box 2.3 below.

Box 2.3 Standard error estimation

- (1) The larger the sample size, T , the smaller will be the coefficient standard errors. T appears explicitly in $SE(\hat{\alpha})$ and implicitly in $SE(\hat{\beta})$. T appears implicitly since the sum $\sum (x_t - \bar{x})^2$ is from $t = 1$ to T . The reason for this is simply that, at least for now, it is assumed that every observation on a series represents a piece of useful information which can be used to help determine the coefficient estimates. So the larger the size of the sample, the more information will have been used in estimation of the parameters, and hence the more confidence will be placed in those estimates.
- (2) Both $SE(\hat{\alpha})$ and $SE(\hat{\beta})$ depend on s^2 (or s). Recall from above that s^2 is the estimate of the error variance. The larger this quantity is, the more dispersed are the residuals, and so the greater is the uncertainty in the model. If s^2 is large, the data points are collectively a long way away from the line.
- (3) The sum of the squares of the x_t about their mean appears in both formulae – since $\sum (x_t - \bar{x})^2$ appears in the denominators. The larger the sum of squares, the smaller the coefficient variances. Consider what happens if $\sum (x_t - \bar{x})^2$ is small or large, as shown in figures 2.8 and 2.9, respectively. In figure 2.8, the data are close together so that $(x_t - \bar{x})^2$ is small. In this first case, it is more difficult to determine with any degree of certainty exactly where the line should be. On the other hand, in figure 2.9, the points are widely dispersed across a long section of the line, so that one could hold

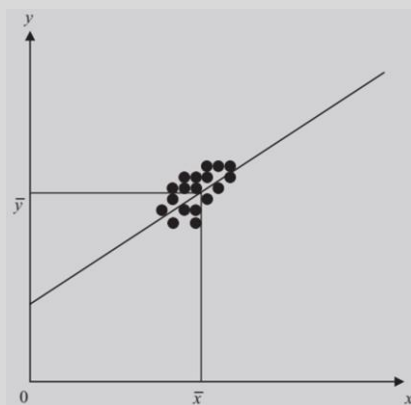


Figure 2.8 Effect on the standard errors of the coefficient estimates when $(x_t - \bar{x})^2$ are narrowly dispersed.

more confidence in the estimates in this case.

- (4) The term x_i^2 affects only the intercept standard error and not the slope standard error. The reason is that x_i^2 measures how far the points are away from the y-axis. Consider figures 2.10 and 2.11. In figure 2.10, all of the points are bunched a long way from the y-axis, which makes it more difficult to accurately estimate the point at which the estimated line crosses the y-axis (the intercept). In figure 2.11, the points collectively are closer to the y-axis and hence it will

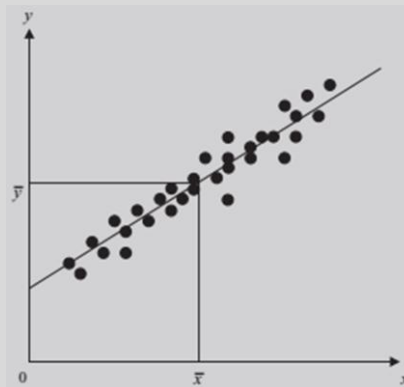


Figure 2.9 Effect on the standard errors of the coefficient estimates when $(x_i - \bar{x})^2$ are widely dispersed.

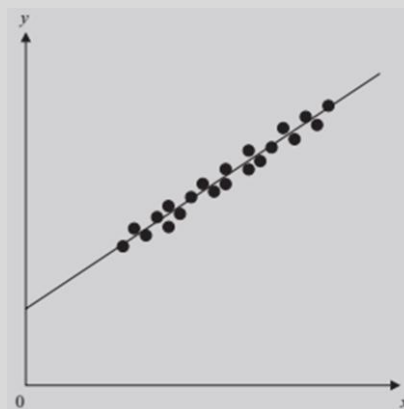


Figure 2.10 Effect on the standard errors of x_i^2 large

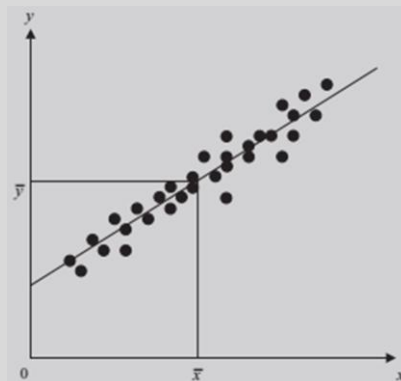


Figure 2.11 Effect on the standard errors of x_t^2 small.

be easier to determine where the line actually crosses the axis. Note that this intuition will work only in the case where all of the x_t are positive!

EXAMPLE 2.7

Wage and education

Assume that the following data have been calculated from a regression of y on a single variable x and a constant over 22 observations

$$\begin{aligned} \sum x_t y_t &= 830102, & T &= 22, & \bar{x} &= 416.5, & \bar{y} &= 86.65 \\ \sum x_t^2 &= 3919654, & RSS &= 130.6 \end{aligned}$$

Determine the appropriate values of the coefficient estimates and their standard errors. This question can simply be answered by plugging the appropriate numbers into the formulae given above. The calculations are

$$\hat{\beta} = \frac{830102 - (22 \times 416.5 \times 86.65)}{3919654 - 22 \times (416.5)^2} = 0.35$$

$$\hat{\alpha} = 86.65 - 0.35 \times 416.5 = -59.12$$

The sample regression function would be written as

$$\hat{y}_t = -59.12 + 0.35x_t$$

Now, turning to the standard error calculations, it is necessary to obtain an estimate, s , of the error variance

$$\begin{aligned} SE(\text{regression}), s &= \sqrt{\frac{\sum u_t^2}{T-2}} = \sqrt{\frac{130.6}{20}} = 2.55 \\ SE(\hat{a}) &= 2.55 \times \sqrt{\frac{3919654}{22 \times (3919654 - 22 \times 416.5^2)}} = 3.35 \\ SE(\hat{\beta}) &= 2.55 \times \sqrt{\frac{1}{22 \times (3919654 - 22 \times 416.5^2)}} = 0.0079 \end{aligned}$$

With the standard errors calculated, the results are written as

$$\hat{y}_t = -\underset{(3.35)}{59.12} + \underset{(0.0079)}{0.35}x_t \quad (2.39)$$

The standard error estimates are usually placed in parentheses under the relevant coefficient estimates.

2.5 Regression through the origin

In rare cases, we wish to impose the restriction that, when $x=0$, the expected value of y is zero. There are certain relationships for which this is reasonable. For example, if income (x) is zero, then income tax revenues (y) must also be zero. In addition, there are exercises where a model that originally has a nonzero intercept is transformed into a model without an intercept. Formally, we now choose a slope estimator, which we call $\tilde{\beta}$, and a line of the form

$$\tilde{y} = \tilde{\beta}x \quad (2.40)$$

where the tildas over $\tilde{\beta}$ and y are used to distinguish this problem from the much more common problem of estimating an intercept along with a slope. Obtaining (2.40) is called **regression through the origin** because the line (2.40) passes through the point $x=0$, $\tilde{y}=0$. To obtain the slope estimate in (2.40), we still rely on the

method of ordinary least squares, which in this case minimizes the sum of squared residuals

$$\sum_{i=1}^n (y_i - \tilde{\beta}x_i)^2 \quad (2.41)$$

Using calculus, it can be shown that $\tilde{\beta}$ must solve the first order condition

$$\sum_{i=1}^n x_i (y_i - \tilde{\beta}x_i) = 0 \quad (2.42)$$

From this we can solve for $\tilde{\beta}$:

$$\tilde{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \quad (2.43)$$

provided that not all the x_i are zero, a case we rule out. Note how $\tilde{\beta}$ compares with the slope estimate when we also estimate the intercept (rather than set it equal to zero). These two estimates are the same if, and only if, $\bar{x} = 0$. Obtaining an estimate of β using regression through the origin is not done very often in applied work, and for good reason: if the intercept $a \neq 0$ then $\tilde{\beta}$ is a biased estimator of β . You will be asked to prove this in Problem 2.8.

SUMMARY

We have introduced the simple linear regression model in this chapter, and we have covered its basic properties. Given a random sample, the method of ordinary least squares is used to estimate the slope and intercept parameters in the population model. We have demonstrated the algebra of the OLS regression line, including computation of fitted values and residuals, and the obtaining of

predicted changes in the dependent variable for a given change in the independent variable.

In Section 2.3, we discussed possible forms for the regression function. The key assumption is that the error term u has zero mean given any value of the independent variable x . Unfortunately, there are reasons to think this is false in many social science applications of simple regression, where the omitted factors in u are often correlated with x .

When we add the assumption that the variance of the error given x is constant, we get simple formulas for the sampling variances of the OLS estimators. As we saw, the variance of the slope estimator $\hat{\beta}_1$ increases as the error variance increases, and it decreases when there is more sample variation in the independent variable. We also derived an unbiased estimator for $\sigma^2 = \text{var}(u)$.

In Section 2.9, we briefly discussed regression through the origin, where the slope estimator is obtained under the assumption that the intercept is zero. Sometimes this is useful, but it appears infrequently in applied work.

APPENDIX 2:

2A. Minimizing the sum of squared residuals

We show that the OLS estimates $\hat{a}$ and $\hat{\beta}$ do minimize the sum of squared residuals, as asserted in Section 2.2. Formally, the problem is to characterize the solutions $\hat{a}$ and $\hat{\beta}$ to the minimization problem

$$\min_{a, \beta} = \sum_{i=1}^n (y_i - a - \beta x_i)^2,$$

where a and β are the dummy arguments for the optimization problem; for simplicity, call this function $Q(a, \beta)$. By a fundamental result from multivariable calculus, a necessary condition for $\hat{a}$ and $\hat{\beta}$ to solve the minimization problem is that the partial derivatives

of $Q(a, \beta)$ with respect to $\hat{a}$ and $\hat{\beta}$ must be zero when evaluated at $\hat{a}, \hat{\beta}$: $\partial Q(\hat{a}, \hat{\beta}) / \partial a = 0$ and $\partial Q(\hat{a}, \hat{\beta}) / \partial \beta = 0$. Using the chain rule from calculus, these two equations become

$$\begin{aligned} -2 \sum_{i=1}^n (y_i - \hat{a} - \hat{\beta}x_i) &= 0. \\ -2 \sum_{i=1}^n x_i (y_i - \hat{a} - \hat{\beta}x_i) &= 0. \end{aligned}$$

These two equations are just (2.14) and (2.15) multiplied by $2n$ and, therefore, are solved by the same $\hat{a}$, and $\hat{\beta}$. How do we know that we have actually minimized the sum of squared residuals? The first order conditions are necessary but not sufficient conditions. One way to verify that we have minimized the sum of squared residuals is to write, for any a and β ,

$$\begin{aligned} Q(a, \beta) &= \sum_{i=1}^n (y_i - \hat{a} - \hat{\beta}x_i + (\hat{a} - b_0) + (\hat{\beta} - b_1)x_i)^2 = \\ &= \sum_{i=1}^n (\hat{u}_i + (\hat{a} - b_0) + (\hat{\beta} - b_1)x_i)^2 = \\ &= \sum_{i=1}^n \hat{u}_i^2 + n(\hat{a} - b_0)^2 + (\hat{\beta} - b_1)^2 \sum_{i=1}^n x_i^2 + 2(\hat{a} - b_0)(\hat{\beta} - b_1) \sum_{i=1}^n x_i \end{aligned}$$

The sum of squared residuals does not depend on a or β , while the sum of the last three terms can be written as

$$\sum_{i=1}^n [(\hat{a} - b_0) + (\hat{\beta} - b_1)x_i]^2,$$

as can be verified by straightforward algebra. Because this is a sum of squared terms, it can be at most zero. Therefore, it is smallest when $b_0 = \hat{a}$ and $b_1 = \hat{\beta}$.

2B. Mathematical derivations of CLRM results

2B.1 Derivation of the OLS coefficient estimator in the bivariate case

$$L = \sum_{t=1}^T (y_t - \hat{y}_t)^2 = \sum_{t=1}^T (y_t - \hat{\alpha} - \hat{\beta}x_t)^2 \quad (2A.1)$$

It is necessary to minimize L w.r.t. $\hat{\alpha}$ and $\hat{\beta}$, to find the values of α and β that give the line that is closest to the data. So L is differentiated w.r.t. $\hat{\alpha}$ and $\hat{\beta}$, and the first derivatives are set to zero. The first derivatives are given by

$$\frac{\partial L}{\partial \hat{\alpha}} = -2 \sum_t (y_t - \hat{\alpha} - \hat{\beta}x_t) = 0 \quad (2A.2)$$

$$\frac{\partial L}{\partial \hat{\beta}} = -2 \sum_t x_t (y_t - \hat{\alpha} - \hat{\beta}x_t) = 0 \quad (2A.3)$$

The next step is to rearrange (2A.2) and (2A.3) in order to obtain expressions for $\hat{\alpha}$ and $\hat{\beta}$. From (2A.2)

$$\sum_t (y_t - \hat{\alpha} - \hat{\beta}x_t) = 0 \quad (2A.4)$$

Expanding the parentheses and recalling that the sum runs from 1 to T so that there will be T terms in $\hat{\alpha}$

$$\sum y_t - T\hat{\alpha} - \hat{\beta} \sum x_t = 0 \quad (2A.5)$$

But $\sum y_t = T\bar{y}$ and $\sum x_t = T\bar{x}$, so it is possible to write (2A.5) as

$$T\bar{y} - T\hat{\alpha} - T\hat{\beta}\bar{x} = 0 \quad (2A.6)$$

or

$$\bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0 \quad (2A.7)$$

From (2A.3)

$$\sum_t x_t (y_t - \hat{a} - \hat{\beta} x_t) = 0 \quad (2A.8)$$

From (2A.7)

$$\hat{a} = y_t - \hat{\beta} \bar{x} \quad (2A.9)$$

Substituting into (2A.8) for $\hat{a}$ from (2A.9)

$$\sum_t x_t (y_t - \bar{y} + \hat{\beta} \bar{x} - \hat{\beta} x_t) = 0 \quad (2A.10)$$

$$\sum_t x_t y_t - \bar{y} \sum_t x_t + \hat{\beta} \bar{x} \sum_t x_t - \hat{\beta} \sum_t x_t^2 = 0 \quad (2A.11)$$

$$\sum_t x_t y_t - T \bar{x} \bar{y} + \hat{\beta} T \bar{x}^2 - \hat{\beta} \sum_t x_t^2 = 0 \quad (2A.12)$$

Rearranging for $\hat{\beta}$,

$$\hat{\beta} (T \bar{x}^2 - \sum_t x_t^2) = T \bar{x} \bar{y} - \sum_t x_t y_t \quad (2A.13)$$

Dividing both sides of (2A.13) by $(T \bar{x}^2 - \sum_t x_t^2)$ gives

$$\hat{\beta} = \frac{\sum_t x_t y_t - T \bar{x} \bar{y}}{\sum_t x_t^2 - T \bar{x}^2} \quad \text{and} \quad \hat{a} = \bar{y} - \hat{\beta} \bar{x} \quad (2A.14)$$

2B.2 Derivation of the OLS standard error estimators for the intercept and slope in the bivariate case

Recall that the variance of the random variable $\hat{a}$ can be written as

$$\text{var}(\hat{a}) = E(\hat{a} - E(\hat{a}))^2 \quad (2A.15)$$

and since the OLS estimator is unbiased

$$\text{var}(\hat{a}) = E(\hat{a} - a)^2 \quad (2A.16)$$

By similar arguments, the variance of the slope estimator can be written as

$$\text{var}(\hat{\beta}) = E(\hat{\beta} - \beta)^2 \quad (2A.17)$$

Working first with (2A.17), replacing $\hat{\beta}$ with the formula for it given by the OLS estimator

$$\text{var}(\hat{\beta}) = E\left(\frac{\sum (x_t - \bar{x})(y_t - \bar{y})}{\sum (x_t - \bar{x})^2} - \beta\right)^2 \quad (2A.18)$$

Replacing y_t with $a + \beta x_t + u_t$, and replacing $\bar{y}$ with $a + \beta \bar{x}$ in (2A.18)

$$\text{var}(\hat{\beta}) = E\left(\frac{\sum (x_t - \bar{x})(a + \beta x_t + u_t - a - \beta \bar{x})}{\sum (x_t - \bar{x})^2} - \beta\right)^2 \quad (2A.19)$$

Cancelling a and multiplying the last β term in (2A.19) by $\frac{\sum (x_t - \bar{x})^2}{\sum (x_t - \bar{x})^2}$

$$\text{var}(\hat{\beta}) = E\left(\frac{\sum (x_t - \bar{x})(\beta x_t + u_t - \beta \bar{x}) - \beta \sum (x_t - \bar{x})^2}{\sum (x_t - \bar{x})^2}\right)^2 \quad (2A.20)$$

Rearranging

$$\text{var}(\hat{\beta}) = E\left(\frac{\sum (x_t - \bar{x})\beta(x_t - \bar{x}) + \sum u_t(x_t - \bar{x}) - \beta \sum (x_t - \bar{x})^2}{\sum (x_t - \bar{x})^2}\right)^2 \quad (2A.21)$$

$$\text{var}(\hat{\beta}) = E \left(\frac{\beta \sum (x_t - \bar{x})^2 + \sum u_t (x_t - \bar{x}) - \beta \sum (x_t - \bar{x})^2}{\sum (x_t - \bar{x})^2} \right)^2 \quad (2A.22)$$

Now the β terms in (2A.22) will cancel to give

$$\text{var}(\hat{\beta}) = E \left(\frac{\sum u_t (x_t - \bar{x})}{\sum (x_t - \bar{x})^2} \right)^2 \quad (2A.23)$$

Now let x_t^* denote the mean-adjusted observation for x_t , i.e. $(x_t - \bar{x})$. Equation (2A.23) can be written

$$\text{var}(\hat{\beta}) = E \left(\frac{\sum u_t x_t^*}{\sum x_t^{*2}} \right)^2 \quad (2A.24)$$

The denominator of (2A.24) can be taken through the expectations operator under the assumption that x is fixed or non-stochastic

$$\text{var}(\hat{\beta}) = \frac{1}{(\sum x_t^{*2})^2} E(\sum u_t x_t^*)^2 \quad (2A.25)$$

Writing the terms out in the last summation of (2A.25)

$$\text{var}(\hat{\beta}) = \frac{1}{(\sum x_t^{*2})^2} E(u_1 x_1^* + u_2 x_2^* \dots + u_T x_T^*)^2 \quad (2A.26)$$

Now expanding the brackets of the squared term in the expectations operator of (2A.26)

$$\text{var}(\hat{\beta}) = \frac{1}{(\sum x_t^{*2})^2} E(u_1^2 x_1^{*2} + u_2^2 x_2^{*2} \dots + u_T^2 x_T^{*2} + \text{cross products}) \quad (2A.27)$$

where ‘cross-products’ in (2A.27) denotes all of the terms $u_i x_i^*, u_j x_j^*$ ($i \neq j$). These cross-products can be written as $u_i u_j x_i^* x_j^*$ ($i \neq j$) and their expectation will be zero under the assumption that the error terms are uncorrelated with one another. Thus, the ‘cross products’ term in (2A.27) will drop out. Recall also from the chapter text that $E(u_t^2)$ is the error variance, which is estimated using s^2

$$\text{var}(\hat{\beta}) = \frac{1}{\left(\sum x_t^{*2}\right)^2} E\left(s^2 x_1^{*2} + s^2 x_2^{*2} \dots + s^2 x_T^{*2}\right) \quad (2A.28)$$

which can also be written

$$\text{var}(\hat{\beta}) = \frac{s^2}{\left(\sum x_t^{*2}\right)^2} \left(x_1^{*2} + x_2^{*2} \dots + x_T^{*2}\right) = \frac{s^2 \sum x_t^{*2}}{\left(\sum x_t^{*2}\right)^2} \quad (2A.29)$$

A term in x_t^{*2} can be cancelled from the numerator and denominator of (2A.29), and recalling that $x_t^* = (x_t - \bar{x})$, this gives the variance of the slope coefficient as

$$\text{var}(\hat{\beta}) = \frac{s^2}{\sum (x_t - \bar{x})^2} \quad (2A.30)$$

so that the standard error can be obtained by taking the square root of (2A.30)

$$SE(\hat{\beta}) = s \sqrt{\frac{1}{\sum (x_t - \bar{x})^2}} \quad (2A.31)$$

Turning now to the derivation of the intercept standard error, this is in fact much more difficult than that of the slope standard error. In fact, both are very much easier using matrix algebra as shown below. Therefore, this derivation will be offered in summary form. It is possible to express $\hat{a}$ as a function of the true a and of the disturbances, u_t

$$\hat{a} = a + \frac{\sum u_t [\sum x_t^2 - x_t \sum x_t]}{[T \sum x_t^2 - (\sum x_t)^2]} \quad (2A.32)$$

Denoting all of the elements in square brackets as g_t (2A.32) can be written

$$\hat{a} - a = \sum u_t g_t \quad (2A.33)$$

From (2A.15), the intercept variance would be written

$$(\text{var } \hat{a}) = E(\sum u_t g_t)^2 = \sum g_t^2 E(u_t^2) = s^2 \sum g_t^2 \quad (2A.34)$$

Writing (2A.34) out in full for g_t^2 and expanding the brackets

$$(\text{var } \hat{a}) = \frac{s^2 [T(\sum x_t^2)^2 - 2\sum x_t (\sum x_t^2) \sum x_t + (\sum x_t^2)(\sum x_t)^2]}{[T \sum x_t^2 - (\sum x_t)^2]^2} \quad (2A.35)$$

This looks rather complex, but fortunately, if we take x_t^2 outside the square brackets in the numerator, the remaining numerator cancels with a term in the denominator to leave the required result

$$SE(\hat{a}) = s \sqrt{\frac{\sum x_t^2}{T \sum (x_t - \bar{x})^2}} \quad (2A.36)$$

CHAPTER 3. MULTIPLE REGRESSION ANALYSIS: ESTIMATION

In Chapter 2, we learned how to use simple regression analysis to explain a dependent variable, y , as a function of a single independent variable, x . The primary drawback in using simple regression analysis for empirical work is that it is very difficult to draw *ceteris paribus* conclusions about how x affects y : the key assumption 1 of the CLRM, - that all other factors affecting y are uncorrelated with x - is often unrealistic.

Multiple regression analysis is more amenable to *ceteris paribus* analysis because it allows us to *explicitly* control for many other factors which simultaneously affect the dependent variable. This is important both for testing economic theories and for evaluating policy effects when we must rely on non-experimental data. Because multiple regression models can accommodate many explanatory variables that may be correlated, we can hope to infer causality in cases where simple regression analysis would be misleading.

Naturally, if we add more factors to our model that are useful for explaining y , then more of the variation in y can be explained. Thus, multiple regression analysis can be used to build better models for predicting the dependent variable. An additional advantage of multiple regression analysis is that it can incorporate fairly general functional form relationships. In the simple regression model, only one function of a single explanatory variable can appear in the equation. As we will see, the multiple regression model allows for much more flexibility. Section 3.1 formally introduces the multiple regression model and further discusses the advantages of multiple regression over simple regression. In Section 3.2, we demonstrate how to estimate the parameters in the multiple regression model using the method of ordinary least squares. In Sections 3.3, 3.4, and 3.5, we describe various statistical properties of the OLS estimators, including unbiasedness and efficiency.

The multiple regression model is still the most widely used vehicle for empirical analysis in economics and other social sciences. Likewise, the method of ordinary least squares is popularly used for estimating the parameters of the multiple regression model.

3.1 Motivation for multiple regression

The model with two independent variables

We begin with some simple examples to show how multiple regression analysis can be used to solve exercises that cannot be solved by simple regression.

The first example is a simple variation of the wage equation introduced in Chapter 2 for obtaining the effect of education on hourly wage:

$$wage = a + \beta_1 educ + \beta_2 exper + u \quad (3.1)$$

where *exper* is years of labor market experience. Thus, *wage* is determined by the two explanatory or independent variables, education and experience, and by other unobserved factors, which are contained in *u*. We are still primarily interested in the effect of *educ* on *wage*, holding fixed all other factors affecting *wage*; that is, we are interested in the parameter β_1 .

Compared with a simple regression analysis relating *wage* to *educ*, equation (3.1) effectively takes *exper* out of the error term and puts it explicitly in the equation. Because *exper* appears in the equation, its coefficient, β_2 , measures the ceteris paribus effect of *exper* on *wage*, which is also of some interest. Not surprisingly, just as with simple regression, we will have to make assumptions about how *u* in (3.1) is related to the independent variables, *educ* and *exper*. However, as we will see in Section 3.2, there is one thing of which we can be confident: since (3.1) contains experience explicitly, we will be able to measure the effect of education on wage, holding experience fixed. In a simple regression analysis -

which puts *exper* in the error term - we would have to assume that experience is uncorrelated with education, a tenuous assumption.

As a second example, consider the problem of explaining the effect of per student spending (*expend*) on the average standardized test score (*avgscore*) at the high school level. Suppose that the average test score depends on funding, average family income (*avginc*), and other unobservables:

$$avgscore = a + \beta_1 \text{expend} + \beta_2 \text{avginc} + u \quad (3.2)$$

The coefficient of interest for policy purposes is β_1 , the ceteris paribus effect of *expend* on *avgscore*. By including *avginc* explicitly in the model, we are able to control for its effect on *avgscore*. This is likely to be important because average family income tends to be correlated with per student spending: spending levels are often determined by both property and local income taxes. In simple regression analysis, *avginc* would be included in the error term, which would likely be correlated with *expend*, causing the OLS estimator of β_1 in the two-variable model to be biased.

In the two previous similar examples, we have shown how observable factors other than the variable of primary interest [*educ* in equation (3.1), *expend* in equation (3.2)] can be included in a regression model. Generally, we can write a model with two independent variables as

$$y = a + \beta_1 x_1 + \beta_2 x_2 + u \quad (3.3)$$

where a is the intercept, β_1 measures the change in y with respect to x_1 , holding other factors fixed, and β_2 measures the change in y with respect to x_2 , holding other factors fixed.

Multiple regression analysis is also useful for generalizing functional relationships between variables. As an example, suppose family consumption (*cons*) is a quadratic function of family income (*inc*):

$$cons = a + \beta_1 inc + \beta_2 inc^2 + u \quad (3.4)$$

where u contains other factors affecting consumption. In this model, consumption depends on only one observed factor, income; so it might seem that it can be handled in a simple regression framework. But the model falls outside simple regression because it contains two functions of income, inc and inc^2 (and therefore three parameters, a , β_1 , and β_2). Nevertheless, the consumption function is easily written as a regression model with two independent variables by letting $x_1=inc$ and $x_2=inc^2$.

Mechanically, there will be *no* difference in using the method of ordinary least squares (introduced in Section 3.2) to estimate equations as different as (3.1) and (3.4). Each equation can be written as (3.3), which is all that matters for computation. There is, however, an important difference in how one *interprets* the parameters. In equation (3.1), β_1 is the ceteris paribus effect of *educ* on *wage*. The parameter β_1 has no such interpretation in (3.4). In other words, it makes no sense to measure the effect of *inc* on *cons* while holding inc^2 fixed, because if *inc* changes, then so must inc^2 ! Instead, the change in consumption with respect to the change in income - the marginal propensity to consume - is approximated by

$$\frac{\Delta cons}{\Delta inc} \approx \beta_1 + 2\beta_2 inc$$

In other words, the marginal effect of income on consumption depends on 2 as well as on 1 and the level of income. This example shows that, in any particular application, the definition of the independent variables are crucial. But for the theoretical development of multiple regression, we can be vague about such details. We will study examples like this more completely in Chapter 6.

In the model with two independent variables, the key assumption about how u is related to x_1 and x_2 is

$$E(u \mid x_1, x_2) = 0 \tag{3.5}$$

The interpretation of condition (3.5) is similar to the interpretation of Assumption 1 of the CLRM for simple regression analysis. It means that, for any values of x_1 and x_2 in the population, the average unobservable is equal to zero. As with simple regression, the important part of the assumption is that the expected value of u is the same for all combinations of x_1 and x_2 ; that this common value is zero is no assumption at all as long as the intercept a is included in the model (see Section 2.1). How can we interpret the zero conditional mean assumption in the previous examples? In equation (3.1), the assumption is $E(u|educ,exper)=0$. This implies that other factors affecting *wage* are not related on average to *educ* and *exper*. Therefore, if we think innate ability is part of u , then we will need average ability levels to be the same across all combinations of education and experience in the working population. This may or may not be true, but, as we will see in Section 3.3, this is the question we need to ask in order to determine whether the method of ordinary least squares produces unbiased estimators.

The example measuring student performance (equation (3.2)) is similar to the wage equation. The zero conditional mean assumption is $E(u|expend,avginc)=0$, which means that other factors affecting test scores-school or student characteristics - are, on average, unrelated to per student funding and average family income.

When applied to the quadratic consumption function in (3.4), the zero conditional mean assumption has a slightly different interpretation. Written literally, equation (3.5) becomes $E(u|inc, inc^2)=0$. Since inc^2 is known when inc is known, including inc^2 in the expectation is redundant: $E(u|inc, inc^2)=0$ is the same as $E(u|inc)=0$. Nothing is wrong with putting inc^2 along with inc in the expectation when stating the assumption, but $E(u|inc)=0$ is more concise.

The model with k independent variables

Once we are in the context of multiple regression, there is no need to stop with two independent variables. Multiple regression analysis allows many observed factors to affect y . In the wage

example, we might also include amount of job training, years of tenure with the current employer, measures of ability, and even demographic variables like number of siblings or mother's education. In the school funding example, additional variables might include measures of teacher quality and school size. The general **multiple linear regression model** (also called the multiple regression model) can be written in the population as

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + u \quad (3.6)$$

where a is the **intercept**, β_1 is the parameter associated with x_1 , β_2 is the parameter associated with x_2 , and so on. Since there are k independent variables and an intercept, equation (3.6) contains $k+1$ (unknown) population parameters. For shorthand purposes, we will sometimes refer to the parameters other than the intercept as **slope parameters**, even though this is not always literally what they are. [See equation (3.4), where neither β_1 nor β_2 is itself a slope, but together they determine the slope of the relationship between consumption and income]. The terminology for multiple regression is similar to that for simple regression and is given in Table 3.1. Just as in simple regression, the variable u is the **error term** or **disturbance**. It contains factors other than $x_1, x_2, \dots, x_k$ that affect y .

No matter how many explanatory variables we include in our model, there will always be factors we cannot include, and these are collectively contained in u .

When applying the general multiple regression model, we must know how to interpret the parameters. We will get plenty of practice now and in subsequent chapters, but it is useful at this point to be reminded of some things we already know.

Table 3.1 Terminology for multiple regression.

y	$x_1, x_2, x_3, \dots, x_k$
Dependent variable	Independent variables
Explained variable	Explanatory variables

Response variable	Control variables
Predicted variable	Predictor variables
Regressand	Regressors

Suppose that CEO salary (*salary*) is related to firm sales and CEO tenure with the firm by

$$\log(\text{salary}) = a + \beta_1 \log(\text{sales}) + \beta_2 \text{ceoten} + \beta_3 \text{ceoten}^2 + u \quad (3.7)$$

This fits into the multiple regression model (with $k=3$) by defining $y=\log(\text{salary})$, $x_1=\log(\text{sales})$, $x_2=\text{ceoten}$, and $x_3=\text{ceoten}^2$. As we know from Chapter 2, the parameter β_1 is the (ceteris paribus) *elasticity* of *salary* with respect to *sales*. If $\beta_3 = 0$, then $100\beta_2$ is approximately the ceteris paribus percentage increase in *salary* when *ceoten* increases by one year. When $\beta_3 \neq 0$, the effect of *ceoten* on *salary* is more complicated. We will postpone a detailed treatment of general models with quadratics until Chapter 8.

Equation (3.7) provides an important reminder about multiple regression analysis. The term “linear” in multiple linear regression model means that equation (3.6) is linear in the *parameters*, β_j . Equation (3.7) is an example of a multiple regression model that, while linear in the β_j , is a nonlinear relationship between *salary* and the variables *sales* and *ceoten*. Many applications of multiple linear regression involve nonlinear relationships among the underlying variables. The key assumption for the general multiple regression model is easy to state in terms of a conditional expectation:

$$E(u \mid x_1, x_2, \dots, x_k) = 0 \quad (3.8)$$

At a minimum, equation (3.8) requires that all factors in the unobserved error term be uncorrelated with the explanatory variables. It also means that we have correctly accounted for the functional relationships between the explained and explanatory variables. Any problem that allows u to be correlated with any of the independent variables causes (3.8) to fail. In Section 3.3, we will

show that assumption (3.8) implies that OLS is unbiased and will derive the bias that arises when a key variable has been omitted from the equation. In Chapters 15 and 16, we will study other reasons that might cause (3.8) to fail and show what can be done in cases where it does fail.

3.2 Mechanics and interpretation of ordinary least squares

We now summarize some computational and algebraic features of the method of ordinary least squares as it applies to a particular set of data. We also discuss how to interpret the estimated equation.

Obtaining the OLS estimates

We first consider estimating the model with two independent variables. The estimated OLS equation is written in a form similar to the simple regression case:

$$\hat{y} = \hat{a} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \quad (3.9)$$

where $\hat{a}$ is the estimate of a , $\hat{\beta}_1$ is the estimate of β_1 , and $\hat{\beta}_2$ is the estimate of β_2 . But how do we obtain $\hat{a}$, $\hat{\beta}_1$, and $\hat{\beta}_2$? The method of **ordinary least squares** chooses the estimates to minimize the sum of squared residuals. That is, given n observations on y , x_1 , and x_2 , $\{(x_{i1}, x_{i2}, y_i): i=1, 2, \dots, n\}$, the estimates $\hat{a}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are chosen simultaneously to make

$$\sum_{i=1}^n (y_i - \hat{a} - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \quad (3.10)$$

as small as possible.

In order to understand what OLS is doing, it is important to master the meaning of the indexing of the independent variables in (3.10). The independent variables have two subscripts here, i followed by either 1 or 2. The i subscript refers to the observation number. Thus, the sum in (3.10) is over all $i=1$ to n observations. The second index is simply a method of distinguishing between

different independent variables. In the example relating *wage* to *educ* and *exper*, $x_{i1} = \text{educ}_i$ is education for person i in the sample, and $x_{i2} = \text{exper}_i$ is experience for person i . The sum of squared residuals in equation (3.10) is $\sum_{i=1}^n (\text{wage}_i - \hat{a} - \beta_1 \text{educ}_i - \hat{\beta}_2 \text{exper}_i)^2$. In what follows, the i subscript is reserved for indexing the observation number. If we write x_{ij} , then this means the i^{th} observation on the j^{th} independent variable. (Some authors prefer to switch the order of the observation number and the variable number, so that x_{1i} is observation i on variable one. But this is just a matter of notational taste.)

In the general case with k independent variables, we seek estimates $\hat{a}, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ in the equation

$$\hat{y} = \hat{a} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k \quad (3.11)$$

The OLS estimates, $k+1$ of them, are chosen to minimize the sum of squared residuals:

$$\sum_{i=1}^n (y_i - \hat{a} - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik})^2 \quad (3.12)$$

This minimization problem can be solved using multivariable calculus. This leads to $k+1$ linear equations in $k+1$ unknowns $\hat{a}, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$:

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{a} - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) &= 0 \\ \sum_{i=1}^n x_{i1} (y_i - \hat{a} - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) &= 0 \\ \cdot & \\ \cdot & \\ \cdot & \\ \sum_{i=1}^n x_{ik} (y_i - \hat{a} - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) &= 0 \end{aligned} \quad (3.13)$$

These are often called the OLS **first order conditions**. As with the simple regression model in Section 2.2, the OLS first order conditions can be motivated by the method of moments: under assumption (3.8), $E(u)=0$ and $E(x_j u)=0$, where $j = 1, 2, \dots, k$. The equations in (3.13) are the sample counterparts of these population moments.

For even moderately sized n and k , solving the equations in (3.13) by hand calculations is tedious. Nevertheless, modern computers running standard statistics and econometrics software can solve these equations with large n and k very quickly.

There is only one slight caveat: we must assume that the equations in (3.13) can be solved *uniquely* for the $\hat{\beta}_j$. For now, we just assume this, as it is usually the case in well specified models. In Section 3.3, we state the assumption needed for unique OLS estimates to exist (see Assumption MLR.4).

As in simple regression analysis, equation (3.11) is called the **OLS regression line**, or the **sample regression function (SRF)**. We will call $\hat{a}$ the **OLS intercept estimate** and $\hat{\beta}_1, \dots, \hat{\beta}_k$ the **OLS slope estimates** (corresponding to the independent variables $x_1, x_2, \dots, x_k$).

In order to indicate that an OLS regression has been run, we will either write out equation (3.11) with y and $x_1, \dots, x_k$ replaced by their variable names (such as *wage*, *educ*, and *exper*), or we will say that “we ran an OLS regression of y on $x_1, x_2, \dots, x_k$ ” or that “we regressed y on $x_1, x_2, \dots, x_k$.” These are shorthand for saying that the method of ordinary least squares was used to obtain the OLS equation (3.11). Unless explicitly stated otherwise, we always estimate an intercept along with the slopes.

Interpreting the OLS regression equation

More important than the details underlying the computation of the $\hat{\beta}_j$ is the *interpretation* of the estimated equation. We begin with the case of two independent variables:

$$\hat{y} = \hat{a} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \quad (3.14)$$

The intercept $\hat{a}$ in equation (3.14) is the predicted value of y when $x_1=0$ and $x_2=0$. Sometimes setting x_1 and x_2 both equal to zero is an interesting scenario, but in other cases it will not make sense. Nevertheless, the intercept is always needed to obtain a prediction of y from the OLS regression line, as (3.14) makes clear.

The estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ have **partial effect**, or **ceteris paribus**, interpretations. From equation (3.14), we have

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2,$$

so we can obtain the predicted change in y given the changes in x_1 and x_2 . (Note how the intercept has nothing to do with the changes in y .) In particular, when x_2 is held fixed, so that $\Delta x_2 = 0$, then

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1,$$

holding x_2 fixed. The key point is that, by including x_2 in our model, we obtain a coefficient on x_1 with a ceteris paribus interpretation. This is why multiple regression analysis is so useful. Similarly,

$$\Delta \hat{y} = \hat{\beta}_2 \Delta x_2,$$

holding x_1 fixed.

EXAMPLE 3.1

Child mortality in relation to per capita GNP and female literacy rate

Let's consider multiple regression example as behavior of child mortality (CM) in relation to per capita GNP (PGNP) and female literacy rate (FLR). Thus, we formulate our model as

$$CM_i = a + \beta_1 PGNP_i + \beta_2 FLR_i + u_i \quad (3.15)$$

The necessary data are given in Table 3.2. Keep in mind that CM is the number of deaths of children under five per 1000 live births, PGNP is per

capita GNP in 1980, and FLR is measured in percent. Our sample consists of 64 countries.

Using the Eviews statistical package, we obtained the following results:

$$\begin{aligned} \hat{CM}_i &= 263.6416 - 0.0056PGNP_i - 2.2316FLR_i \\ se &= (11.5932) \quad (0.0019) \quad (0.2099) \quad R^2 = 0.7077 \quad (3.16) \\ &\quad \bar{R}^2 = 0.6981^* \end{aligned}$$

where Figures in parentheses are the estimated standard errors. Before we interpret this regression, observe the partial slope coefficient of PGNP, namely, -0.0056. But should that surprise you? Not only that, but the two standard errors are precisely the same, which is again unsurprising. But we did so without the three-step cumbersome procedure.

Let us now interpret these regression coefficients: -0.0056 is the partial regression coefficient of PGNP and tells us that with the influence of FLR held constant, as PGNP increases, say, by a dollar, on average, child mortality goes down by 0.0056 units. To make it more economically interpretable, if the per capita GNP goes up by a thousand dollars, on average, the number of deaths of children under age 5 goes down by about 5.6 per thousand live births.

Table 3.2 Fertility and other data for 64 countries.

FERTILITY AND OTHER DATA FOR 64 COUNTRIES									
Observation	CM	FLFP	PGNP	TFR	Observation	CM	FLFP	PGNP	TFR
1	128	37	1870	6.66	33	142	50	8640	7.17
2	204	22	130	6.15	34	104	62	350	6.60
3	202	16	310	7.00	35	287	31	230	7.00
4	197	65	570	6.25	36	41	66	1620	3.91
5	96	76	2050	3.81	37	312	11	190	6.70
6	209	26	200	6.44	38	77	88	2090	4.20
7	170	45	670	6.19	39	142	22	900	5.43
8	240	29	300	5.89	40	262	22	230	6.50
9	241	11	120	5.89	41	215	12	140	6.25
10	55	55	290	2.36	42	246	9	330	7.10
11	75	87	1180	3.93	43	191	31	1010	7.10
12	129	55	900	5.99	44	182	19	300	7.00
13	24	93	1730	3.50	45	37	88	1730	3.46
14	165	31	1150	7.41	46	103	35	780	5.66
15	94	77	1160	4.21	47	67	85	1300	4.82
16	96	80	1270	5.00	48	143	78	930	5.00
17	148	30	580	5.27	49	83	85	690	4.74
18	98	69	660	5.21	50	223	33	200	8.49
19	161	43	420	6.50	51	240	19	450	6.50
20	118	47	1080	6.12	52	312	21	280	6.50
21	269	17	290	6.19	53	12	79	4430	1.69
22	189	35	270	5.05	54	52	83	270	3.25
23	126	58	560	6.16	55	79	43	1340	7.17
24	12	81	4240	1.80	56	61	88	670	3.52
25	167	29	240	4.75	57	168	28	410	6.09
26	135	65	430	4.10	58	28	95	4370	2.86
27	107	87	3020	6.66	59	121	41	1310	4.88
28	72	63	1420	7.28	60	115	62	1470	3.89
29	128	49	420	8.12	61	186	45	300	6.90
30	27	63	19830	5.23	62	47	85	3630	4.10
31	152	84	420	5.79	63	178	45	220	6.09
32	224	23	530	6.50	64	142	67	560	7.20

Note: CM = Child mortality, the number of deaths of children under age 5 in a year per 1000 live births.
FLFP = Female literacy rate, percent.
PGNP = per capita GNP in 1980.
TFR = total fertility rate, 1980–1985, the average number of children born to a woman, using age-specific fertility rates for a given year.
Source: Chandan Mukherjee, Howard White, and Marc Whyte, *Econometrics and Data Analysis for Developing Countries*, Routledge, London, 1998, p. 456.

The coefficient -2.2316 tells us that holding the influence of PGNP constant, on average, the number of deaths of children under 5 goes down by about 2.23 per thousand live births as the female literacy rate increases by one percentage point. The intercept value of about 263, mechanically interpreted, means that if the values of PGNP and FLR rate were fixed at zero, the mean child mortality would be about 263 deaths per thousand live births. Of course, such an interpretation should be taken with a grain of salt. All one could infer is that if the two regressors were fixed at zero, child mortality will be quite high, which makes practical sense. The R^2 value of about 0.71 means that about 71 percent of the variation in child mortality is explained by PGNP and FLR, a fairly high value considering that the maximum value of R^2 can at most be 1. All told, the regression results make sense.

EXAMPLE 3.2

Hourly wage education

Using the 526 observations on workers in WAGE1.RAW, we include *educ* (years of education), *exper* (years of labor market experience), and *tenure* (years with the current employer) in an equation explaining $\log(\text{wage})$. The estimated equation is

$$\log(\hat{\text{wage}}) = 0.284 + 0.092\text{educ} + 0.0041\text{exper} + 0.022\text{tenure} \quad (3.17)$$

As in the simple regression case, the coefficients have a percentage interpretation. The only difference here is that they also have a *ceteris paribus* interpretation. The coefficient .092 means that, holding *exper* and *tenure* fixed, another year of education is predicted to increase $\log(\text{wage})$ by .092, which translates into an approximate 9.2% [$100(.092)$] increase in *wage*. Alternatively, if we take two people with the same levels of experience and job tenure, the coefficient on *educ* is the proportionate difference in predicted wage when their education levels differ by one year. This measure of the return to education at least keeps two important productivity factors fixed; whether it is a good estimate of the *ceteris paribus* return to another year of education requires us to study the statistical properties of OLS (see Section 3.3).

Changing more than one independent variable simultaneously

Sometimes we want to change more than one independent variable at the same time to find the resulting effect on the dependent variable. For example, in equation (3.19), we can obtain the estimated effect on *wage* when an individual stays at the same firm for another year: *exper* (general workforce experience) and *tenure* both increase by one year.

The total effect (holding *educ* fixed) is

$$\begin{aligned}\Delta \log(\hat{wage}) &= 0.0041\Delta \text{exper} + 0.022\Delta \text{tenure} = \\ &= 0.0041 + 0.022 = 0.0261\end{aligned}$$

or about 2.6%. Since *exper* and *tenure* each increase by one year, we just add the coefficients on *exper* and *tenure* and multiply by 100 to turn the effect into a percent.

OLS Fitted Values and Residuals

After obtaining the OLS regression line (3.11), we can obtain a *fitted* or *predicted value* for each observation. For observation *i*, the fitted value is simply

$$y = \hat{a} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_k x_{ik} \quad (3.18)$$

which is just the predicted value obtained by plugging the values of the independent variables for observation *i* into equation (3.11). We should not forget about the intercept in obtaining the fitted values; otherwise, the answer can be very misleading. As an example, if in (3.16), $PGNP_i=3.5$, and $FLR_i=22$,

$$CM_i = 263.6416 - 0.0056 \times 3.5 - 2.2316 \times 22 = 214.5268.$$

Normally, the actual value y_i for any observation *i* will not equal the predicted value, $\hat{y}_i$: OLS minimizes the *average* squared prediction error, which says nothing about the prediction error for

any particular observation. The **residual** for observation i is defined just as in the simple regression case,

$$\hat{u}_i = y_i - \hat{y}_i \quad (3.19)$$

There is a residual for each observation. If $\hat{u}_i > 0$, then $\hat{y}_i$ is below y_i , which means that, for this observation, y_i is underpredicted. If $\hat{u}_i < 0$, then $y_i < \hat{y}_i$, and y_i is overpredicted.

Box 3.1 The OLS properties

The OLS fitted values and residuals have some important properties that are immediate extensions from the single variable case:

1. The sample average of the residuals is zero.
2. The sample covariance between each independent variable and the OLS residuals is zero. Consequently, the sample covariance between the OLS fitted values and the OLS residuals is zero.
3. The point $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k, \bar{y})$ is always on the OLS regression line:

$$\bar{y} = \hat{\alpha} + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \dots + \hat{\beta}_k \bar{x}_k.$$

4. In the sample (and therefore in the population), none of the independent variables is constant, and there are no *exact linear* relationships among the independent variables.

The first two properties are immediate consequences of the set of equations used to obtain the OLS estimates. The first equation in (3.13) says that the sum of the residuals is zero. The remaining equations are of the form $\sum_{i=1}^n x_{ij} \hat{u}_i = 0$, which imply that the each independent variable has zero sample covariance with $\hat{u}_i$. Property 3 follows immediately from Property 1.

Goodness-of-fit statistics

Coefficient of determination R^2

It is desirable to have some measure of how well the regression model actually fits the data. In other words, it is desirable to have an answer to the question, ‘how well does the model containing the explanatory variables that was proposed actually explain variations in the dependent variable?’ Quantities known as *goodness of fit statistics* are available to test how well the sample regression function (SRF) fits the data - that is, how ‘close’ the fitted regression line is to all of the data points taken together. Note that it is not possible to say how well the sample regression function fits the population regression function - i.e. how the estimated model compares with the true relationship between the variables, since the latter is never known.

But what measures might make plausible candidates to be goodness of fit statistics? A first response to this might be to look at the residual sum of squares (*RSS*). Recall that OLS selected the coefficient estimates that minimized this quantity, so the lower was the minimized value of the *RSS*, the better the model fitted the data. Consideration of the *RSS* is certainly one possibility, but *RSS* is unbounded from above (strictly, *RSS* is bounded from above by the total sum of squares - see below) - i.e. it can take any (non-negative) value. So, for example, if the value of the *RSS* under OLS estimation was 136.4, what does this actually mean? It would therefore be very difficult, by looking at this number alone, to tell whether the regression line fitted the data closely or not. The value of *RSS* depends to a great extent on the scale of the dependent variable. Thus, one way to pointlessly reduce the *RSS* would be to divide all of the observations on y by 10!

In fact, a *scaled version* of the residual sum of squares is usually employed. The most common goodness of fit statistic is known as R^2 . One way to define R^2 is to say that it is the square of the correlation coefficient between y and $\hat{y}$ - that is, the square of the correlation between the values of the dependent variable and the corresponding fitted values from the model. A correlation coefficient must lie between -1 and +1 by definition. Since R^2 defined in this way is the square of a correlation coefficient, it must

lie between 0 and 1. If this correlation is high, the model fits the data well, while if the correlation is low (close to zero), the model is not providing a good fit to the data.

Another definition of R^2 requires a consideration of what the model is attempting to explain. What the model is trying to do in effect is to explain variability of y about its mean value, $\bar{y}$. This quantity, $\bar{y}$, which is more specifically known as the unconditional mean of y , acts like a benchmark since, if the researcher had no model for y , he could do no worse than to regress y on a constant only. In fact, the coefficient estimate for this regression would be the mean of y . So, from the regression

$$y_t = \beta_1 + u_t \quad (3.20)$$

the coefficient estimate $\hat{\beta}_1$, will be the mean of y , i.e. $\bar{y}$. The total variation across all observations of the dependent variable about its mean value is known as the total sum of squares, TSS , which is given by:

$$TSS = \sum_t (y_t - \bar{y})^2 \quad (3.21)$$

The TSS can be split into two parts: the part that has been explained by the model (known as the explained sum of squares, ESS) and the part that the model was not able to explain (the RSS). That is

$$TSS = ESS + RSS \quad (3.22)$$

$$\sum_t (y_t - \bar{y})^2 = \sum_t (\hat{y}_t - \bar{y})^2 + \sum_t \hat{u}_t^2 \quad (3.23)$$

Recall also that the residual sum of squares can also be expressed as

$$\sum_t (y_t - \hat{y}_t)^2$$

since a residual for observation t is defined as the difference between the actual and fitted values for that observation. The goodness of fit statistic is given by the ratio of the explained sum of squares to the total sum of squares:

$$R^2 = \frac{ESS}{TSS} \quad (3.24)$$

but since $TSS = ESS + RSS$, it is also possible to write

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \quad (3.25)$$

and it is interpreted as the proportion of the sample variation in y_i that is explained by the OLS regression line.

R^2 can also be shown to equal the squared correlation coefficient between the actual y_i and the fitted values $\hat{y}_i$. That is,

$$R^2 = \frac{\left(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}) \right)^2}{\left(\sum_{i=1}^n (y_i - \bar{y})^2 \right) \left(\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 \right)} \quad (3.25a)$$

By definition, R^2 is a number between zero and one (provided that there is a constant term in the regression). This is intuitive from the correlation interpretation of R^2 given above, but for another explanation, consider two extreme cases

$$\begin{aligned} RSS &= TSS \text{ i.e. } ESS = 0 \text{ so } R^2 = ESS/TSS = 0 \\ ESS &= TSS \text{ i.e. } RSS = 0 \text{ so } R^2 = ESS/TSS = 1 \end{aligned}$$

In the first case, the model has not succeeded in explaining any of the variability of y about its mean value, and hence the residual and total sums of squares are equal. This would happen only where the estimated values of all of the coefficients were exactly zero. In the second case, the model has explained all of the variability of y about its mean value, which implies that the residual sum of squares

will be zero. This would happen only in the case where all of the observation points lie exactly on the fitted line. Neither of these two extremes is likely in practice, of course, but they do show that R^2 is bounded to lie between zero and one, with a higher R^2 implying, everything else being equal, that the model fits the data better.

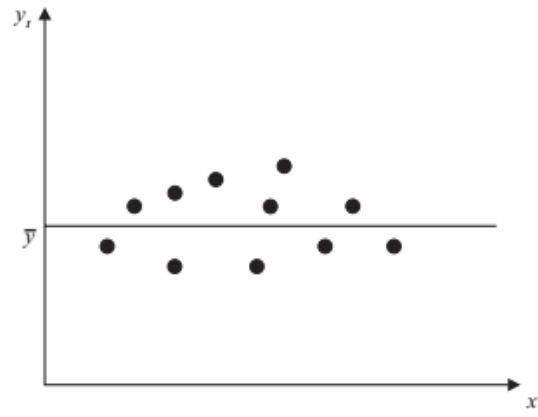


Figure 3.1 $R^2=0$ demonstrated by a flat estimated line, i.e. a zero slope coefficient

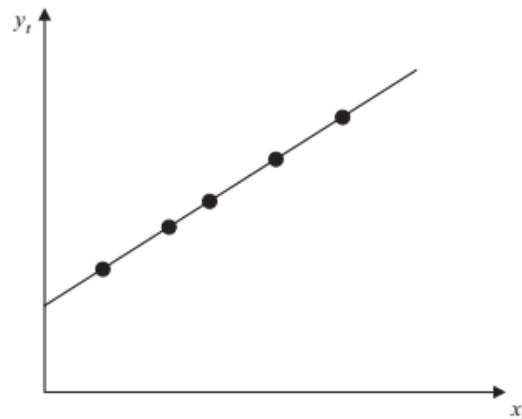


Figure 3.2 $R^2=1$ when all data points lie exactly on the estimated line

To sum up, a simple way (but crude, as explained next) to tell whether the regression line fits the data well is to look at the value of R^2 . A value of R^2 close to 1 indicates that the model explains nearly all of the variability of the dependent variable about its mean value, while a value close to zero indicates that the model fits the data poorly. The two extreme cases, where $R^2 = 0$ and $R^2 = 1$, are

indicated in Figures 3.1 and 3.2 in the context of a simple bivariate regression.

Exercises with R^2 as a goodness of fit measure

R^2 is simple to calculate, intuitive to understand, and provides a broad indication of the fit of the model to the data. However, there are a number of exercises with R^2 as a goodness of fit measure:

(1) R^2 is defined in terms of variation about the mean of y so that if a model is reparameterised (rearranged) and the dependent variable changes, R^2 will change, even if the second model was a simple rearrangement of the first, with identical RSS . Thus it is not sensible to compare the value of R^2 across models with different dependent variables.

(2) R^2 never falls if more regressors are added to the regression. For example, consider the following two models:

$$\text{Regression 1: } y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u \quad (3.41)$$

$$\text{Regression 2: } y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u \quad (3.42)$$

R^2 will always be at least as high for regression 2 relative to regression 1. The R^2 from regression 2 would be exactly the same as that for regression 1 only if the estimated value of the coefficient on the new variable were exactly zero, i.e. $\hat{\beta}_4 = 0$. In practice, $\hat{\beta}_4$ will always be nonzero, even if not significantly so, and thus in practice R^2 always rises as more variables are added to a model. This feature of R^2 essentially makes it impossible to use as a determinant of whether a given variable should be present in the model or not.

(3) R^2 can take values of 0.9 or higher for time series regressions, and hence it is not good at discriminating between models, since a wide array of models will frequently have broadly similar (and high) values of R^2 .

Adjusted R^2

In order to get around the second of these three exercises, a modification to R^2 is often made which takes into account the loss of degrees of freedom associated with adding extra variables. This is known as $\bar{R}^2$, or adjusted R^2 , which is defined as

$$\bar{R}^2 = 1 - \left[\frac{T-1}{T-k} (1 - R^2) \right] \quad (3.26)$$

So if an extra regressor (variable) is added to the model, k increases and unless R^2 increases by a more than off-setting amount, $\bar{R}^2$ will actually fall. Hence $\bar{R}^2$ can be used as a decision-making tool for determining whether a given variable should be included in a regression model or not, with the rule being: include the variable if $\bar{R}^2$ rises and do not include it if $\bar{R}^2$ falls.

However, there are still exercises with the maximisation of $\bar{R}^2$ as criterion for model selection, and principal among these is that it is a ‘soft’ rule, implying that by following it, the researcher will typically end up with a large model, containing a lot of marginally significant or insignificant variables. Also, while R^2 must be at least zero if an intercept is included in the regression, its adjusted counterpart may take negative values, even with an intercept in the regression, if the model fits the data very poorly.

Now reconsider the results from the previous exercises using Eviews in the previous chapter and earlier in this chapter. If we first consider the hedging model from chapter 2, the R^2 value for the returns regression was only 0.01, indicating that a mere 1% of the variation in spot returns is explained by the futures returns - a very poor model fit indeed.

The fit is no better for the Ford stock *CAPM* regression described in chapter 2, where the R^2 is less than 1% and the adjusted R^2 is actually negative. The conclusion here would be that for this stock and this sample period, almost none of the monthly movement in the excess returns can be attributed to movements in the market as a whole, as measured by the S&P500.

Finally, if we look at the results from the recent regressions for Microsoft, we find a considerably better fit. It is of interest to compare the model fit for the original regression that included all of the variables with the results of the stepwise procedure. We can see that the raw R^2 is slightly higher for the original regression (0.204 versus 0.200 for the stepwise regression, to three decimal places), exactly as we would expect. Since the original regression contains more variables, the R^2 -value must be at least as high. But comparing the $\bar{R}^2$ s, the stepwise regression value (0.187) is slightly higher than for the full regression (0.181), indicating that the additional regressors in the full regression do not justify their presence, at least according to this criterion.

Regression through the origin

Sometimes, an economic theory or common sense suggests that 0 should be zero, and so we should briefly mention OLS estimation when the intercept is zero. Specifically, we now seek an equation of the form

$$\tilde{y} = \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \dots + \tilde{\beta}_k x_k \quad (3.31)$$

where the symbol “ $\sim$ ” over the estimates is used to distinguish them from the OLS estimates obtained along with the intercept [as in (3.11)]. In (3.31), when $x_1 = 0, x_2 = 0, \dots, x_k = 0$, the predicted value is zero. In this case, $\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_k$ are said to be the OLS estimates from the regression of y on $x_1, x_2, \dots, x_k$ *through the origin*.

The OLS estimates in (3.31), as always, minimize the sum of squared residuals, but with the intercept set at zero. You should be warned that the properties of OLS that we derived earlier no longer hold for regression through the origin. In particular, the OLS residuals no longer have a zero sample average. Further, if R^2 is defined as $1 - \text{RSS}/\text{TSS}$, where TSS is given in (3.22) and RSS is now $\sum_{i=1}^n (y_i - \tilde{\beta}_1 x_{i1} - \dots - \tilde{\beta}_k x_{ik})^2$, then R^2 can actually be negative. This means that the sample average, $\bar{y}$, “explains” more of the variation in the y_i than the explanatory variables. Either we should include an intercept in the regression or conclude that the explanatory variables poorly explain y . In order to always have a nonnegative R -squared, some economists prefer to calculate R^2 as the squared correlation coefficient between the actual and fitted values of y , as in (3.25a). (In this case, the average fitted value must be computed directly since it no longer equals to $\bar{y}$). However, there is no set rule on computing R -squared for regression through the origin.

One serious drawback with regression through the origin is that, if the intercept α in the population model is different from zero, then the OLS estimators of the slope parameters will be biased. The bias can be severe in some cases. The cost of estimating

an intercept when a is truly zero is that the variances of the OLS slope estimators are larger.

3.3 Comparison of simple and multiple regression estimates

Two special cases exist in which the simple regression of y on x_1 will produce the *same* OLS estimate on x_1 as the regression of y on x_1 and x_2 . To be more precise, write the simple regression of y on x_1 as $\tilde{y} = \tilde{a} + \tilde{\beta}x_1$ and write the multiple regression as $y = \hat{a} + \hat{\beta}_1x_1 + \hat{\beta}_2x_2$. We know that the simple regression coefficient $\tilde{a}$ does not usually equal the multiple regression coefficient $\hat{a}$. There are two distinct cases where $\tilde{a}$ and $\hat{a}$ are identical:

1. The partial effect of x_2 on y is zero in the sample. That is, $\hat{\beta}_2 = 0$.
2. x_1 and x_2 are uncorrelated in the sample.

The first assertion can be proven by looking at two of the equations used to determine $\hat{a}$, $\hat{\beta}_1$, and $\hat{\beta}_2$: $\sum_{i=1}^n x_{i1}(y_i - \hat{a} - \hat{\beta}_1x_{i1} - \hat{\beta}_2x_{i2}) = 0$ and $\hat{a} = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2$. Setting $\hat{\beta}_2 = 0$ gives the same intercept and slope as does the regression of y on x_1 . The second assertion follows from equation

$$\hat{\beta}_1 = \left(\sum_{i=1}^n \hat{r}_{i1}y_i \right) / \left(\sum_{i=1}^n \hat{r}_{i1}^2 \right) \quad (3.32)$$

If x_1 and x_2 are uncorrelated in the sample, then regressing x_1 on x_2 results in no partialling out, and so the simple regression of y on x_1 and the multiple regression of y on x_1 and x_2 produce identical estimates on x_1 .

Even though simple and multiple regression estimates are almost never identical, we can use the previous characterizations to explain why they might be either very different or quite similar. For example, if $\hat{\beta}_2$ is small, we might expect the simple and multiple regression estimates of β_1 to be similar.

EXAMPLE 3.3

Participation in 401(k) pension plans

We use the data in 401K.RAW to estimate the effect of a plan's match rate (*mrte*) on the participation rate (*prate*) in its 401(k) pension plan. The match rate is the amount the firm contributes to a worker's fund for each dollar the worker contributes (up to some limit); thus, *mrte*=0.75 means that the firm contributes 75 cents for each dollar contributed by the worker. The participation rate is the percentage of eligible workers having a 401(k) account. The variable *age* is the age of the 401(k) plan. There are 1,534 plans in the data set, the average *p-rate* is 87.36, the average *m-rate* is .732, and the average *age* is 13.2.

Regressing *prate* on *m-rate*, *age* gives

$$\hat{prate} = 80.12 + 5.52mrte + 0.243age \quad (3.33)$$

Thus, both *mrte* and *age* have the expected effects. What happens if we do not control for *age*? The estimated effect of *age* is not trivial, and so we might expect a large change in the estimated effect of *mrte* if *age* is dropped from the regression. However, the simple regression of *prate* on *mrte* yields $\hat{prate} = 83.08 + 5.86mrte$. The simple regression estimate of the effect of *mrte* on *prate* is clearly different from the multiple regression estimate, but the difference is not very big. (The simple regression estimate is only about 6.2 percent larger than the multiple regression estimate.) This can be explained by the fact that the sample correlation between *mrte* and *age* is only 0.12.

In the case with *k* independent variables, the simple regression of *y* on *x*₁ and the multiple regression of *y* on *x*₁, *x*₂, ..., *x*_{*k*} produce an identical estimate of *x*₁ only if (1) the OLS coefficients on *x*₂ through *x*_{*k*} are all zero or (2) *x*₁ is uncorrelated with *each of* *x*₂, ..., *x*_{*k*}. Neither of these is very likely in practice. But if the coefficients on *x*₂ through *x*_{*k*} are small, or the sample correlations between *x*₁ and the other independent variables are insubstantial, then the simple and multiple regression estimates of the effect of *x*₁ on *y* can be similar.

SUMMARY

1. The multiple regression model allows us to effectively hold other factors fixed while examining the effects of a particular independent variable on the dependent variable. It explicitly allows the independent variables to be correlated.
2. Although the model is linear in its *parameters*, it can be used to model nonlinear relationships by appropriately choosing the dependent and independent variables.
3. The method of ordinary least squares is easily applied to the multiple regression model. Each slope estimate measures the partial effect of the corresponding independent variable on the dependent variable, holding all other independent variables fixed.
4. R^2 is the proportion of the sample variation in the dependent variable explained by the independent variables, and it serves as a goodness-of-fit measure. It is important not to put too much weight on the value of R^2 when evaluating econometric models.

APPENDIX 3:

3A.1 Mathematical derivations of CLRM results

Derivation of the OLS coefficient estimator in the multiple regression context

In the multiple regression context, in order to obtain the parameter estimates, $\beta_1, \beta_2, \dots, \beta_k$, the RSS would be minimized with respect to all the elements of β . Now the residuals are expressed in a vector:

$$\hat{u} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_T \end{bmatrix} \quad (3A.1)$$

The RSS is still the relevant loss function, and would be given in a matrix notation by expression (3A.2)

$$L = \hat{u}'\hat{u} = [\hat{u}_1, \hat{u}_2, \dots, \hat{u}_T] \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_T \end{bmatrix} = \hat{u}_1^2 + \hat{u}_2^2 + \dots + \hat{u}_T^2 = \sum \hat{u}_t^2 \quad (3A.2)$$

Denoting the vector of estimated parameters as $\hat{\beta}$, it is also possible to write

$$L = \hat{u}'\hat{u} = (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \quad (3A.3)$$

It turns out that $\hat{\beta}_{xy}$ is $(1 \times k) \times (k \times T) \times (T \times 1) = 1 \times 1$, and also that $y'X\hat{\beta}$ is $(1 \times T) \times (T \times k) \times (k \times 1) = 1 \times 1$, so in fact $\hat{\beta}_{xy} = y'X\hat{\beta}$. Thus (3A.3) can be written

$$L = \hat{u}'\hat{u} = (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \quad (3A.4)$$

Differentiating this expression with respect to $\hat{\beta}$ and setting it to zero in order to find the parameter values that minimize the residual sum of squares would yield

$$\frac{\partial L}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0 \quad (3A.5)$$

This expression arises since the derivative of $y'y$ is zero with respect to $\hat{\beta}$, and $\hat{\beta}'X'X\hat{\beta}$ acts like a square of $X\hat{\beta}$, which is differentiated to $2X'X\hat{\beta}$. Rearranging (3A.5)

$$2X'y = 2X'X\hat{\beta} \quad (3A.6)$$

$$X'y = X'X\hat{\beta} \quad (3A.7)$$

Pre-multiplying both sides of (3A.7) by the inverse of $X'X$

$$\hat{\beta} = (X'X)^{-1} X'y \quad (3A.8)$$

Thus, the vector of OLS coefficient estimates for a set of k parameters is given by

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (X'X)^{-1} X'y \quad (3A.9)$$

Derivation of the OLS standard error estimator in the multiple regression context

The variance of a vector of random variables $\hat{\beta}$ is given by the formula $E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$. Since $y = X\beta + u$, it can also be stated, given (3A.9), that

$$\hat{\beta} = (X'X)^{-1} X'(X\beta + u) \quad (3A.10)$$

Expanding the parentheses

$$\hat{\beta} = (X'X)^{-1} X'X\beta + (X'X)^{-1} X'u \quad (3A.11)$$

$$\hat{\beta} = \beta + (X'X)^{-1} X'u \quad (3A.12)$$

Thus, it is possible to express the variance of $\hat{\beta}$ as

$$E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = E[(\beta + (X'X)^{-1} X'u - \beta)(\beta + (X'X)^{-1} X'u - \beta)'] \quad (3A.13)$$

Cancelling the β terms in each set of parentheses

$$E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = E[(X'X)^{-1} X'u(X'X)^{-1} X'u'] \quad (3A.14)$$

Expanding the parentheses on the RHS of (3A.14) gives

$$E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] = E\left[(X'X)^{-1} X'uu'X(X'X)^{-1}\right] \quad (3A.15)$$

$$E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] = (X'X)^{-1} X'E[uu']X(X'X)^{-1} \quad (3A.16)$$

Now $E[uu']$ is estimated by s^2I , so that

$$E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] = (X'X)^{-1} X's^2IX(X'X)^{-1} \quad (3A.17)$$

where I is a $k \times k$ identity matrix. Rearranging further,

$$E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] = s^2(X'X)^{-1} X'X(X'X)^{-1} \quad (3A.18)$$

The $X'X$ and the last $(X'X)^{-1}$ term cancel out to leave

$$\text{var}(\hat{\beta}) = s^2(X'X)^{-1} \quad (3A.19)$$

as the expression for the parameter variance--covariance matrix. This quantity, $s^2(X'X)^{-1}$, is known as the estimated variance-covariance matrix of the coefficients. The leading diagonal terms give the estimated coefficient variances while the off-diagonal terms give the estimated covariances between the parameter estimates. The variance of $\hat{\beta}_1$ is the first diagonal element, the variance of $\hat{\beta}_2$ is the second element on the leading diagonal, ..., and the variance of $\hat{\beta}_k$ is the k th diagonal element, etc. as discussed in the body of the chapter.

3A.2 Derivation of the first order conditions

The analysis is very similar to the simple regression case. We must characterize the solutions to the problem

$$a, b_1, b_2, \dots, b_k = \min \sum_{i=1}^n (y_i - a - b_1 x_{i1} - b_2 x_{i2} - \dots - b_k x_{ik})^2$$

Taking the partial derivatives with respect to each of the b_j (see Appendix A), evaluating them at the solutions, and setting them equal to zero gives

$$\begin{aligned} -2 \sum_{i=1}^n (y_i - \hat{a} - b_1 x_{i1} - \dots - b_k x_{ik}) &= 0 \\ -2 \sum_{i=1}^n x_{ij} (y_i - \hat{a} - b_1 x_{i1} - \dots - b_k x_{ik}) &= 0, j = 1, \dots, k \end{aligned}$$

Cancelling the -2 gives the first order conditions in (3.13).

3B. A brief introduction to factor models and principal components analysis.

Factor models are employed primarily as dimensionality reduction techniques in situations where we have a large number of closely related variables and where we wish to allow for the most important influences from all of these variables at the same time. Factor models decompose the structure of a set of series into factors that are common to all series and a proportion that is specific to each series (idiosyncratic variation). There are broadly two types of such models, which can be loosely characterized as either macroeconomic or mathematical factor models. The key distinction between the two is that the factors are observable for the former but are latent (unobservable) for the latter. Observable factor models include the APT model of Ross (1976). The most common mathematical factor model is principal components analysis (PCA). PCA is a technique that may be useful where explanatory variables are closely related - for example, in the context of near multicollinearity. Specifically, if there are k explanatory variables in the regression model, PCA will transform them into k uncorrelated new variables. To elucidate, suppose that the original explanatory variables are denoted $x_1, x_2, \dots, x_k$, and denote the principal components by $p_1, p_2, \dots, p_k$. These principal components are independent linear combinations of the original data

$$\begin{aligned}
 p_1 &= \alpha_{11}x_1 + \alpha_{12}x_2 + \dots + \alpha_{1k}x_k \\
 p_2 &= \alpha_{21}x_1 + \alpha_{22}x_2 + \dots + \alpha_{2k}x_k \\
 &\dots \quad \dots \quad \dots \quad \dots \\
 p_k &= \alpha_{k1}x_1 + \alpha_{k2}x_2 + \dots + \alpha_{kk}x_k
 \end{aligned}
 \tag{3B.1}$$

where α_{ij} are coefficients to be calculated, representing the coefficient on the j th explanatory variable in the i th principal component. These coefficients are also known as factor loadings. Note that there will be T observations on each principal component if there were T observations on each explanatory variable. It is also required that the sum of the squares of the coefficients for each component is one, i.e.

$$\begin{aligned}
 \alpha_{11}^2 + \alpha_{12}^2 + \dots + \alpha_{1k}^2 &= 1 \\
 \vdots \quad \quad \quad \vdots & \\
 \alpha_{k1}^2 + \alpha_{k2}^2 + \dots + \alpha_{kk}^2 &= 1
 \end{aligned}
 \tag{3B.2}$$

This requirement could also be expressed using sigma notation

$$\sum_{j=1}^k \alpha_{ij}^2 = 1 \quad \forall \quad i=1,2,\dots,k$$

Constructing the components is a purely mathematical exercise in constrained optimization, and thus no assumption is made concerning the structure, distribution, or other properties of the variables.

The principal components are derived in such a way that they are in descending order of importance. Although there are k principal components, the same as the number of explanatory variables, if there is some collinearity between these original explanatory variables, it is likely that some of the (last few) principal components will account for so little of the variation that they can be discarded. However, if all of the original explanatory variables were already essentially uncorrelated, all of the components would be required, although in such a case there would have been little motivation for using PCA in the first place.

The principal components can also be understood as the eigenvalues of $(X'X)$, where X is the matrix of observations on the original variables. Thus the number of eigenvalues will be equal to the number of variables, k . If the ordered eigenvalues are denoted λ_i ($i=1, \dots, k$), the ratio

$$\phi_i = \frac{\lambda_i}{\sum_{i=1}^k \lambda_i}$$

gives the proportion of the total variation in the original data explained by the principal component i . Suppose that only the first r ($0 < r < k$) principal components are deemed sufficiently useful in explaining the variation of (XX) , and that they are to be retained, with the remaining $k - r$ components being discarded. The regression finally estimated, after the principal components have been formed, would be one of y on the r principal components

$$y_t = \gamma_0 + \gamma_1 p_{1t} + \dots + \gamma_r p_{rt} + u_t \quad (3B.3)$$

In this way, the principal components are argued to keep most of the important information contained in the original explanatory variables, but are orthogonal. This may be particularly useful for independent variables that are very closely related. The principal component estimates $\hat{\gamma}_i, i=1,2,\dots,r$ will be biased estimates, although they will be more efficient than the OLS estimators since redundant information has been removed. In fact, if the OLS estimator for the original regression of y on x is denoted $\hat{\beta}$, it can be shown that

$$\hat{\gamma}_r = P_r' \hat{\beta} \quad (3B.4)$$

where $\hat{\gamma}_r$ are the coefficient estimates for the principal components, and P_r is a matrix of the first r principal components. The principal component coefficient estimates are thus simply linear combinations of the original OLS estimates.

An application of principal components to interest rates

Many economic and financial models make use of interest rates in some form or another as independent variables. Researchers may wish to include interest rates on a large number of different assets in order to reflect the variety of investment opportunities open to investors. However, market interest rates could be argued to be not sufficiently independent of one another to make the inclusion of several interest rate series in an econometric model statistically sensible. One approach to examining this issue would be to use PCA on several related interest rate series to determine whether they did move independently of one another over some historical time period or not. Fase (1973) conducted such a study in the context of monthly Dutch market interest rates from January 1962 until December 1970 (108 months). Fase examined both ‘money market’ and ‘capital market’ rates, although only the money market results will be discussed here in the interests of brevity. The money market instruments investigated were:

- Call money
- Three-month Treasury paper
- One-year Treasury paper
- Two-year Treasury paper
- Three-year Treasury paper
- Five-year Treasury paper
- Loans to local authorities: three-month
- Loans to local authorities: one-year
- Eurodollar deposits
- Netherlands Bank official discount rate.

Prior to analysis, each series was standardized to have zero mean and unit variance by subtracting the mean and dividing by the standard deviation in each case. The three largest of the ten eigenvalues are given in Table 3B.1.

The results in Table 3A.1 are presented for the whole period using the monthly data, for two monthly sub-samples, and for the whole period using data sampled quarterly instead of monthly. The results show clearly that the first principal component is sufficient to describe the common variation in these Dutch interest rate series.

The first component is able to explain over 90% of the variation in all four cases, as given in the last row of Table 3A.1. Clearly, the estimated eigenvalues are fairly stable across the sample periods and are relatively invariant to the frequency of sampling of the data. The factor loadings (coefficient estimates) for the first two ordered components are given in Table 3A.2.

Table 3B.1

Principal component ordered eigenvalues for Dutch interest rates, 1962–1970

	Monthly data			Quarterly data
	Jan 62–Dec 70	Jan 62–Jun 66	Jul 66–Dec 70	Jan 62–Dec 70
λ_1	9.57	9.31	9.32	9.67
λ_2	0.20	0.31	0.40	0.16
λ_3	0.09	0.20	0.17	0.07
ϕ_1	95.7%	93.1%	93.2%	96.7%

Source: Fase (1973). Reprinted with the permission of Elsevier Science.

Table 3B.2

Factor loadings of the first and second principal components for Dutch interest rates, 1962–1970

j	Debt instrument	α_{j1}	α_{j2}
1	Call money	0.95	−0.22
2	3-month Treasury paper	0.98	0.12
3	1-year Treasury paper	0.99	0.15
4	2-year Treasury paper	0.99	0.13
5	3-year Treasury paper	0.99	0.11
6	5-year Treasury paper	0.99	0.09
7	Loans to local authorities: 3-month	0.99	−0.08
8	Loans to local authorities: 1-year	0.99	−0.04
9	Eurodollar deposits	0.96	−0.26
10	Netherlands Bank official discount rate	0.96	−0.03
	Eigenvalue, λ_i	9.57	0.20
	Proportion of variability explained by eigenvalue i , $\phi_i(\%)$	95.7	2.0

Source: Fase (1973). Reprinted with the permission of Elsevier Science.

As Table 3A.2 shows, the loadings on each factor making up the first principal component are all positive. Since each series has been standardized to have zero mean and unit variance, the coefficients a_{j1} and a_{j2} can be interpreted as the correlations between the interest rate j and the first and second principal components, respectively. The factor loadings for each interest rate series on the first component are all very close to one. Fase (1973) therefore argues that the first component can be interpreted simply as an equally weighted combination of all of the market interest rates. The second component, which explains much less of the variability of the rates, shows a factor loading pattern of positive coefficients for the Treasury paper series and negative or almost zero values for the other series. Fase (1973) argues that this is owing to the characteristics of the Dutch Treasury instruments that they rarely change hands and have low transactions costs, and therefore have less sensitivity to general interest rate movements. Also, they are not subject to default risks in the same way as, for example Eurodollar deposits. Therefore, the second principal component is broadly interpreted as relating to default risk and transactions costs. Principal components can be useful in some circumstances, although the technique has limited applicability for the following reasons:

- A change in the units of measurement of x will change the principal components. It is thus usual to transform all of the variables to have zero mean and unit variance prior to applying PCA.
- The principal components usually have no theoretical motivation or interpretation whatsoever.
- The r principal components retained from the original k are the ones that explain most of the variation in x , but these components might not be the most useful as explanations for y .

CHAPTER 4. LEARNING TO USE REGRESSION ANALYSIS

4.1 Steps in applied regression analysis

Although there are no hard and fast rules for conducting econometric research, most investigators commonly follow a standard method for applied regression analysis. The relative emphasis and effort expended on each step will vary, but normally all the steps are necessary for successful research. Note that we don't discuss the selection of the dependent variable; this choice is determined by the purpose of the research. We'll cover choosing a dependent variable in Chapter 12. Once a dependent variable is chosen, however, it's logical to follow these **six steps in applied regression analysis**.

1. Review the literature and develop the theoretical model.
2. Specify the model: Select the independent variables and the functional form.
3. Hypothesize the expected signs of the coefficients.
4. Collect the data. Inspect and clean the data.
5. Estimate and evaluate the equation.
6. Document the results.

The purpose of suggesting these steps is not to discourage the use of innovative or unusual approaches but rather to develop in the reader a sense of how regression ordinarily is done by professional economists and business analysts.

Step 1: Review the literature and develop the theoretical model

The first step in any applied research is to get a good theoretical grasp of the topic to be studied. That's right: the best data analysts don't start with data, but with theory! This is because many econometric decisions, ranging from which variables to include to which functional form to employ, are determined by the underlying theoretical model. It's virtually impossible to build a good econometric model without a solid understanding of the topic you're studying.

For most topics, this means that it's smart to review the scholarly literature before doing anything else. If a professor has investigated the theory behind your topic, you want to know about it. If other researchers have estimated equations for your dependent variable, you might want to apply one of their models to your data set. On the other hand, if you disagree with the approach of previous authors, you might want to head off in a new direction. In either case, you shouldn't have to "reinvent the wheel." You should start your investigation where earlier researchers left off. Any academic paper on an empirical topic should begin with a summary of the extent and quality of previous research.

The most convenient approaches to reviewing the literature are to obtain several recent issues of the *Journal of Economic Literature* or a business-oriented publication of abstracts, or to run an Internet search or an *EconLit* search on your topic. Using these resources, find and read several recent articles on your topic. Pay attention to the bibliographies of these articles. If an older article is cited by a number of current authors, or if its title hits your topic on the head, trace back through the literature and find this article as well.

In some cases, a topic will be so new or so obscure that you won't be able to find any articles on it. What then? We recommend two possible strategies. First, try to transfer theory from a similar topic to yours. For example, if you're trying to build a model of the demand for a new product, read articles that analyze the demand for similar, existing products. Second, if all else fails, contact someone who works in the field you're investigating. For example, if you're building a model of housing in an unfamiliar city, call a real estate agent who works there.

Step 2: Specify the model: select the independent variables and the functional form

The most important step in applied regression analysis is the specification of the theoretical regression model. After selecting the dependent variable, the **specification** of a model involves choosing the following components:

1. the independent variables and how they should be measured,
2. the functional (mathematical) form of the variables, and
3. the properties of the stochastic error term.

A regression equation is specified when each of these elements has been treated appropriately. We'll discuss the details of these specification decisions in Chapters 7, 8, and 5, respectively. Each of the elements of specification is determined primarily on the basis of economic theory. A mistake in any of the three elements results in a **specification error**. Of all the kinds of mistakes that can be made in applied regression analysis, specification error is usually the most disastrous to the validity of the estimated equation. Thus, the more attention paid to economic theory at the beginning of a project, the more satisfying the regression results are likely to be.

The emphasis in this text is on estimating behavioral equations, those that describe the behavior of economic entities. We focus on selecting independent variables based on the economic theory concerning that behavior. An explanatory variable is chosen because it is a theoretical determinant of the dependent variable; it is expected to explain at least part of the variation in the dependent variable. Recall that regression gives evidence but does not prove economic causality. Just as an example does not prove the rule, a regression result does not prove the theory.

There are dangers in specifying the wrong independent variables. Our goal should be to specify only relevant explanatory variables, those expected theoretically to assert a substantive influence on the dependent variable. Variables suspected of having little effect should be excluded unless their possible impact on the dependent variable is of some particular (e.g., policy) interest.

For example, an equation that explains the quantity demanded of a consumption good might use the price of the product and consumer income or wealth as likely variables. Theory also indicates that complementary and substitute goods are important. Therefore, you might decide to include the prices of complements and substitutes, but which complements and substitutes? Of course, selection of the closest complements and/or substitutes is appropriate, but how far should you go? The choice must be based

on theoretical judgment, and such judgments are often quite subjective.

When researchers decide, for example, that the prices of only two other goods need to be included, they are said to impose their *priors* (i.e., previous theoretical belief) or their working hypotheses on the regression equation. Imposition of such priors is a common practice that determines the number and kind of hypotheses that the regression equation has to test. The danger is that a prior may be wrong and could diminish the usefulness of the estimated regression equation. Each of the priors therefore should be explained and justified in detail.

Step 3: Hypothesize the expected signs of the coefficients

Once the variables have been selected, it's important to hypothesize the expected signs of the slope coefficients before you collect any data. In many cases, the basic theory is general knowledge, so you don't need to discuss the reasons for the expected sign. However, if any doubt surrounds the choice of an expected sign, then you should document the opposing theories and your reasons for hypothesizing a positive or a negative slope³ coefficient. For example, suppose that you're interested in the impact of class size on student learning at the elementary level in the United States. A reasonable dependent variable (y) might be the student score on a test of grammar, math, and science. Likely independent variables would include the income level of the student's family (x_1) and the size (in students per teacher) of the student's class (x_2).

$$y = a + \beta_1^+ x_1 + \beta_2^- x_2 + u \quad (4.1)$$

The signs above the coefficients in Equation 3.1 indicate the hypothesized sign of that particular coefficient. Take another look at the equation. Do you agree with the hypothesized signs? The

³ Note that while we hypothesize signs for the slope coefficients, we don't hypothesize an expected sign for the intercept. We'll explain why in Section 6.1.

expectation that higher income will improve test scores (holding constant class size) seems reasonable because of the extra learning opportunities that the money might allow, but the hypothesized sign for β_2 is a little trickier.

Step 4: Collect the data. Inspect and clean the data

Obtaining an original data set and properly preparing it for regression is a surprisingly difficult task. This step entails more than a mechanical recording of data, because the type and size of the sample also must be chosen.

A general rule regarding sample size is “the more observations the better,” as long as the observations are from the same general population. Ordinarily, researchers take all the roughly comparable observations that are readily available. In regression analysis, all the variables must have the same number of observations. They also should have the same frequency (monthly, quarterly, annual, etc.) and time period. Often, the frequency selected is determined by the availability of data.

The reason there should be as many observations as possible concerns the statistical concept of *degrees of freedom* first mentioned in Section 2.4. Consider fitting a straight line to two points on an x, y coordinate system as in Figure 4.1 below. Such an exercise can be done mathematically without error. Both points lie on the line, so there is no estimation of the coefficients involved. The two points determine the two parameters, the intercept and the slope, precisely. Estimation takes place only when a straight line is fitted to constant term and measures of overall fit remain unchanged. Such a multiplicative factor does change the slope coefficient, but only by the exact amount necessary to compensate for the change in the units of measurement of the independent variable. Similarly, a constant factor added to a variable alters only the intercept term without changing the slope coefficient itself.

The final step before estimating your equation is to inspect and clean the data. You should make it a point always to look over your data set to see if you can find any errors. The reason is obvious: why

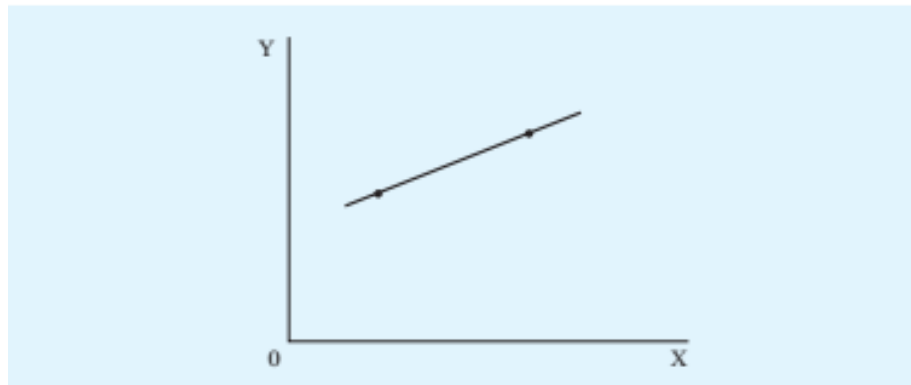


Figure 4.1 Mathematical fit of a line to two points

If there are only two points in a data set, as in Figure 3.1, a straight line can be fitted to those points mathematically without error, because two points completely determine a straight line.

bother using sophisticated regression analysis if your data are incorrect?

To inspect the data, obtain a plot (graph) of the data and look for outliers. An *outlier* is an observation that lies outside the range of the rest of the observations, and looking for outliers is an easy way to find data entry errors. In addition, it's a good habit to look at the mean, maximum, and minimum of each variable and then think about possible inconsistencies in the data. Are any observations impossible or unrealistic? Did GDP double in one year? Does a student have a 7.0 GPA on a 4.0 scale? Is consumption negative?

Typically, the data can be cleaned of these errors by replacing an incorrect number with the correct one. In extremely rare circumstances, an observation can be dropped from the sample, but only if the correct number can't be found or if that particular observation clearly isn't from the same population as the rest of the sample. Be careful! The mere existence of an outlier is not a justification for dropping that observation from the sample. A regression needs to be able to explain all the observations in a sample, not just the well-behaved ones.

Step 5: Estimate and evaluate the equation

Believe it or not, it can take months to complete steps 1–4 for a regression equation, but a computer program like Stata or Eviews can estimate that equation in less than a second! Typically, estimation is done using OLS, as discussed in Section 2.1, but if another estimation technique is used, the reasons for that alternative technique should be carefully explained and evaluated.

You might think that once your equation has been estimated, your work is finished, but that's hardly the case. Instead, you need to evaluate your results in a variety of ways. How well did the equation fit the data? Were the signs and magnitudes of the estimated coefficients what you expected? Most of the rest of this book is concerned with the evaluation of estimated econometric equations, and beginning researchers should be prepared to spend a considerable amount of time doing this evaluation.

Once this evaluation is complete, don't automatically go to step 6. Regression results are rarely what one expects, and additional model development often is required. For example, an evaluation of your results might indicate that your equation is missing an important variable. In such a case, you'd go back to step 1 to review the literature and add the appropriate variable to your equation. You'd then go through each of the steps in order until you had estimated your new specification in step 5. You'd move on to step 6 only if you were satisfied with your estimated equation. Don't be too quick to make such adjustments, however, because we don't want to adjust the theory merely to fit the data. A researcher has to walk a fine line between making appropriate changes and avoiding inappropriate ones, and making these choices is one of the artistic elements of applied econometrics.

Finally, it's often worthwhile to estimate additional specifications of an equation in order to see how stable your observed results are. This approach, called *sensitivity analysis*, will be discussed in Section 7.4.

Step 6: Document the results

A standard format usually is used to present estimated regression results:

$$\begin{aligned}\hat{y}_i &= 103.40 + 6.38x_i \\ &\quad (0.88) \\ t &= 7.22 \\ N &= 20, \quad \bar{R}^2 = 0.73\end{aligned}\tag{4.2}$$

The number in parentheses is the estimated standard error of the estimated coefficient, and the t -value is the one used to test the hypothesis that the true value of the coefficient is different from zero. These and other measures of the quality of the regression will be discussed in later chapters.⁴ What is important to note is that the documentation of regression results using an easily understood format is considered part of the analysis itself. For time-series data sets, the documentation also includes the frequency (e.g., quarterly or annual) and the time period of the data.

One of the important parts of the documentation is the explanation of the model, the assumptions, and the procedures and data used. The written documentation must contain enough information so that the entire study could be replicated by others.⁵ Unless the variables have been defined in a glossary or Table, short definitions should be presented along with the equations. If there is a series of estimated regression equations, then Tables should provide the relevant information for each equation. All data manipulations as well as data sources should be documented fully. When there is much to explain, this documentation usually is relegated to a data appendix. If the data are not available generally or are available only after computation, the data set itself might be included in this appendix.

⁴ The standard error of the coefficient is discussed in more detail in Section 5.2; the t -value is developed in Section 5.2.

⁵ For example, the *Journal of Money, Credit, and Banking* and the *American Economic Review* have requested authors to submit their actual data sets so that regression results can be verified. See W. G. Dewald et al., "Replication in Empirical Economics," *American Economic Review*, Vol. 76, No. 4, pp. 587–603 and Daniel S. Hamermesh, "Replication in Economics," NBER Working Paper 13026, April 2007.

4.2 Using regression analysis to pick restaurant locations

To solidify your understanding of the six basic steps of applied regression analysis, let's work through a complete regression example. Suppose that you've been hired to determine the best location for the next Rohat's restaurant, where Rohat's is a moderately priced, 24-hour, family restaurant chain. You decide to build a regression model to explain the gross sales volume at each of the restaurants in the chain as a function of various descriptors of the location of that branch. If you can come up with a sound equation to explain gross sales as a function of location, then you can use this equation to help Rohat's decide where to build their newest eatery. Given data on land costs, building costs, and local building and restaurant municipal codes, the owners of Rohat's will be able to make an informed decision.

1. Review the literature and develop the theoretical model.

You do some reading about the restaurant industry, but your review of the literature consists mainly of talking to various experts within the firm. They give you some good ideas about the attributes of a successful Rohat's location. The experts tell you that all of the chain's restaurants are identical (indeed, this is sometimes a criticism of the chain) and that all the locations are in what might be called "suburban, retail, or residential" environments (as distinguished from central cities or rural areas, for example). Because of this, you realize that many of the reasons that might help explain differences in sales volume in other chains do not apply in this case because all the Rohat's locations are similar. (If you were comparing Rohat's to another chain, such variables might be appropriate.) In addition, discussions with the people in the Rohat's strategic planning department convince you that price differentials and consumption differences between locations are not as important as the number of customers a particular location attracts. This causes you concern for a while because the variable you had planned to study originally, gross sales volume, would vary as prices changed between locations. Since your company controls these prices, you feel that you would rather have an estimate of the "potential" for such sales. As a result, you decide to specify your dependent variable as the number of customers served (measured by the number of checks or bills that the servers handed out) in a given

location in the most recent year for which complete data are available.

2. Specify the model: Select the independent variables and the functional form.

Your discussions lead to a number of suggested variables. After a while, you realize that there are three major determinants of sales (customers) on which virtually everyone agrees. These are the number of people who live near the location, the general income level of the location, and the number of direct competitors close to the location. In addition, there are two other good suggestions for potential explanatory variables. These are the number of cars passing the location per day and the number of months that the particular restaurant has been open.

After some serious consideration of your alternatives, you decide not to include the last possibilities. All the locations have been open long enough to have achieved a stable clientele. In addition, it would be very expensive to collect data on the number of passing cars for all the locations. Should population prove to be a poor measure of the available customers in a location, you'll have to decide whether to ask your boss for the money to collect complete traffic data.

The exact definitions of the independent variables you decide to include are:

N = Competition: the number of direct market competitors within a two-mile radius of the Rohat's location,

P = Population: the number of people living within a three-mile radius of the Rohat's location,

I = Income: the average household income of the population measured in variable P .

Since we have yet to develop any functional forms other than a linear functional form and a typical stochastic error term, that's what you decide to use.

3. Hypothesize the expected signs of the coefficients.

After thinking about which variables to include, you expect hypothesizing signs will be easy. For two of the variables, you're right. Everyone expects that the more competition there is, the

fewer customers (holding constant the population and income of an area) there will be, and also that the more people there are who live near a particular restaurant, the more customers (holding constant the competition and income) the restaurant will have. You expect that the greater the income is in a particular area, the more people will choose to eat in a family restaurant. However, people in especially high-income areas might want to eat in a restaurant that has more “atmosphere” than a family restaurant like Rohat’s. As a result, you worry that the income variable might be only weakly positive in its impact. To sum, you expect:

$$y_i = a + \beta_N N_i + \beta_P P_i + \beta_I I_i + u_i \quad (4.3)$$

where the signs above the coefficients indicate the expected impact of that particular independent variable on the dependent variable, holding constant the other two explanatory variables, and u_i is a typical stochastic error term.

4. Collect the data. Inspect and clean the data.

You want to include every local restaurant in the Rohat’s chain in your study, and, after some effort, you come up with data for your dependent variable and your independent variables for all 33 locations. You inspect the data, and you’re confident that the quality of your data is excellent for three reasons: each manager measured each variable identically, you’ve included each restaurant in the sample, and all the information is from the same year. [The data set is included in this section, along with a sample computer output for the regression estimated by Stata (Tables 4.1 and 4.2 below).].

Table 4.1 Data for the Rohat's restaurant example

	Y	N	P	I
1.	107919	3	65044	13240
2.	118866	5	101376	22554
3.	98579	7	124989	16916
4.	122015	2	55249	20967
5.	152827	3	73775	19576
6.	91259	5	48484	15039
7.	123550	8	138809	21857
8.	160931	2	50244	26435
9.	98496	6	104300	24024
10.	108052	2	37852	14987
11.	144788	3	66921	30902
12.	164571	4	166332	31573
13.	105564	3	61951	19001
14.	102568	5	100441	20058
15.	103342	2	39462	16194
16.	127030	5	139900	21384
17.	166755	6	171740	18800
18.	125343	6	149894	15289
19.	121886	3	57386	16702
20.	134594	6	185105	19093
21.	152937	3	114520	26502
22.	109622	3	52933	18760
23.	149884	5	203500	33242
24.	98388	4	39334	14988
25.	140791	3	95120	18505
26.	101260	3	49200	16839
27.	139517	4	113566	28915
28.	115236	9	194125	19033
29.	136749	7	233844	19200
30.	105067	7	83416	22833
31.	136872	6	183953	14409
32.	117146	3	60457	20307
33.	163538	2	65065	20111

(obs=33)

	Y	N	P	I
Y	1.0000			
N	-0.1442	1.0000		
P	0.3926	0.7263	1.0000	
I	0.5370	-0.0315	0.2452	1.0000

Table 4.2 Actual computer output (using the stata program)

Source	SS	df	MS	Number of obs	=	33
Model	9.9289e + 09	3	3.3096e + 09	F(3, 29)	=	15.65
Residual	6.1333e + 09	29	211492485	Prob > F	=	0.0000
Total	1.6062e + 10	32	501943246	R-squared	=	0.6182
				Adj R-squared	=	0.5787
				Root MSE	=	14543

Y	Coef.	Std. Err.	t	P > t	[95% Conf. Interval]
N	−9074.674	2052.674	−4.42	0.000	−13272.86 −4876.485
P	.3546684	.0726808	4.88	0.000	.2060195 .5033172
I	1.287923	.5432938	2.37	0.025	.1767628 2.399084
_cons	102192.4	12799.83	7.98	0.000	76013.84 128371

	Y	Yhat	residuals
1.	107919	115089.6	−7170.56
2.	118866	121821.7	−2955.74
3.	98579	104785.9	−6206.864
4.	122015	130642	−8627.041
5.	152827	126346.5	26480.55
6.	91259	93383.88	−2124.877
7.	123550	106976.3	16573.66
8.	160931	135909.3	25021.71
9.	98496	115677.4	−17181.36
10.	108052	116770.1	−8718.094
11.	144788	138502.6	6285.425
12.	164571	165550	−979.0342
13.	105564	121412.3	−15848.3
14.	102568	118275.5	−15707.47
15.	103342	118895.6	−15553.63
16.	127030	133978.1	−6948.114
17.	166755	132868.1	33886.91
18.	125343	120598.1	4744.898
19.	121886	116832.3	5053.7
20.	134594	137985.6	−3391.591
21.	152937	149717.6	3219.428
22.	109622	117903.5	−8281.508
23.	149884	171807.2	−21923.22
24.	98388	99147.65	−759.6514
25.	140791	132537.5	8253.518
26.	101260	114105.4	−12845.43
27.	139517	143412.3	−3895.303
28.	115236	113883.4	1352.599
29.	136749	146334.9	−9585.905
30.	105067	97661.88	7405.122
31.	136872	131544.4	5327.621
32.	117146	122564.5	−5418.45
33.	163538	133021	30517

5. Estimate and evaluate the equation.

You take the data set and enter it into the computer. You then run an OLS regression on the data, but you do so only after thinking through your model once again to see if there are hints that you've

made theoretical mistakes. You end up admitting that although you cannot be sure you are right, you've done the best you can, so you estimate the equation, obtaining:

$$\hat{y}_i = 102.192 + 9075N_i + 0.355P_i + 1.288I_i$$

(2053)	(0.073)	(0.543)	
$t = -4.42$	4.88	2.37	(4.4)
$N = 33$	$\bar{R}^2 = 0.579$		

This equation satisfies your needs in the short run. In particular, the estimated coefficients in the equation have the signs you expected. The overall fit, although not outstanding, seems reasonable for such a diverse group of locations. To predict Y, you obtain the values of N, P, and I for each potential new location and then plug them into Equation 4.4. Other things being equal, the higher the predicted Y, the better the location from Rohat's point of view.

6. Document the results.

The results summarized in Equation 4.4 meet our documentation requirements. (Note that we include the standard errors of the estimated coefficients and t -values⁶ for completeness, even though we won't make use of them until Chapter 6.) However, it's not easy for a beginning researcher to wade through a computer's regression output to find all the numbers required for documentation. You'll probably have an easier time reading your own computer system's printout if you take the time to "walk through" the sample computer output for the Rohat's model in Tables 4.1–4.2. This sample output was produced by the Stata computer program, but it's similar to those produced by Eviews, Statistica, Microfit, and others.

⁶ Throughout the text, the number in parentheses below a coefficient estimate typically will be the standard error of that estimated coefficient. Some authors put the t -value in parentheses, though, so be alert when reading journal articles or other books.

The first items listed are the actual data. These are followed by the simple correlation coefficients between all pairs of variables in the data set. Next comes a listing of the estimated coefficients, their estimated standard errors, and the associated t -values, and follows with R^2 , $\bar{R}^2$, RSS , the F -ratio, and other items that we will explain in later chapters. Finally, we have a listing of the observed y s, the predicted Y s, and the residuals for each observation. Numbers followed by “ $e + 06$ ” or “ $e - 01$ ” are expressed in a scientific notation indicating that the printed decimal point should be moved six places to the right or one place to the left, respectively. In future sections, we’ll return to this example in order to apply various tests and ideas as we learn them.

4.3 Dummy variables

Some concepts (for example, gender) might seem impossible to include in an equation because they’re inherently qualitative in nature and can’t be expressed as a number. Luckily, such concepts can be quantified by using dummy (or binary) variables. A dummy variable takes on the value of one or zero (and only those values) depending on whether a specified condition is met.

As an illustration of a dummy variable, suppose that Y_i represents the salary of the i th high school teacher and that salaries depend primarily on the experience of the teacher and the type of degree that the teacher has earned. All teachers have a B.A., but some also have a graduate degree like an M.A. An equation representing the relationship between earnings and these variables would be:

$$y_i = a + \beta_1^+ x_i + \beta_2^+ d_i + u_i \quad (4.5)$$

where y_i = the income of the i th teacher in dollars
 x_i = the number of years of teaching experience of the i th

$$d_i = \begin{cases} 1 & \text{if the } i\text{th teacher has a graduate degree} \\ 0 & \text{otherwise} \end{cases}$$

The variable d_i takes on values of only zero or one, so d_i is called a dummy variable, or just a “dummy.” Needless to say, the term has generated many a pun. In this case, the dummy variable represents the condition of having a graduate degree. The coefficient β_2 indicates the additional salary that can be attributed to having a graduate degree, holding teaching experience constant. Since more experience and a graduate degree can be expected to increase the earnings of teachers, we expect positive coefficients for both variables, as indicated by the signs above the coefficients in Equation 4.5. Think for a second about what those expected signs would be if we had instead defined d_i to be equal to one if the i th teacher has no graduate degree and equal to zero otherwise. This change shouldn’t impact the expected sign of β_1 , but do you see that the expected sign of β_2 now would be negative?⁷ As can be seen in Figure 4.2, the dummy changes the intercept depending on the value of d , but the slopes remain constant no matter what value d takes. This is true even if we define the dummy variable “reversed” and have $d=0$ if the particular condition is met and $d=1$ otherwise. The slopes still remain constant.

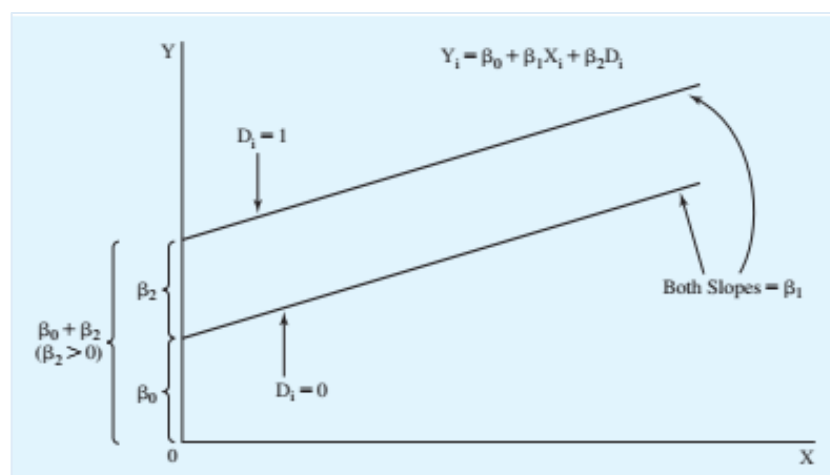


Figure 4.2 Dummy variable representation

⁷ The constant term will change as well.

If a dummy ($\beta_2 d_i$) is added to an equation, a graph of the equation will have different intercepts for the two qualitative conditions specified by the dummy variable. The difference between the two intercepts is β_2 . The slopes are constant with respect to the qualitative condition.

Note that in this example only one dummy variable is used even though there were two conditions. This is because one fewer dummy variable is constructed than conditions. The event not explicitly represented by a dummy variable, the **omitted condition**, forms the basis against which the included conditions are compared. Thus, for dual situations only one dummy variable is entered as an independent variable; the coefficient is interpreted as the effect of the included condition relative to the omitted condition. Be careful never to use two dummy variables to describe the two conditions. If you were to make this mistake, sometimes called a *dummy variable trap*, you'd have perfect multicollinearity (to be described in Section 9.1).

For another example of the meaning of the coefficient of a dummy variable, let's look at a study of the relationship between fraternity/sorority membership and grade point average (GPA). Most noneconometricians would approach this research problem by calculating the mean grades of fraternity/sorority (so-called Greek) members and comparing them to the mean grades of nonmembers. However, such a technique would ignore the relationship that grades have to characteristics other than Greek membership.

Instead, we'd want to build a regression model that explains college GPA. Independent variables would include not only Greek membership but also other predictors of academic performance such as SAT scores and high school grades. Being a member of a social organization is a qualitative variable, however, so we'd have to create a dummy variable to represent fraternity or sorority membership quantitatively in a regression equation:

$$d_i = \begin{cases} 1 & \text{if the } i\text{th student is an active member} \\ & \text{of a fraternity or sorority} \\ 0 & \text{otherwise} \end{cases}$$

If we collect data from all the students in our class and estimate the equation implied in this example, we obtain:

$$CG_i = 0.37 + 0.81HG_i + 0.00001S_i - 0.38D_i \quad (4.6)$$

$$\bar{R}^2 = 0.45 \quad N = 25$$

where CG_i = the cumulative college GPA (4-point scale) of the i th student

HG_i = the cumulative high school GPA (4-point scale) of the i th student

S_i = the sum of the highest verbal and mathematics SAT scores earned by the i th student.

The meaning of the estimated coefficient of D_i in Equation 4.6 is very specific. Stop for a second and Figure it out for yourself. What is it? The estimate that $\hat{\beta}_D = -0.38$ means that, for this sample, the GPA of fraternity/sorority members is 0.38 lower than for nonmembers, holding SATs and high school GPA constant. Thus, Greek members are doing about a third of a grade worse than otherwise might be expected. To understand this example better, try using Equation 3.6 to predict your own GPA; how close does it come?

Before you rush out and quit whatever social organization you're in, however, note that this sample is quite small and that we've surely omitted some important determinants of academic success from the equation. As a result, we shouldn't be too quick to conclude that Greeks are dummies.

Up to this point, we've used dummy variables to represent only those qualitative variables that have exactly two possibilities (such as gender). What about situations where a qualitative variable has three or more alternatives? For example, in our study of the salaries of high school teachers in Equation 3.5, what if we learn that some of the teachers have Ph.D.s? We now need to be able to distinguish teachers whose highest degree is a Ph.D. from teachers

whose highest degree is an M.A. from teachers whose highest degree is a B.A. What can we do?

Well, the answer certainly isn't to define a variable such that Ph.D. = 2, M.A. = 1, and B.A. = 0, because we have no reason to think that the impact of having a Ph.D. is exactly twice that of having an M.A. If not that, then what?

The answer is to create one fewer dummy variable than there are possibilities (conditions) and to use each dummy to represent only one of the possible conditions. In the high school salary case, you'd create *two* dummy variables to represent the three conditions, for example:

$$PhD_i = \begin{cases} 1 & \text{if the } i\text{th teacher's highest degree is PhD} \\ 0 & \text{otherwise} \end{cases}$$

$$MA_i = \begin{cases} 1 & \text{if the } i\text{th teacher's highest degree is MA} \\ 0 & \text{otherwise} \end{cases}$$

The omitted condition (when a B.A. is the highest degree) is represented by having both dummies equal to 0. This way you can measure the impact of each degree independently without having to link the impacts of having an M.A. and a Ph.D. Thus Equation 4.5 now would look like this:

$$y_i = a + \beta_1^+ x_i + \beta_2^+ PhD_i + \beta_3^? MA_i + u_i \quad (4.7)$$

However, be careful! The interpretation of the coefficients when there are two or more related dummy variables is tricky. The coefficient tells you the increase in the dependent variable caused by the condition being met *compared to the omitted condition*. Thus β_3 measures the impact of having the highest degree be an M.A. (holding x and PHD constant) compared to the omitted condition, which is when the highest degree is a B.A. To make sure that you understand this, go back to Equation 4.7 and determine the expected sign of β_3 . Did you decide it should be positive? That's right! We'd expect a high school teacher whose highest degree is an M.A. to

have a higher salary than one whose highest degree is a B.A. (holding x and PHD constant).

A dummy variable that has only a single observation with a value of 1 while the rest of the observations are 0 (or vice versa) is to be avoided unless the variable is required by theory. Such a *one-time dummy* acts merely to eliminate that observation from the data set, improving the fit artificially by setting the dummy's coefficient equal to the residual for that observation. One would obtain exactly the same estimates of the other coefficients if that observation were deleted, but the deletion of an observation is rarely, if ever, appropriate.

While this is the end of the section, it's not the end of our coverage of dummy variables. In Section 8.4, we'll discuss slope dummy variables, and in Chapter 13 we'll analyze what happens when the *dependent* variable is a dummy.

SUMMARY

1. Six steps typically taken in applied regression analysis for a given dependent variable are:
 - a. Review the literature and develop the theoretical model.
 - b. Specify the model: Select the independent variables and the functional form.
 - c. Hypothesize the expected signs of the coefficients.
 - d. Collect the data. Inspect and clean the data.
 - e. Estimate and evaluate the equation.
 - f. Document the results.
2. A dummy variable takes on only the values of 1 or 0, depending on whether some condition is met. An example of a dummy variable would be x equals 1 if a particular individual is female and 0 if the person is male.

CHAPTER 5. THE CLASSICAL MODEL

The Classical model of econometrics has nothing to do with ancient Greece or even the classical economic thinking of Adam Smith. Instead, the term *classical* refers to a set of fairly basic assumptions required to hold in order for OLS to be considered the “best” estimator available for regression models. When one or more of these assumptions do not hold, other estimation techniques (such as generalized least squares, to be explained in Chapter 9) may be better than OLS.

As a result, one of the most important jobs in regression analysis is to decide whether the classical assumptions hold for a particular equation. If so, the OLS estimation technique is the best available. Otherwise, the pros and cons of alternative estimation techniques must be weighed. These alternatives usually are adjustments to OLS that take account of the particular assumption that has been violated. In a sense, most of the rest of this book deals in one way or another with the question of what to do when one of the classical assumptions is not met. Since econometricians spend so much time analyzing violations of them, it is crucial that they know and understand these assumptions.

5.1 The Classical assumptions

The Classical assumptions must be met in order for OLS estimators to be the best available. Because of their importance in regression analysis, the assumptions are presented following:

1. The regression model is linear, is correctly specified, and has an additive error term.
2. The error term has a zero population mean.
3. All explanatory variables are uncorrelated with the error term.
4. Observations of the error term are uncorrelated with each other (no serial correlation).

5. The error term has a constant variance (no heteroskedasticity).
6. No explanatory variable is a perfect linear function of any other explanatory variable(s) (no perfect multicollinearity).
7. The error term is normally distributed (this assumption is optional but usually is invoked).

Subsequent chapters will investigate major violations of the assumptions and introduce estimation techniques that may provide better estimates in such cases. An error term satisfying Assumptions I through V is called a classical error term, and if Assumption 7 is added, the error term is called a classical normal error term.

1. ***The regression model is linear, is correctly specified, and has an additive error term.*** The regression model is assumed to be linear:

$$y_i = a + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + u_i \quad (5.1)$$

The assumption that the regression model is linear⁸ does not require the underlying theory to be linear. For example, an exponential function:

$$y_i = e^{\beta_0} x_i^{\beta_1} e^{u_i} \quad (5.2)$$

where e is the base of the natural log, can be transformed by taking the natural log of both sides of the equation:

$$\ln(y_i) = a + \beta_1 \ln(x_i) + u_i \quad (5.3)$$

If the variables are relabeled as $y_i^* = \ln(y_i)$ and $x_i^* = \ln(x_i)$, then the form of the equation becomes linear:

⁸ The Classical Assumption that the regression model is “linear” technically requires the model to be “linear in the coefficients.” You’ll learn what it means for a model to be linear in the coefficients, particularly in contrast to being linear in the variables, in Section 8.2. We’ll cover the application of regression analysis to equations that are nonlinear in the variables in that same section, but the application of regression analysis to equations that are nonlinear in the coefficients is beyond the scope of this textbook.

$$y_i^* = a + \beta_1 x_i^* + u_i \quad (5.4)$$

In Equation 5.4, the properties of the OLS estimator of the β s still hold because the equation is linear. Two additional properties also must hold. First, we assume that the equation is correctly specified. If an equation has an omitted variable or an incorrect functional form, the odds are against that equation working well. Second, we assume that a stochastic error term has been added to the equation. This error term must be an additive one and cannot be multiplied by or divided into any of the variables in the equation.

2. ***The error term has a zero population mean.*** As was pointed out in Section 2.1, econometricians add a stochastic (random) error term to regression equations to account for variation in the dependent variable that is not explained by the model. The specific value of the error term for each observation is determined purely by chance. Probably the best way to picture this concept is to think of each observation of the error term as being drawn from a random variable distribution such as the one illustrated in Figure 5.1.

Classical assumption 2 says that the mean of this distribution is zero. That is, when the entire population of possible values for the stochastic error term is considered, the average value of that population is zero. For a small sample, it is not likely that the mean is exactly zero, but as the size of the sample approaches infinity, the mean of the sample approaches zero.

What happens if the mean doesn't equal zero in a sample? As long as you have a constant term in the equation, the estimate of a will absorb the nonzero mean. In essence, the constant term equals the *fixed* portion of y that cannot be explained by the independent variables, and the error term equals the *stochastic* portion of the unexplained value of y .

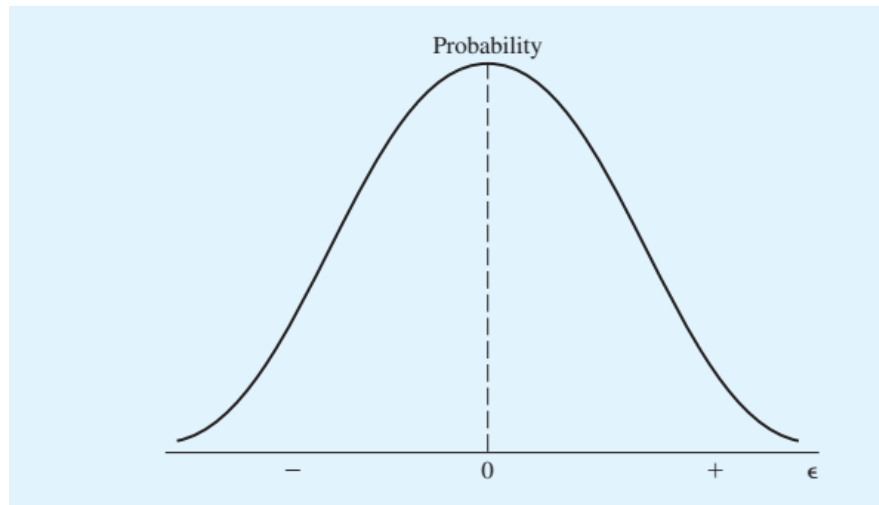


Figure 5.1 An error term distribution with a mean of zero.

Observations of stochastic error terms are assumed to be drawn from a random variable distribution with a mean of zero. If Classical assumption II is met, the expected value (the mean) of the error term is zero.

Although it's true that the error term never can be observed, it's instructive to pretend that we can do so to see how the constant term absorbs the non-zero mean of the error term in a sample. Consider a typical regression equation:

$$y_i = a + \beta_1 x_i + u_i \quad (5.5)$$

Suppose that the mean of u_i is 3 instead of 0, then⁹ $E(u_i - 3) = 0$. If we add 3 to the constant term and subtract it from the error term, we obtain:

$$y_i = (a + 3) + \beta_1 x_i + (u_i - 3) \quad (5.6)$$

Since equations (5.5) and (5.6) are equivalent (do you see why?), and since $E(u_i - 3) = 0$, then Equation 5.6 can be written in a form that has a zero mean for the error term u_i^* :

⁹ Here, as in Chapter 2, the “E” refers to the expected value (mean) of the item in parentheses after it. Thus $E(u_i - 3)$ equals the expected value of the stochastic error term epsilon minus 3. In this specific example, since we've defined $E(u_i) = 3$, we know that $E(u_i - 3) = 0$. One way to think about expected value is as our best guess of the long-run average value a specific random variable will have.

$$y_i = a^* + \beta_1 x_i + u_i^* \quad (5.7)$$

where $a^* = a + 3$ and $u_i^* = u_i - 3$. As can be seen, equation 5.7 conforms to Assumption 2. In essence, if Classical assumption 2 is violated in an equation that includes a constant term, then the estimate of a absorbs the nonzero mean of the error term, and the estimates of the other coefficients are unaffected.

3. All explanatory variables are uncorrelated with the error term. It is assumed that the observed values of the explanatory variables are independent of the values of the error term.

If an explanatory variable and the error term were instead correlated with each other, the OLS estimates would be likely to attribute to the X some of the variation in Y that actually came from the error term. If the error term and X were positively correlated, for example, then the estimated coefficient would probably be higher than it would otherwise have been (biased upward), because the OLS program would mistakenly attribute the variation in y caused by u to x instead. As a result, it's important to ensure that the explanatory variables are uncorrelated with the error term.

Classical assumption 3 is violated most frequently when a researcher omits an important independent variable from an equation.¹⁰ As you learned in Chapter 2, one of the major components of the stochastic error term is omitted variables, so if a variable has been omitted, then the error term will change when the omitted variable changes. If this omitted variable is correlated with an included independent variable (as often happens in economics), then the error term is correlated with that independent variable as well. We have violated Assumption 3! Because of this violation, OLS will attribute the impact of the omitted variable to the included variable, to the extent that the two variables are correlated.

4. Observations of the error term are uncorrelated with each other. The observations of the error term are drawn independently from each other. If a systematic correlation exists between one observation of the error term and another, then OLS estimates will

¹⁰ Another important economic application that violates this assumption is any model that is simultaneous in nature. This will be considered in Chapter 14.

be less precise than estimates that account for the correlation. For example, if the fact that the u from one observation is positive increases the probability that the u from another observation also is positive, then the two observations of the error term are positively correlated. Such a correlation would violate Classical assumption 4.

In economic applications, this assumption is most important in time-series models. In such a context, Assumption 4 says that an increase in the error term in one time period (a random shock, for example) does not show up in or affect in any way the error term in another time period. In some cases, though, this assumption is unrealistic, since the effects of a random shock sometimes last for a number of time periods. For example, a natural disaster like the 2015 earthquake in Nepal will have a negative impact on a region long after the time period in which it was truly a random event. If, over all the observations of the sample, u_{t+1} is correlated with u_t , then the error term is said to be *serially correlated* (or *autocorrelated*), and Assumption 4 is violated. Violations of this assumption are considered in more detail in Chapter 10.

5. The error term has a constant variance. The variance (or dispersion) of the distribution from which the observations of the error term are drawn is constant.¹¹ That is, the observations of the error term are assumed to be drawn continually from identical distributions (for example, the one pictured in Figure 5.1).

¹¹ This is a simplification. The actual assumption (that error terms have positive finite second moments) is equivalent to this simplification in all but a few extremely rare cases.

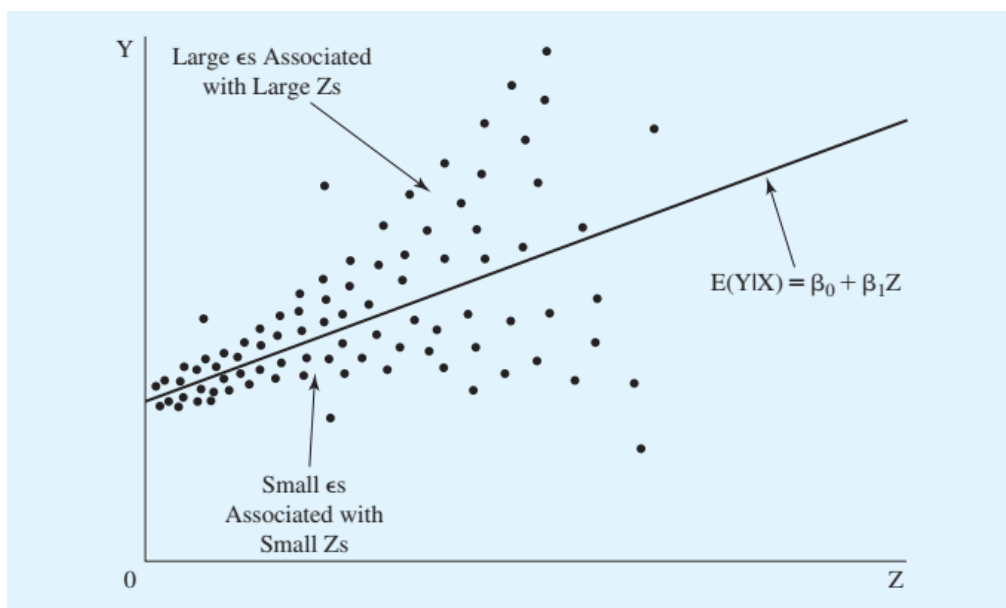


Figure 5.2 An error term whose variance increases as Z increases.

One example of Classical assumption V not being met is when the variance of the error term increases as Z increases. In such a situation (called heteroskedasticity), the observations are on average farther from the true regression line for large values of Z than they are for small values of Z .

The alternative would be for the variance of the distribution of the error term to change for each observation or range of observations. In Figure 5.2, for example, the variance of the error term is shown to increase as the variable Z increases; such a pattern violates Classical assumption 5. The actual values of the error term are not directly observable, but the lack of a constant variance for the distribution of the error term causes OLS to generate inaccurate estimates of the standard error of the coefficients.¹²

For example, suppose that you're studying the amount of money that the districts of Tajikistan spend on education. Dushanbe and Khujand are more heavily populated than Vahdat and Hissor, so it's probable that the variance of the error term for big states is larger than it is for small states. The amount of unexplained variation in educational expenditures seems likely to be larger in big

¹² Because some observations have errors with a large variance, those observations are not as reliable and so should be given less weight when minimizing the sum of squares. OLS, however, gives equal weight to each observation, so it will be less precise than estimators that weigh the observations more appropriately.

states like Dushanbe and Khujand than in small districts like Vahdat. The violation of Assumption 5 is referred to as *heteroskedasticity* and will be discussed in more detail in Chapter 11.

6. No explanatory variable is a perfect linear function of any other explanatory variable(s). Perfect *collinearity* between two independent variables implies that they are really the same variable, or that one is a multiple of the other, and/or that a constant has been added to one of the variables. That is, the relative movements of one explanatory variable will be matched exactly by the relative movements of the other even though the absolute size of the movements might differ. Because every movement of one of the variables is matched exactly by a relative movement in the other, the OLS estimation procedure will be incapable of distinguishing one variable from the other.

Many instances of perfect collinearity (or *multicollinearity* if more than two independent variables are involved) are the result of the researcher not accounting for identities (definitional equivalences) among the independent variables. This problem can be corrected easily by dropping one of the perfectly collinear variables from the equation.

What's an example of perfect multicollinearity? Suppose that you decide to build a model of the profits of tire stores in your city and you include annual sales of tires (in dollars) at each store and the annual sales tax paid by each store as independent variables. Since the tire stores are all in the same city, they all pay the same percentage sales tax, so the sales tax paid will be a constant percentage of their total sales (in dollars). If the sales tax rate is 7%, then the total taxes paid will be 7% of sales for each and every tire store. Thus sales tax will be a perfect linear function of sales, and you'll have perfect multicollinearity!

Perfect multicollinearity also can occur when two independent variables always sum to a third or when one of the explanatory variables doesn't change within the sample. With perfect multicollinearity, OLS (or any other estimation technique) will be unable to estimate the coefficients of the collinear variables (unless there is a rounding error). While it's quite unusual for an experienced researcher to encounter perfect multicollinearity, even

imperfect multicollinearity can cause exercises for estimation, as you will see in Chapter 8.

7. The error term is normally distributed. Although we have already assumed that observations of the error term are drawn independently (Assumption 4) from a distribution that has a zero mean (Assumption 2) and that has a constant variance (Assumption 5), we have said little about the shape of that distribution. Assumption 7 states that the observations of the error term are drawn from a distribution that is normal (that is, bell-shaped, and generally following the symmetrical pattern portrayed in Figure 5.3).

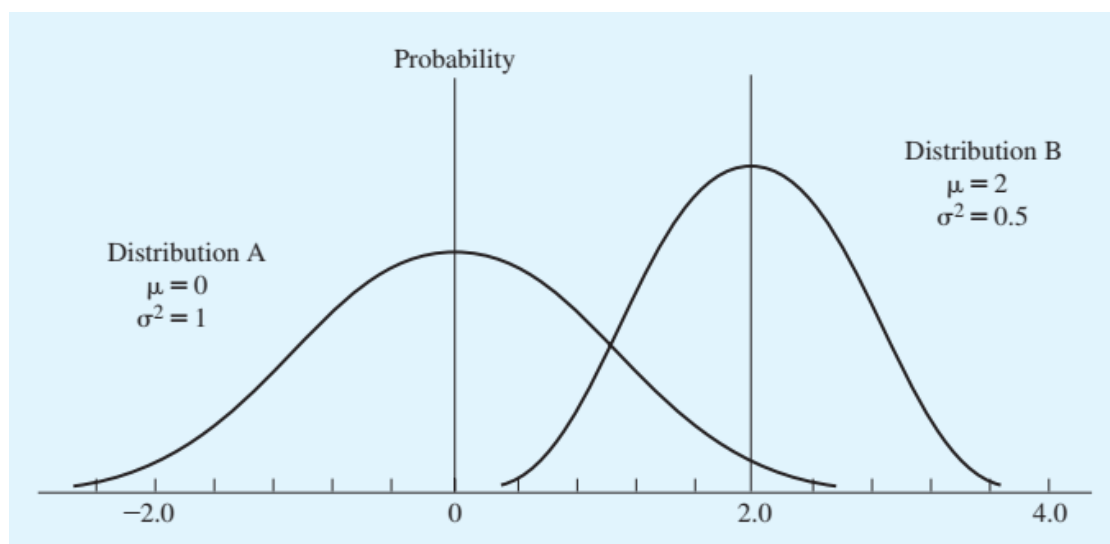


Figure 5.3 Normal distribution

Although all normal distributions are symmetrical and bell-shaped, they do not necessarily have the same mean and variance. Distribution A has a mean of 0 and a variance of 1, whereas distribution B has a mean of 2 and a variance of 0.5. As can be seen, the whole distribution shifts when the mean changes, and the distribution gets fatter as the variance increases.

This assumption of normality is not required for OLS estimation. Its major application is in *hypothesis testing* and *confidence intervals*, which use the estimated regression coefficient to investigate hypotheses about economic behavior. Hypothesis testing is the subject of Chapter 6, and without the normality assumption, most of the small sample tests of that chapter would be

invalid. Even though Assumption 7 is optional, it's usually advisable to add the assumption of normality to the other six assumptions for two reasons:

1. The error term e_i can be thought of as the sum of a number of minor influences or errors. As the number of these minor influences gets larger, the distribution of the error term tends to approach the normal distribution.

2. The t -statistic and the F -statistic, which will be developed in Chapter 6, are not truly applicable unless the error term is normally distributed. A quick look at Figure 5.3 shows how normal distributions differ when the means and variances are different. In normal distribution A (a **standard normal distribution**), the mean is 0 and the variance is 1; in normal distribution B, the mean is 2, and the variance is 0.5. When the mean is different, the entire distribution shifts. When the variance is different, the distribution becomes fatter or skinnier.

5.2 The Sampling distribution of $\hat{\beta}$

Just as the error term follows a probability distribution, so too do the estimates of β . In fact, each different sample of data typically produces a different estimate of β . The probability distribution of these $\hat{\beta}$ values across different samples is called the **sampling distribution of $\hat{\beta}$** .

Recall that an *estimator* is a formula, such as the OLS formula in Equation 2.4 that tells you how to compute $\hat{\beta}$, while an *estimate* is the value of $\hat{\beta}$ computed by the formula for a given sample. Since researchers usually have only one sample, beginning econometricians often assume that regression analysis can produce only one estimate of β for a given population. In reality, however, each different sample from the same population will produce a different estimate of β . The collection of all the possible samples has a distribution, with a mean and a variance, and we need to discuss the properties of this sampling distribution of $\hat{\beta}$, even

though in most real applications we will encounter only a single draw from it. Be sure to remember that a sampling distribution refers to the distribution of different values of $\hat{\beta}$ across different samples, not within one. These $\hat{\beta}$ s usually are assumed to be normally distributed because the normality of the error term implies that the OLS estimates of β are normally distributed as well.

Let's look at an example of a sampling distribution of $\hat{\beta}$ by going back to the height and weight example of Chapter 2 (with weight measured in pounds and height measured in inches above five feet).

$$weight_i = a + \beta_1^+ height_i + u_i \quad (5.8)$$

Suppose you take a sample of six students and apply OLS to equation 5.8 to get an estimate of β_1 . So far, so good.

But what will happen if you select a second sample of six students and do the same thing? Will you get the exact same $\hat{\beta}_1$? Nope! Your second $\hat{\beta}_1$ will depend on the second sample, which almost surely will be different from your first sample. If your random sample includes a couple of football linemen, you're likely to get a really large coefficient. If you randomly choose some cross country runners, you'll get a low estimate. Even if there's nothing unusual about your second sample, you'll almost certainly get a different $\hat{\beta}_1$. Why? Different data yield different estimates, so if you collect 100 samples, you're likely to get 100 different $\hat{\beta}_1$ s.

All these $\hat{\beta}_1$ estimates come from a distribution with its own mean and variance, called a *sampling distribution*. To help you understand this, we took 100 different samples of six students and estimated equation (5.8) 100 times. Take a look at Figure 5.4. With the help of a histogram, we've graphed all 100 $\hat{\beta}_1$ s so that we can get an idea of what the sampling distribution looks like.

While the histogram in Figure 5.4 isn't perfectly normally distributed (represented by the thin line), it's close. Notice how the

estimates are clustered in the middle (near the mean of 7.75), with fewer and fewer estimates in the tails. With many more estimates of β_1 , we'd expect the histogram to look even more like a normal curve.

For an estimation technique to be “good,” the mean of the sampling distribution of the $\hat{\beta}$ s it produces should equal the true population β . This property has a special name in econometrics: *unbiasedness*. There's much more to come on this idea.

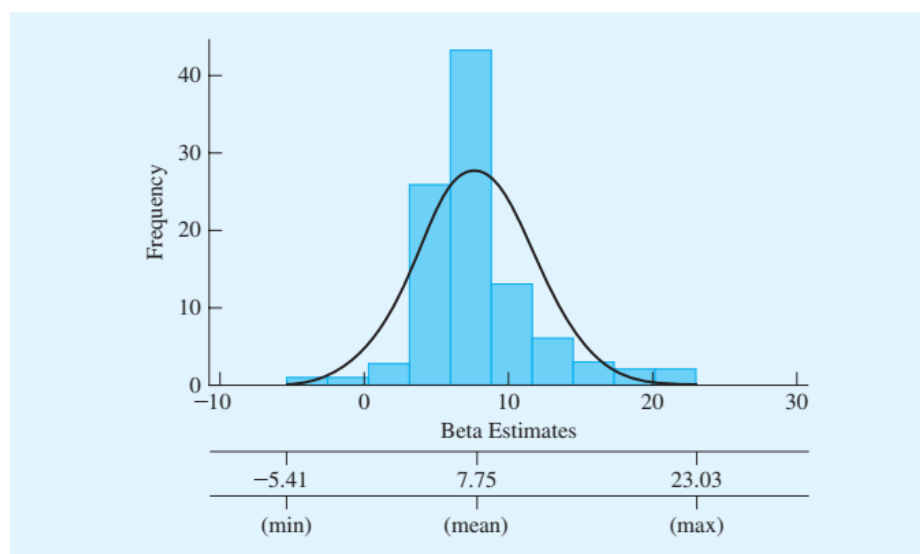


Figure 5.4 A height and weight sampling distribution of $\hat{\beta}$

We estimated Equation 5.8 (the height weight equation) with 100 samples of six students each and plotted the 100 $\hat{\beta}$ s in Figure 5.4. The result is a sampling distribution of $\hat{\beta}$ with a mean of 7.75 and a pattern that is reasonably close to being normally distributed (the thin line).

Although we don't know the true β in this case, it's likely that if we took enough samples - thousands perhaps - the mean of the $\hat{\beta}_1$ s would approach the true β . For example, when we took 1,000 samples of six, the mean of the $\hat{\beta}_1$ s was 6.88. The chances are that 6.88 is closer to the true β than is 7.75, the mean of the 100 estimates shown in Figure 4.4.

The moral of the story is that while a single sample provides a single estimate of β_1 , that estimate comes from a sampling

distribution with a mean and a variance. Other estimates from that sampling distribution will most likely be different. When we discuss the properties of estimators in the next section, it will be important to remember that we are discussing the properties of a sampling distribution, not the properties of one sample.

Properties of the mean

A desirable property of a distribution of estimates is that its mean equals the true mean of the variable being estimated. An estimator that yields such estimates is called an unbiased estimator. An estimator $\hat{\beta}$ is an **unbiased estimator** if its sampling distribution has as its expected value the true value of β .

$$E(\hat{\beta}) = \beta \quad (5.9)$$

Only one value of $\hat{\beta}$ is obtained in practice, but the property of unbiasedness is useful because a single estimate drawn from an unbiased distribution is more likely to be near the true value (assuming identical variances) than one taken from a distribution not centered around the true value. If an estimator produces $\hat{\beta}$ s that are not centered around the true β , the estimator is referred to as a **biased estimator**.

We cannot ensure that every estimate from an unbiased estimator is better than every estimate from a biased one, because a particular unbiased estimate¹³ could, by chance, be farther from the true value than a biased estimate might be. This could happen by chance or because the biased estimator had a smaller variance. For example, a broken clock is a biased estimator of the time of day, but it has a zero variance and happens to be exactly right twice a day. Without any other information about the distribution of the

¹³ Technically, since an estimate has just one value, an estimate cannot be unbiased (or biased). On the other hand, the phrase “estimate produced by an unbiased estimator” is cumbersome, especially if repeated 10 times on a page. As a result, many econometricians use “unbiased estimate” as shorthand for “a single estimate produced by an unbiased estimator.”

estimates, however, we would always rather have an unbiased estimate than a biased one.

Properties of the variance

Just as we would like the distribution of the β_N s to be centered around the true population β , we would also like that distribution to be as narrow (or precise) as possible. A distribution centered around the truth but with an extremely large variance might be of very little use because any given estimate would quite likely be far from the true β value. For a $\hat{\beta}$ distribution with a small variance, the estimates are likely to be close to the mean of the sampling distribution. To see this more clearly, compare distributions A and B (both of which are unbiased) in Figure 4.5. Distribution A, which has a larger variance than distribution B, is less precise than distribution B. Figure 5.5 distributions of $\hat{\beta}$.

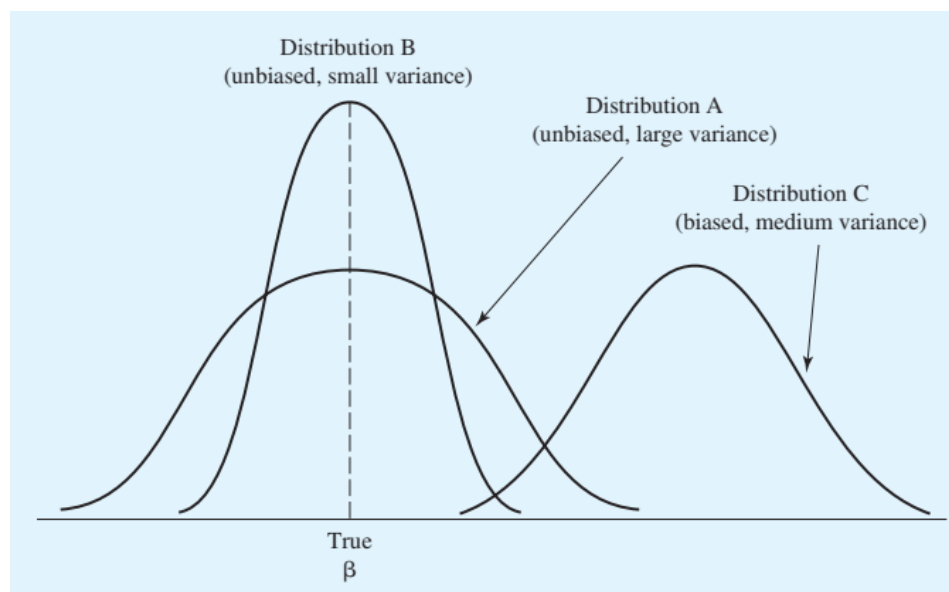


Figure 5.5 Distributions of $\hat{\beta}$

Different distributions of $\hat{\beta}$ can have different means and variances. Distributions A and B, for example, are both unbiased, but distribution A has a larger variance than does distribution B. Distribution C has a smaller variance than distribution A, but it is biased.

For comparison purposes, a biased distribution (distribution C) is also pictured; note that bias implies that the expected value of the distribution is to the right or left of the true β .

The variance of the distribution of the $\hat{\beta}$ s can be decreased by increasing the size of the sample. This also increases the degrees of freedom, since the number of degrees of freedom equals the sample size minus the number of coefficients or parameters estimated. As the number of observations increases, other things held constant, the variance of the sampling distribution tends to decrease. Although it is not true that a sample of 60 will always produce estimates closer to the true β than a sample of 6, it is quite likely to do so; such larger samples should be sought. Figure 5.6 presents illustrative sampling distributions of $\hat{\beta}$ s for 6, 60, and 600 observations for OLS estimators of β when the true β equals 1. The larger samples do indeed produce sampling distributions that are more closely centered around β .

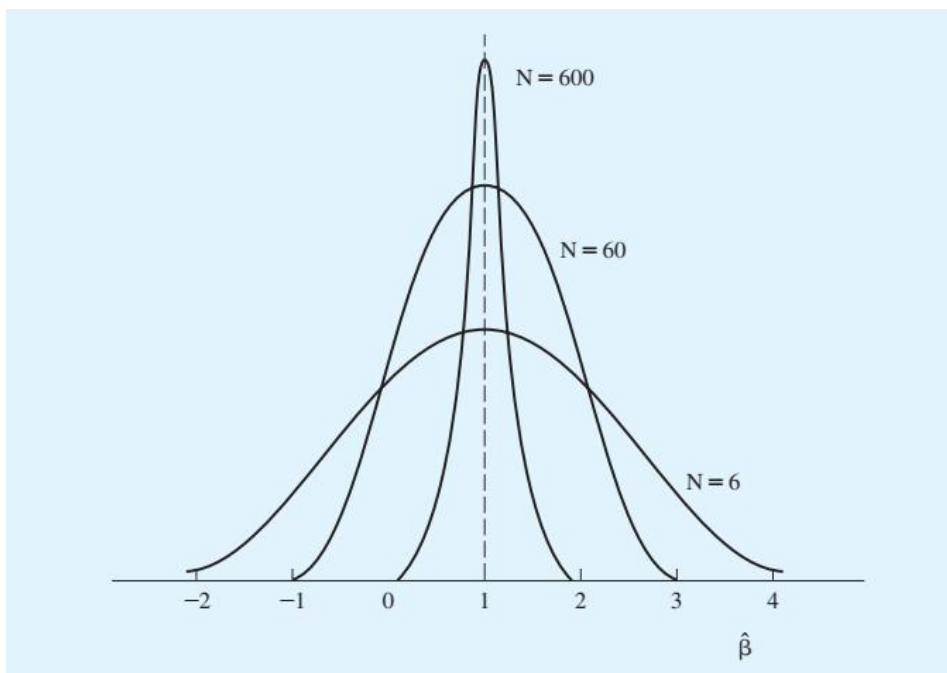


Figure 5.6 Sampling distribution of $\hat{\beta}$ for various observations (N).

As the size of the sample increases, the variance of the distribution of $\hat{\beta}$ s calculated from that sample tends to decrease. In the extreme case (not shown), a sample equal to the population would yield only an estimate equal

to the mean of that distribution, which (for unbiased estimators) would equal the true β , and the variance of the estimates would be zero.

The powerful lesson illustrated by Figure 5.6 is that if you want to maximize your chances of getting an estimate close to the true value, apply OLS to a large sample. There's no guarantee that you'll get a more accurate estimate from a large sample, but your chances are better. Larger samples, all else equal, tend to result in more precise estimates. And if the estimator is unbiased, more precise estimates are more accurate estimates.

Think of it this way. Having a couple of cross country runners might lead to a pretty wacky estimate from a sample of 6, but their influence on $\hat{\beta}_1$ is going to be much smaller in a sample of 60. You could imagine getting 2 cross country runners in a random sample of 6 students, but it would be virtually impossible to get 20 cross country runners in a sample of 60. So try to get larger samples.

In econometrics, we must rely on general tendencies. The element of chance, a random occurrence, is always present in estimating regression coefficients, and some estimates may be far from the true value no matter how good the estimating technique. However, if the distribution is centered on the true value and has as small a variance as possible, the element of chance is less likely to induce a poor estimate. If the sampling distribution is centered around a value other than the true β (that is, if βN is *biased*) then a lower variance implies that most of the sampling distribution of βN is concentrated on the wrong value. However, if this value is not very different from the true value, which is usually not known in practice, then the greater precision will still be valuable.

One method of deciding whether this decreased variance in the distribution of the $\hat{\beta}$ s is valuable enough to offset the bias is to compare different estimation techniques by using a measure called the Mean Square Error (MSE). The *mean square error* is equal to the variance plus the square of the bias. The lower the MSE, the better. A final item of importance is that as the variance of the error term increases, so too does the variance of the distribution of $\hat{\beta}$. The reason for the increased variance of $\hat{\beta}$ is that with the larger variance of u_i , the more extreme values of e_i are observed with more

frequency, and the error term becomes more important in determining the values of y_i .

The standard error of $\hat{\beta}$

Since the standard error of the estimated coefficient, $SE(\hat{\beta})$, is the square root of the estimated variance of the $\hat{\beta}$ s, it is similarly affected by the size of the sample and the other factors we've mentioned. For example, an increase in sample size will cause $SE(\hat{\beta})$ to fall; the larger the sample, the more precise our coefficient estimates will be.

5.3 The Gauss–Markov theorem and the properties of OLS estimators

The Gauss–Markov theorem proves two important properties of OLS estimators. This theorem is proven in all advanced econometrics textbooks, but for a regression user, it's more important to know what the theorem implies than to be able to prove it. The **Gauss–Markov theorem** states that:

Given Classical assumptions 1 through 6 (Assumption 7, normality, is not needed for this theorem), the ordinary least squares estimator of β_k is the minimum variance estimator from among the set of all linear unbiased estimators of β_k , for $k = 0, 1, 2, \dots, K$.

The Gauss–Markov theorem is perhaps most easily remembered by stating that “OLS is BLUE” where BLUE stands for “Best (meaning minimum variance) Linear Unbiased Estimator.” Students who might forget that “best” stands for minimum variance might be better served by remembering “OLS is MvLUE,” but such a phrase is hardly catchy or easy to remember.

If an equation's coefficient estimation is unbiased (that is, if each of the estimated coefficients is produced by an unbiased estimator of the true population coefficient), then:

$$E(\hat{\beta}_k) = \beta_k \quad (k = 0, 1, 2, \dots, K)$$

Best means that each $\hat{\beta}_k$ has the smallest variance possible (in this case, out of all the linear unbiased estimators of β_k). An unbiased estimator with the smallest variance is called *efficient*, and that estimator is said to have the property of **efficiency**. Since the variance typically falls as the sample size increases, larger samples almost always produce more accurate coefficient estimates than do smaller ones.

The Gauss–Markov theorem requires that just the first six of the seven classical assumptions be met. What happens if we add in the seventh assumption, that the error term is normally distributed? In this case, the result of the Gauss–Markov theorem is strengthened because the OLS estimator can be shown to be the best (minimum variance) unbiased estimator out of *all* the possible estimators, not just out of the linear estimators. In other words, if all seven assumptions are met, OLS is “BLUE”.

Given all seven classical assumptions, the OLS coefficient estimators can be shown to have the following properties:

1. *They are unbiased.* That is, $E(\hat{\beta})$ is β . This means that the OLS estimates of the coefficients are centered around the true population values of the parameters being estimated.

2. *They are minimum variance.* The distribution of the coefficient estimates around the true parameter values is as tightly or narrowly distributed as is possible for an unbiased distribution. No other unbiased estimator has a lower variance for each estimated coefficient than OLS.

3. *They are consistent.* As the sample size approaches infinity, the estimates converge to the true population parameters. Put differently, as the sample size gets larger, the variance gets smaller, and each estimate approaches the true value of the coefficient being estimated.¹⁴

4. *They are normally distributed.* The $\hat{\beta}$ s are $N(\beta, \text{var}[\hat{\beta}])$. Thus various statistical tests based on the normal distribution may indeed be applied to these estimates, as will be done in Chapter 6.

¹⁴ Technically, OLS estimates are consistent only if the independent variables continue to fluctuate as the sample size increases. See Halbert White, *Asymptotic Theory for Econometricians* (Orlando: Academic Press, 1984), p. 20.

5.4 Standard econometric notation

This section presents the standard notation used throughout the econometrics literature. Table 5.1 presents various alternative notational devices used to represent the different population (true) parameters and their corresponding estimates (based on samples).

The measure of the central tendency of the sampling distribution of $\hat{\beta}$, which can be thought of as the mean of the $\hat{\beta}$ s, is denoted as $E(\hat{\beta})$, read as “the expected value of beta-hat.” The variance of $\hat{\beta}$ is the typical measure of dispersion of the sampling distribution of $\hat{\beta}$. The variance (or, alternatively, the square root of the variance, called the *standard deviation*) has several alternative notational representations, including $\text{var}(\hat{\beta})$ and $\sigma^2(\hat{\beta})$, read as the “variance of beta-hat.”

The variance of the estimates is a population parameter that is never actually observed in practice; instead, it is estimated with $\hat{\sigma}^2(\hat{\beta}_k)$, also written as $s^2(\hat{\beta}_k)$. Note, by the way, that the variance of the true β , $\sigma^2(\beta)$, is zero, since there is only one true β_k with no distribution around it. Thus, the estimated variance of the estimated coefficient is

Table 5.1 Notation convention

Population Parameter (True Values, but Unobserved)		Estimate (Observed from Sample)	
Name	Symbol(s)	Name	Symbol(s)
Regression coefficient	β_k	Estimated regression coefficient	$\hat{\beta}_k$
Expected value of the estimated coefficient	$E(\hat{\beta}_k)$		
Variance of the error term	σ^2 or $\text{VAR}(\epsilon_i)$	Estimated variance of the error term	s^2 or $\hat{\sigma}^2$
Standard deviation of the error term	σ	An estimate of the standard deviation of the error term	s or SE
Variance of the estimated coefficient	$\sigma^2(\hat{\beta}_k)$ or $\text{VAR}(\hat{\beta}_k)$	Estimated variance of the estimated coefficient	$s^2(\hat{\beta}_k)$ or $\widehat{\text{VAR}}(\hat{\beta}_k)$
Standard deviation of the estimated coefficient	$\sigma_{\hat{\beta}_k}$ or $\sigma(\hat{\beta}_k)$	Standard error of the estimated coefficient	$\hat{\sigma}(\hat{\beta}_k)$ or $\text{SE}(\hat{\beta}_k)$
Error or disturbance term	ϵ_i	Residual (estimate of error in a loose sense)	e_i

defined and observed, the true variance of the estimated coefficient is unobservable, and the true variance of the true coefficient is zero. The square root of the estimated variance of the coefficient estimate is the standard error of $\hat{\beta}$, $SE(\hat{\beta}_k)$, which we will use extensively in hypothesis testing.

SUMMARY

1. The seven Classical assumptions state that the regression model is linear with an additive error term that has a mean of zero, is uncorrelated with the explanatory variables and other observations of the error term, has a constant variance, and is normally distributed (optional). In addition, explanatory variables must not be perfect linear functions of each other.
2. The two most important properties of an estimator are unbiasedness and minimum variance. An estimator is unbiased when the expected value of the estimated coefficient is equal to the true value. Minimum variance holds when the estimating distribution has the smallest variance of all the estimators in a given class of estimators (for example, unbiased estimators).
3. Given the Classical assumptions, OLS can be shown to be the minimum variance, linear, unbiased estimator (or BLUE, for best linear unbiased estimator) of the regression coefficients. This is the Gauss-Markov theorem. When one or more of the classical properties do not hold (excluding normality), OLS is no longer BLUE, although it still may provide better estimates in some cases than the alternative estimation techniques discussed in subsequent chapters.
4. Because the sampling distribution of the OLS estimator of $\hat{\beta}_k$ is BLUE, it has desirable properties. Moreover, the variance, or the measure of dispersion of the sampling distribution of $\hat{\beta}_k$, decreases as the number of observations increases.
5. There is a standard notation used in the econometric literature. Table 5.1 presents this fairly complex set of notational conventions for use in regression analysis. This Table should be reviewed periodically as a refresher.

CHAPTER 6. AN INTRODUCTION TO STATISTICAL INFERENCE

Often, economic theory in general and finance theory in particular, will suggest that certain coefficients should take on particular values, or values within a given range. It is thus of interest to determine whether the relationships expected from economic or financial theory are upheld by the data to hand or not. Estimates of α and β have been obtained from the sample, but these values are not of any particular interest; the population values that describe the true relationship between the variables would be of more interest, but are never available. Instead, inferences are made concerning the likely population values from the regression parameters that have been estimated from the sample of data to hand. In doing this, the aim is to determine whether the differences between the coefficient estimates that are actually obtained, and expectations arising from financial theory, are a long way from one another in a statistical sense.

EXAMPLE 6.1

Suppose the following regression results have been calculated:

$$\hat{y}_t = 20.3 + 0.5091x_t \quad (6.1)$$

$\hat{\beta} = 0.5091$ is a single (point) estimate of the unknown population parameter, β . As stated above, the reliability of the point estimate is measured by the coefficient's standard error. The information from one or more of the sample coefficients and their standard errors can be used to make inferences about the population parameters. So the estimate of the slope coefficient is $\hat{\beta} = 0.5091$, but it is obvious that this number is likely to vary to some degree from one sample to the next. It might be of interest to answer the question, 'Is it plausible, given this estimate, that the true population parameter, β , could be 0.5? Is it plausible that β could be 1?', etc. Answers to these questions can be obtained through *hypothesis testing*.

In this chapter, we return to the essence of econometrics - an effort to quantify economic relationships by analyzing sample data - and ask what conclusions we can draw from this quantification. Hypothesis testing goes beyond calculating estimates of the true population parameters to a much more complex set of questions. Hypothesis testing and statistical inference allow us to answer important questions about the real world from a sample. Is it likely that our result could have been obtained by chance? Would the results generated from our sample lead us to reject our original theories? How confident can we be that our estimate is close to the true value of the parameter?

This chapter starts with a brief introduction to the topic of hypothesis testing. We then examine the t -test, typically used for hypothesis tests of individual regression coefficients. We next look at the confidence interval, a tool for evaluating the precision of our estimates, and we end the chapter by learning how to use the F -test to determine whether whole groups of coefficients affect the dependent variable. Hypothesis testing and the t -test should be familiar topics to readers with strong backgrounds in statistics, who are encouraged to skim this chapter and focus on only those applications that seem somewhat new. The development of hypothesis testing procedures is explained here in terms of the regression model, however, so parts of the chapter may be instructive even to those already skilled in statistics. Our approach will be classical in nature, since we assume that the sample data are our best and only information about the population. An alternative, *Bayesian statistics*, uses a completely different definition of probability and does not use the sampling distribution concept.¹⁵

6.1 What is hypothesis testing?

Hypothesis testing is used in a variety of settings. The Food and Drug Administration (FDA), for example, tests new products before allowing their sale. If the sample of people exposed to the new

¹⁵ Bayesians, by being forced to state explicitly their prior expectations, tend to do most of their thinking before estimation, which is a good habit for a number of important reasons.

product shows some side effect significantly more frequently than would be expected to occur by chance, the FDA is likely to withhold approval of marketing that product. Similarly, economists have been statistically testing various relationships between consumption and income for almost a century; theories developed by John Maynard Keynes and Milton Friedman, among others, have been tested on macroeconomic and microeconomic data sets.

Although researchers are always interested in learning whether the theory in question is supported by estimates generated from a sample of real-world observations, it's almost impossible to *prove* that a given hypothesis is correct. All that can be done is to state that a particular sample conforms to a particular hypothesis. Even though we cannot prove that a given theory is “correct” using hypothesis testing, we often can *reject* a given hypothesis with a certain level of confidence. In such a case, the researcher concludes that it is very unlikely that the sample result would have been observed if the hypothesized theory were correct.

Classical null and alternative hypotheses

The first step in hypothesis testing is to state the hypotheses to be tested. This should be done *before* the equation is estimated because hypotheses developed after estimation run the risk of being justifications of particular results rather than tests of the validity of those results. The **null hypothesis** typically is a statement of the values that the researcher does not expect. The notation used to specify the null hypothesis is “H0:” followed by a statement of the range of values you do not expect. For example, if you expect a positive coefficient, then you don't expect a zero or negative coefficient, and the null hypothesis is:

Null hypothesis $H_0: \beta \leq 0$ (the values you do not expect)

The **alternative hypothesis** typically is a statement of the values that the researcher expects. The notation used to specify the alternative hypothesis is “H1:” followed by a statement of the range of values you expect. To continue our previous example, if you expect a positive coefficient, then the alternative hypothesis is:

Alternative hypothesis $H_1: \beta > 0$ (the values you expect)

To test yourself, take a moment and think about what the null and alternative hypotheses will be if you expect a negative coefficient. That's right, they're:

$$H_0: \beta \geq 0$$

$$H_1: \beta < 0$$

The above hypotheses are for a **one-sided test** because the alternative hypotheses have values on only one side of the null hypothesis. Another approach is to use a **two-sided test** (or a *two-tailed test*) in which the alternative hypothesis has values on both sides of the null hypothesis. For a two-sided test around zero, the null and alternative hypotheses are:

$$H_0: \beta = 0$$

$$H_A: \beta \neq 0$$

We should note that there are a few rare cases in which we must violate our rule that the value you expect goes in the alternative hypothesis. Classical hypothesis testing requires that the null hypothesis contain the equal sign in some form (whether it be $=$, $\leq$, or $\geq$). This requirement means that researchers are forced to put the value they expect in the null hypothesis if their expectation includes an equal sign. This typically happens when the researcher specifies a particular value rather than a range. Luckily, such exceptions are unusual in elementary applications. With the exception of the unusual cases previously mentioned, economists always put what they expect in the alternative hypothesis. This allows us to make rather strong statements when we reject a null hypothesis. However, we can never say that we *accept* the null hypothesis; we must always say that we *cannot reject* the null hypothesis. As put by Jan Kmenta: Just as a court pronounces a

verdict as *not guilty* rather than *innocent*, so the conclusion of a statistical test is *do not reject* rather than *accept*.¹⁶

Type I and type II errors

The typical testing technique in econometrics is to hypothesize an expected sign (or value) for each regression coefficient (except the constant term) and then to determine whether to reject the null hypothesis. Since the regression coefficients are only estimates of the true population parameters, it would be unrealistic to think that conclusions drawn from regression analysis will always be right. There are two kinds of errors we can make in such hypothesis testing:

Type I: We reject a true null hypothesis.

Type II: We do not reject a false null hypothesis.

We will refer to these errors as **type I** and **type II errors**, respectively. Suppose we have the following null and alternative hypotheses:

$$H_0: \beta \leq 0$$

$$H_1: \beta > 0$$

Even if the true parameter β is not positive, the particular estimate obtained by a researcher may be sufficiently positive to lead to the rejection of the null hypothesis that $\beta \leq 0$. This is a type I error; we have rejected the truth! Alternatively, it's possible to obtain an estimate of β that is close enough to zero (or negative) to be considered "not significantly positive." Such a result may lead the researcher to "accept"¹⁷ the hypothesis that $\beta \leq 0$ when in truth $\beta > 0$. This is a type II error; we have failed to reject a false null hypothesis! As an example of type I and type II errors, let's suppose that you're on a jury in a murder case.¹⁸ In such a situation, the presumption of "innocent until proven guilty" implies that:

¹⁶ Jan Kmenta, *Elements of Econometrics* (Ann Arbor: University of Michigan Press, © 1986), p. 112. (Emphasis added.)

¹⁷ We will consistently put the word accept in quotes whenever we use it. In essence, "accept" means *do not reject*.

¹⁸ This example comes from and is discussed in much more detail in Ed Leamer, *Specification Searches* (New York: John Wiley and Sons, 1978), pp. 93–98.

H0: The defendant is innocent.

H1: The defendant is guilty.

What would a type I error be? Rejecting the null hypothesis would mean sending the defendant to jail, so a type I error, rejecting a true null hypothesis, would mean:

Type I error = sending an innocent defendant to jail. Similarly,

Type II error = freeing a guilty defendant.

Most reasonable jury members would want both levels of error to be quite small, but such certainty is almost impossible. After all, couldn't there be a mistaken identification or a lying witness? In the real world, decreasing the probability of a type I error (sending an innocent defendant to jail) means increasing the probability of a type II error (freeing a guilty defendant). If we never sent an innocent defendant to jail, we'd be freeing quite a few murderers!

Decision rules of hypothesis testing

A **decision rule** is a method of deciding whether to reject a null hypothesis. Typically, a decision rule involves comparing a sample statistic with a preselected *critical value* found in Tables such as those in the end of this text. A decision rule should be formulated before regression estimates are obtained. The range of possible values of β_N is divided into two regions, an “*acceptance*” region and a *rejection region*, where the terms are expressed relative to the null hypothesis. To define these regions, we must determine a *critical value* (or, for a two-tailed test, two critical values) of $\hat{\beta}$. Thus, a **critical value** is a value that divides the “acceptance” region from the rejection region when testing a null hypothesis. Graphs of these “acceptance” and rejection regions are presented in Figures 6.1 and 6.2.

To use a decision rule, we need to select a critical value. Let's suppose that the critical value is 1.8. If the observed $\hat{\beta}$ is greater than 1.8, we can reject the null hypothesis that β is zero or negative. To see this, take a look at Figure 5.1. Any $\hat{\beta}$ above 1.8 can be seen

to fall into the rejection region, whereas any $\hat{\beta}$ below 1.8 can be seen to fall into the “acceptance” region.

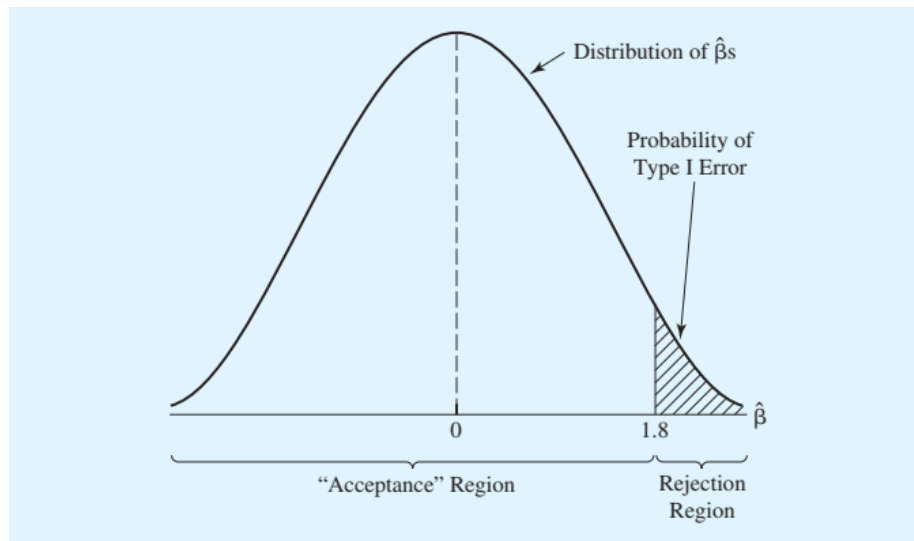


Figure 6.1 “Acceptance” and rejection regions
for a one-sided test of β

For a one-sided test of $H_0: \beta \leq 0$ vs. $H_1: \beta > 0$, the critical value divides the distribution of $\hat{\beta}$ (centered around zero on the assumption that H_0 is true) into “acceptance” and rejection regions.

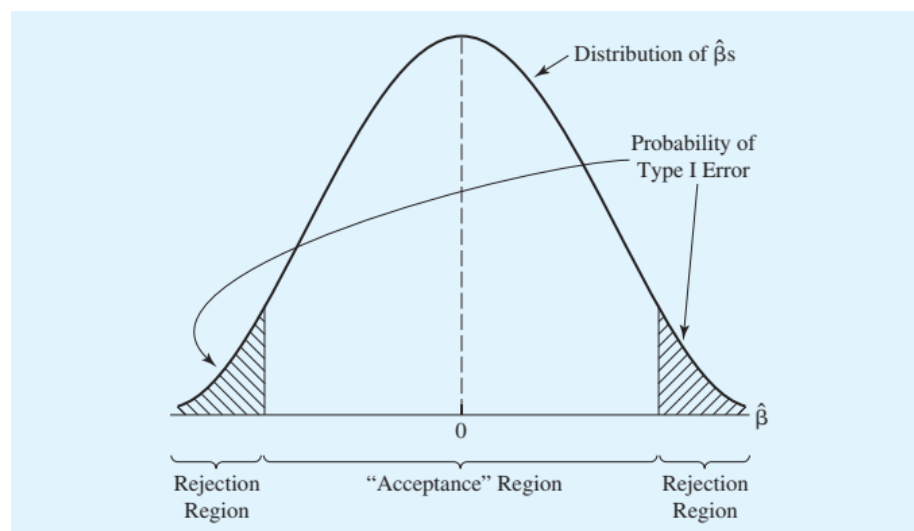


Figure 6.2 “Acceptance” and rejection regions
for a two-sided test of β

For a two-sided test of $H_0: \beta = 0$ vs. $H_1: \beta \neq 0$, we divided the distribution of $\hat{\beta}$ into an “acceptance” region and two rejection regions.

The rejection region measures the probability of a type I error if the null hypothesis is true. Some students react to this news by suggesting that we make the rejection region as small as possible. Unfortunately, decreasing the chance of a type I error means increasing the chance of a type II error (not rejecting a false null hypothesis). If you make the rejection region so small that you almost never reject a true null hypothesis, then you're going to be unable to reject almost every null hypothesis, whether they're true or not! As a result, the probability of a type II error will rise.

Given that, how do you choose between type I and type II errors? The answer is easiest if you know that the cost (to society or the decision maker) of making one kind of error is dramatically larger than the cost of making the other. If you worked for the FDA, for example, you'd want to be very sure that you hadn't released a product that had horrible side effects.

6.2 The t-test

Econometricians generally use the t-test to test hypotheses about individual regression slope coefficients. Tests of more than one coefficient at a time (joint hypotheses) are typically done with the F-test, presented in Section 6.6. The t-test is easy to use because it accounts for differences in the units of measurement of the variables and in the standard deviations of the estimated coefficients. More important, the t-statistic is the appropriate test to use when the stochastic error term is normally distributed and when the variance of that distribution must be estimated. Since these usually are the case, the use of the t-test for hypothesis testing has become standard practice in econometrics.

The t-Statistic

For a typical simple regression equation:

$$y_t = a + \beta_1 x_{1t} + u_t \quad (6.1)$$

we can calculate *t*-values for each of the estimated coefficients in the equation. For reasons that will be explained in Section 7.1, *t*-tests are usually done only on the slope coefficients. The steps involved in doing a test of significance are shown in box 6.1

Box 6.1 Conducting a test of significance

- (1) Estimate $\hat{\alpha}$, $\hat{\beta}$ and $SE(\hat{\alpha})$, $SE(\hat{\beta})$ in the usual way.
- (2) Calculate the test statistic. This is given by the formula

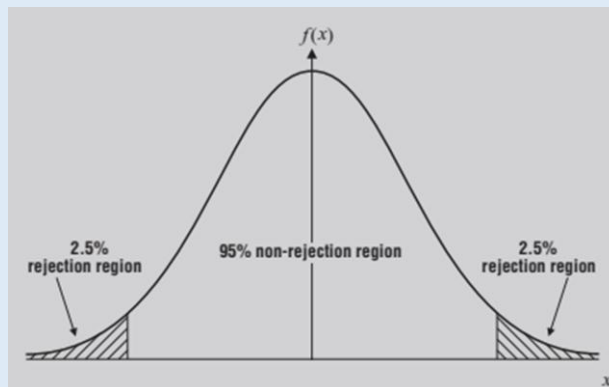
$$\text{test statistics} = \frac{\hat{\beta} - \beta^*}{SE(\hat{\beta})} \quad (6.2)$$

where $\hat{\beta}$ is the value of $\hat{\alpha}$ under the null hypothesis. The null hypothesis is $H_0: \beta = \beta^*$ and the alternative hypothesis is $H_1: \beta \neq \beta^*$ (for a two-sided test).

- (3) A tabulated distribution with which to compare the estimated test statistics is required. Test statistics derived in this way can be shown to follow a t -distribution with $T - 2$ degrees of freedom.

- (4) Choose a ‘significance level’, often denoted α (*not* the same as the regression intercept coefficient). It is conventional to use a significance level of 5%.

- (5) Given a significance level, a *rejection region* and *non-rejection region* can be determined. If a 5% significance level is employed, this means that 5% of the total distribution (5% of the area under the curve) will be in the rejection region. That rejection region can either be split in half (for a two-sided test) or it can all fall on one side of the y -axis, as is the case for a one-sided test. For a two-sided test, the 5% rejection region is split equally between the two tails, as shown in figure 6.1. For a one-sided test, the 5% rejection region is located solely in one tail of the distribution, as shown in figures 6.2 and 6.3, for a test where the alternative is of the ‘less than’ form, and where the alternative is of the ‘greater than’ form, respectively.



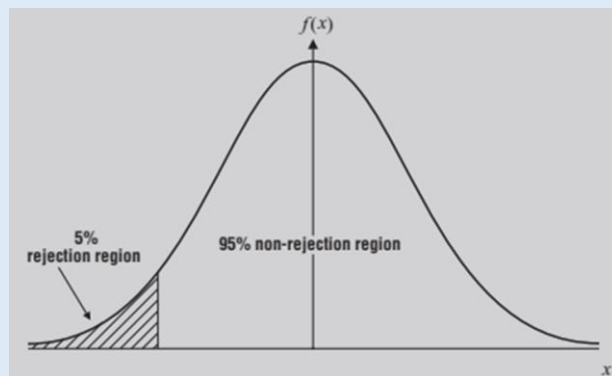


Figure 6.2 Rejection region for a one-sided hypothesis test of the form

$$H_0: \beta = \beta^*, H_1: \beta < \beta^*$$

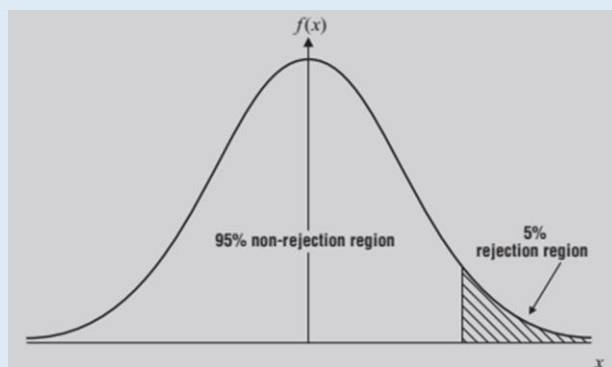


Figure 6.3 Rejection region for a one-sided hypothesis test of the form

$$H_0: \beta = \beta^*, H_1: \beta > \beta^*$$

(6) Use the t -tables to obtain a critical value or values with which to compare the test statistic. The critical value will be that value of x that puts 5% into the rejection region.

(7) Finally perform the test. If the test statistic lies in the rejection region then reject the null hypothesis (H_0), else do not reject H_0 .

Steps 2-7 require further comment. In step 2, the estimated value of β is compared with the value that is subject to test under

the null hypothesis, but this difference is ‘normalised’ or scaled by the standard error of the coefficient estimate. The standard error is a measure of how confident one is in the coefficient estimate obtained in the first stage. If a standard error is small, the value of the test statistic will be large relative to the case where the standard error is large. For a small standard error, it would not require the estimated and hypothesised values to be far away from one another for the null hypothesis to be rejected. Dividing by the standard error also ensures that, under the *CLRM* assumptions, the test statistic follows a tabulated distribution.

In this context, the number of degrees of freedom can be interpreted as the number of pieces of additional information beyond the minimum requirement. If two parameters are estimated (α and β - the intercept and the slope of the line, respectively), a minimum of two observations is required to fit this line to the data. As the number of degrees of freedom increases, the critical values in the Tables decrease in absolute terms, since less caution is required and one can be more confident that the results are appropriate.

The significance level is also sometimes called the *size of the test* (note that this is completely different from the size of the sample) and it determines the region where the null hypothesis under test will be rejected or not rejected. Remember that the distributions in Figures 6.1-6.3 are for a random variable. Purely by chance, a random variable will take on extreme values (either large and positive values or large and negative values) occasionally. More specifically, a significance level of 5% means that a result as extreme as this or more extreme would be expected only 5% of the time as a consequence of chance alone. To give one illustration, if the 5% critical value for a one-sided test is 1.68, this implies that the test statistic would be expected to be greater than this only 5% of the time by chance alone. There is nothing magical about the test -- all that is done is to specify an arbitrary cutoff value for the test statistic that determines whether the null hypothesis would be rejected or not. It is conventional to use a 5% size of test, but 10% and 1% are also commonly used.

However, one potential problem with the use of a fixed (e.g. 5%) size of test is that if the sample size is sufficiently large, any null hypothesis can be rejected. This is particularly worrisome in

finance, where tens of thousands of observations or more are often available. What happens is that the standard errors reduce as the sample size increases, thus leading to an increase in the value of all t -test statistics. This problem is frequently overlooked in empirical work, but some econometricians have suggested that a lower size of test (e.g. 1%) should be used for large samples.

Note also the use of terminology in connection with hypothesis tests: it is said that the null hypothesis is either *rejected* or *not rejected*. It is incorrect to state that if the null hypothesis is not rejected, it is ‘accepted’ (although this error is frequently made in practice), and it is never said that the alternative hypothesis is accepted or rejected. One reason why it is not sensible to say that the null hypothesis is ‘accepted’ is that it is impossible to know whether the null is actually true or not! In any given situation, many null hypotheses will not be rejected. For example, suppose that $H_0: \beta = 0.5$ and $H_0: \beta = 1$ are separately tested against the relevant two-sided alternatives and neither null is rejected. Clearly then it would not make sense to say that ‘ $H_0: \beta = 0.5$ is accepted’ and ‘ $H_0: \beta = 1$ is accepted’, since the true (but unknown) value of β cannot be both 0.5 and 1. So, to summarise, the null hypothesis is either rejected or not rejected on the basis of the available evidence.

The confidence interval approach to hypothesis testing

To decide whether to reject or not to reject a null hypothesis based on a calculated t -value, we use a critical t -value. A *critical t -value* is the value that distinguishes the “acceptance” region from the rejection region. The critical t -value, t_c , is selected from a t -Table (see Statistical Table B-1 in the back of the book) depending on whether the test is one-sided or two-sided, on the level of type I error you specify, and on the degrees of freedom, which we have defined as the number of observations minus the number of coefficients estimated (including the constant) or $N - K - 1$. The level of type I error in a hypothesis test is also called the *level of significance* of that test and will be discussed in more detail later in this section. The t -Table was created to save time during research; it consists of critical t -values given specific areas underneath curves such as those in Figure 6.4 below for type I errors. A critical t -value is thus a function of the probability of type I error that the researcher wants to specify.

Once you have obtained a calculated t -value t_k and a critical t -value t_{crit} , you reject the null hypothesis if the calculated t -value is greater in absolute value than the critical t -value and if the calculated t -value has the sign implied by H_1 . Thus, the rule to apply when testing a single regression coefficient is that you should:

**Reject H_0 if $|t_k| > t_c$ and if t_k also has the sign implied by H_1 .
Do not reject H_0 otherwise.**

To give an example of its usage, one might estimate a parameter, say $\hat{\beta}$, to be 0.93, and a ‘95% confidence interval’ to be (0.77, 1.09). This means that in many repeated samples, 95% of the time, the true value of β will be contained within this interval. Confidence intervals are almost invariably estimated in a two-sided form, although in theory a one-sided interval can be constructed. Constructing a 95% confidence interval is equivalent to using the 5% level in a test of significance.

Box 6.2 Carrying out a hypothesis test using confidence intervals

- (1) Calculate $\hat{a}$, $\hat{\beta}$ and $SE(\hat{a})$, $SE(\hat{\beta})$ as before.
- (2) Choose a significance level, α (again the convention is 5%). This is equivalent to choosing a $(1-\alpha) \hat{a}$ 100% confidence interval i.e.

5% significance level = 95% confidence interval

- (3) Use the t -tables to find the appropriate critical value, which will again have $T-2$ degrees of freedom.
- (4) The confidence interval for β is given by $(\hat{\beta} - t_{crit} \times SE(\hat{\beta}), \hat{\beta} + t_{crit} \times SE(\hat{\beta}))$.
- (5) Perform the test: if the hypothesized value of β (i.e. $\hat{a}$) lies outside the confidence interval, then reject the null hypothesis that $\beta = \beta^*$, otherwise do not reject the null.

Rearranging, the null hypothesis would not be rejected if

$$-t_{crit} \times SE(\hat{\beta}) \leq \hat{\beta} - \beta^* \leq +t_{crit} \times SE(\hat{\beta})$$

i.e. one would not reject if

$$\hat{\beta} - t_{crit} \times SE(\hat{\beta}) \leq \beta^* \leq \hat{\beta} + t_{crit} \times SE(\hat{\beta})$$

But this is just the rule for non-rejection under the confidence interval approach. So it will always be the case that, for a given significance level, the test of significance and confidence interval approaches will provide the same conclusion by construction. One testing approach is simply an algebraic rearrangement of the other.

EXAMPLE 6.2

Given the regression results above:

$$\hat{y}_t = 20.3 + 0.5091x_t \quad T=22 \quad (6.3)$$

(14.38) (0.2561)

Using both the test of significance and confidence interval approaches, test the hypothesis that $\beta = 1$ against a two-sided alternative. This hypothesis might be of interest, for a unit coefficient on the explanatory variable implies a 1:1 relationship between movements in x and movements in y . The null and alternative hypotheses are respectively:

$$H0: \beta = 1$$

$$H1: \beta \neq 1$$

In equation 6.3, the numbers in parentheses underneath the estimated regression coefficients are the estimated standard errors of the estimated $\hat{\beta}$ s, and the numbers below them are t -values calculated according to Equation 6.2. The format used to document equation 6.2 is the one we'll use whenever possible throughout this text. Note that the sign of the t -value is always the same as that of the estimated regression coefficient, and the standard error is always positive.

The results of the test according to each approach are shown in box 6.3. A couple of comments are in order. First, the critical value

from the t -distribution that is required is for 20 degrees of freedom and at the 5% level. This means that 5% of the total distribution will be in the rejection region, and since this is a two-sided test, 2.5% of the distribution is required to be contained in each tail. From the symmetry of the t distribution around zero, the critical values in the upper and lower tail will be equal in magnitude, but opposite in sign, as shown in Figure 6.4

Box 6.3 The test of significance and confidence interval approaches compared

Test of significance approach

$$test\ stat = \frac{\hat{\beta} - \beta^*}{SE(\hat{\beta})} = \frac{0.5091 - 1}{0.2561} = -1.917$$

Find $t_{crit} = t_{20;5\%} = \pm 2.086$

Do not reject H_0 since test statistics lies within non-rejection region

Confidence interval approach

Find $t_{crit} = t_{20;5\%} = \pm 2.086$

$$\begin{aligned} \hat{\beta} \pm t_{crit} \times SE(\hat{\beta}) &= \\ &= 0.5091 \pm 2.086 \times 0.2561 \\ &= (-0.0251, 1.0433) \end{aligned}$$

Do not reject H_0 since test statistics lies within non-rejection region

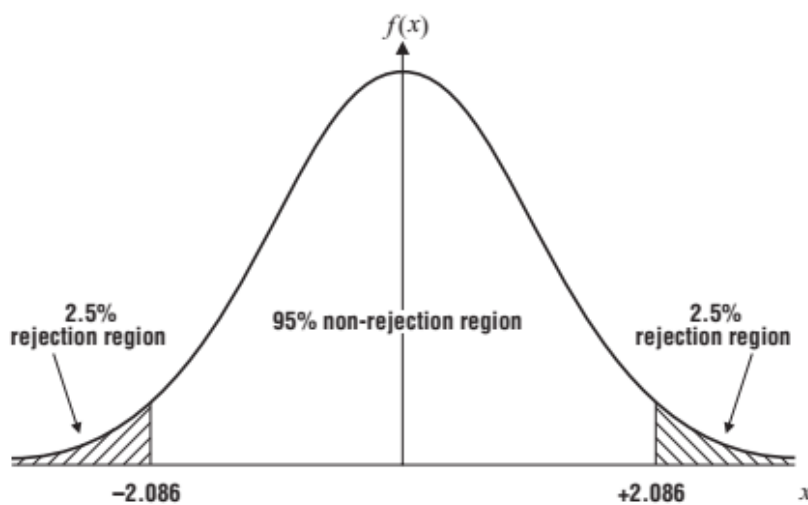


Figure 6.4 Critical values and rejection regions for a $t_{20;5\%}$.

Statistical Table B-1 in the Appendix of tutorial *Econometric Analysis using Eviews: a practical tutorial* contains the critical values t_{crit} for varying degrees of freedom and levels of significance. The columns indicate the levels of significance according to whether the test is one-sided or two-sided, and the rows indicate the degrees of freedom.

For an example of the use of this Table and the decision rule, let's return to the Rohat's restaurant example and, in particular, to the t -value for $\hat{\beta}_p$ calculated in the previous section. Recall that we hypothesized that population's coefficient would be positive, so this is a one-sided test:

$$H_0 : \beta_p \leq 0$$

$$H_1 : \beta_p > 0$$

There are 29 degrees of freedom (equal to $N - K - 1$, or $33 - 3 - 1$) in this regression, so the appropriate t -value with which to test the calculated t -value is a one-tailed critical t -value with 29 degrees of freedom. To find this value, pick a level of significance, say 5%, and turn to Statistical Table B-1.¹⁹ Take a look for yourself. Do you agree that the number there is 1.699? Given that, should you reject the null hypothesis? The decision rule is to reject H_0 if $|t_k| > t_c$ and if t_k has the sign implied by H_1 . Since the 5%, one-sided, 29 degrees of freedom critical t -value is 1.699, and since the sign implied by H_1 is positive, the decision rule (for this specific case) becomes:

Reject H_0 if $|t_p| > 1.699$ and if t_p is positive

or, combining the two conditions:

Reject H_0 if $t_p > 1.699$

What is t_p ? In the previous section, we found that t_p was +4.88, so we would reject the null hypothesis and conclude that population does indeed tend to have a positive relationship with Rohat's check volume (holding the other variables in the equation constant).

¹⁹ Khakimova M.F. *Econometric Analysis using Eviews: Practical tutorial*. Dushanbe, 2020. – 208p.

Note from Statistical Table B-1 that the critical t -value for a one-tailed test at a given level of significance is exactly equal to the critical t -value for a two-tailed test at twice the level of significance as the one-tailed test. This relationship between one-sided and two-sided tests is illustrated in Figure 6.5.

The critical value $t_c = 1.699$ is for a one-sided, 5% level of significance, but it also represents a two-sided, 10% of significance because if one tail represents 5%, then both tails added together represent 10 %.

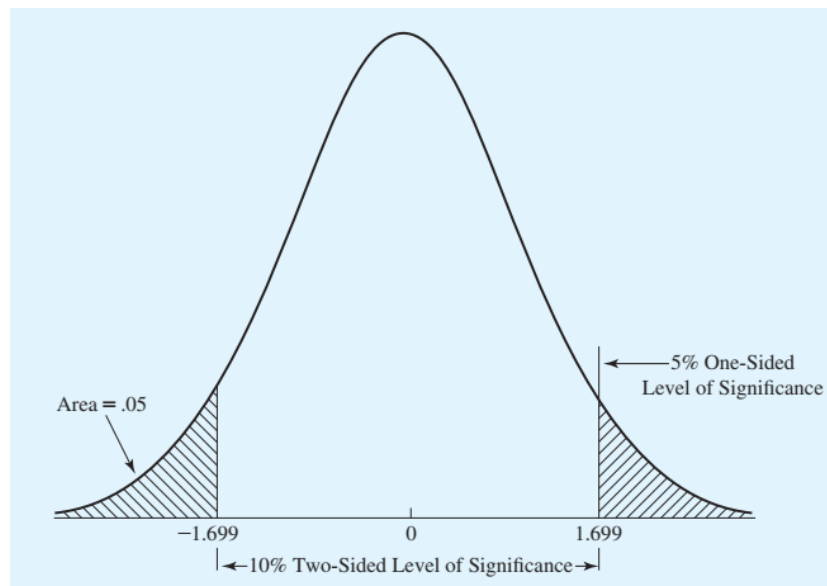


Figure 6.5 One-sided and two-sided t -test

What if instead the researcher wanted to test $H_0 : \beta = 0$ or $H_0 : \beta = 2$? In order to test these hypotheses using the test of significance approach, the test statistic would have to be reconstructed in each case, although the critical value would be the same. On the other hand, no additional work would be required if the confidence interval approach had been adopted, since it effectively permits the testing of an infinite number of hypotheses.

So for example, suppose that the researcher wanted to test

$H_0 : \beta = 0$

Versus

$H_1 : \beta \neq 0$

And

$H_0 : \beta = 2$

Versus

$H_1 : \beta \neq 2$

In the first case, the null hypothesis (that $\beta=0$) would not be rejected since 0 lies within the 95% confidence interval. By the same argument, the second null hypothesis (that $\beta=2$) would be rejected since 2 lies outside the estimated confidence interval.

On the other hand, note that this book has so far considered only the results under a 5% size of test. In marginal cases (e.g. $H_0: \beta = 1$, where the test statistic and critical value are close together), a completely different answer may arise if a different size of test was used. This is where the test of significance approach is preferable to the construction of a confidence interval.

For example, suppose that now a 10% size of test is used for the null hypothesis given in example 6.2. Using the test of significance approach,

$$\begin{aligned} \text{test statistic} &= \frac{\hat{\beta} - \beta^*}{SE(\hat{\beta})} = \\ &= \frac{0.5091 - 1}{0.2561} = -1.917 \end{aligned}$$

as above. The only thing that changes is the critical t -value. At the 10% level (so that 5% of the total distribution is placed in each of the tails for this two-sided test), the required critical value is $t_{20;10\%} = \pm 1.725$. So now, as the test statistic lies in the rejection region, H_0 would be rejected. In order to use a 10% test under the confidence interval approach, the interval itself would have to have been re-estimated since the critical value is embedded in the calculation of the confidence interval.

So the test of significance and confidence interval approaches both have their relative merits. The testing of a number of different hypotheses is easier under the confidence interval approach, while a consideration of the effect of the size of the test on the conclusion is easier to address under the test of significance approach.

Caution should therefore be used when placing emphasis on or making decisions in the context of marginal cases (i.e. in cases where the null is only just rejected or not rejected). In this situation, the appropriate conclusion to draw is that the results are marginal and that no strong inference can be made one way or the other. A thorough empirical analysis should involve conducting a sensitivity analysis on the results to determine whether using a different size of test alters the conclusions. It is worth stating again that it is conventional to consider sizes of test of 10%, 5% and 1%. If the conclusion (i.e. 'reject' or 'do not reject') is robust to changes in the size of the test, then one can be more confident that the conclusions are appropriate. If the outcome of the test is qualitatively altered when the size of the test is modified, the conclusion must be that there is no conclusion one way or the other!

It is also worth noting that if a given null hypothesis is rejected using a 1% significance level, it will also automatically be rejected at the 5% level, so that there is no need to actually state the latter. Here example with a high jumper can be consider. If the high jumper can clear 2 metres, it is obvious that the jumper could also clear 1.5 metres. The 1% significance level is a higher hurdle than the 5% significance level. Similarly, if the null is not rejected at the 5% level of significance, it will automatically not be rejected at any stronger level of significance (e.g. 1%). In this case, if the jumper cannot clear 1.5 metres, there is no way s/he will be able to clear 2 metres.

The exact significance level is also commonly known as the p -value. It gives the *marginal significance level* where one would be indifferent between rejecting and not rejecting the null hypothesis. If the test statistic is 'large' in absolute value, the p -value will be small, and vice versa. For example, consider a test statistic that is distributed as a t_{62} and takes a value of 1.47. Would the null hypothesis be rejected? It would depend on the size of the test. Now, suppose that the p -value for this test is calculated to be 0.12:

• Is the null rejected at the 5% level?	No
• Is the null rejected at the 10% level?	No
• Is the null rejected at the 20% level?	Yes

In fact, the null would have been rejected at the 12% level or higher. To see this, consider conducting a series of tests with size 0.1%, 0.2%, 0.3%, 0.4%, . . . 1%, . . . , 5%, . . . 10%, . . . Eventually, the critical value and test statistic will meet and this will be the p -value. p -values are almost always provided automatically by software packages. Note how useful they are! They provide all of the information required to conduct a hypothesis test without requiring of the researcher the need to calculate a test statistic or to find a critical value from a Table - both of these steps have already been taken by the package in producing the p -value. The p -value is also useful since it avoids the requirement of specifying an arbitrary significance level (α). Sensitivity analysis of the effect of the significance level on the conclusion occurs automatically.

Informally, the p -value is also often referred to as the probability of being wrong when the null hypothesis is rejected. Thus, for example, if a p -value of 0.05 or less leads the researcher to reject the null (equivalent to a 5% significance level), this is equivalent to saying that if the probability of incorrectly rejecting the null is more than 5%, do not reject it. The p -value has also been termed the ‘plausibility’ of the null hypothesis; so, the smaller is the p -value, the less plausible is the null hypothesis.

Choosing a level of significance

To complete the previous example, it was necessary to pick a level of significance before a critical t -value could be found in Statistical Table B-1. The words “significantly positive” usually carry the statistical interpretation that H_0 ($\beta \leq 0$) was rejected in favor of H_1 ($\beta > 0$) according to the preestablished decision rule, which was set up with a given level of significance. The **level of significance** indicates the probability of observing an estimated t -value greater than the critical t -value if the null hypothesis were correct. It measures the amount of type I error implied by a particular critical t -value. If the level of significance is 10% and we reject the null hypothesis at that level, then this result would have occurred only 10% of the time that the null hypothesis was indeed correct.

How should you choose a level of significance? Most beginning econometricians (and many published ones, too) assume

that the lower the level of significance, the better. After all, they say, doesn't a low level of significance guarantee a low probability of making a type I error? Unfortunately, an extremely low level of significance also dramatically increases the probability of making a type II error. Therefore, unless you're in the unusual situation of not caring about mistakenly "accepting" a false null hypothesis, minimizing the level of significance is *not* good standard practice.

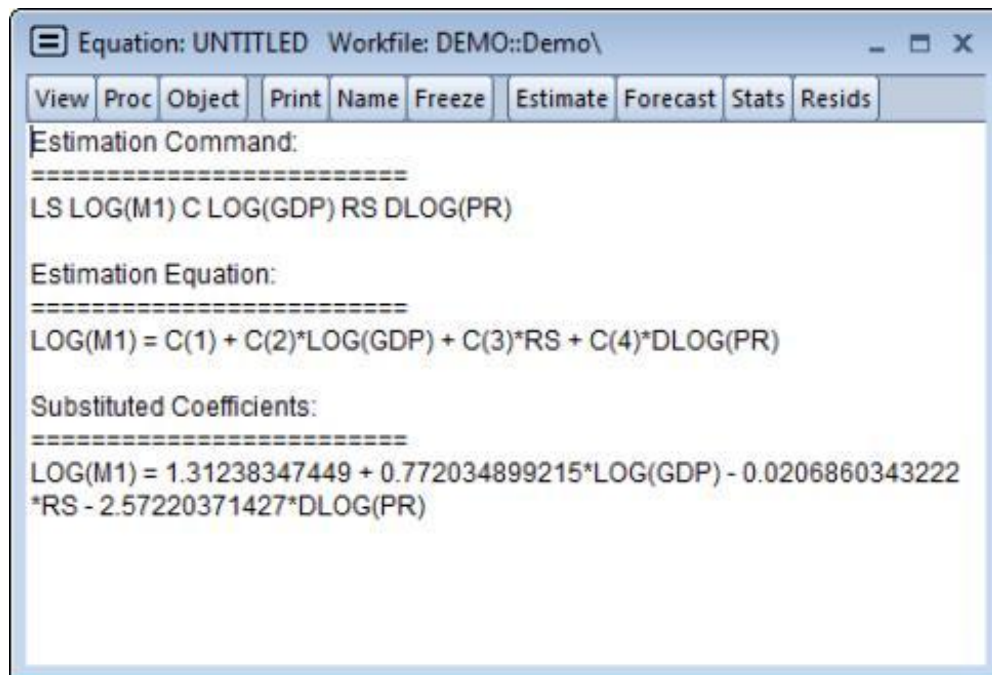
Instead, we recommend using a 5% level of significance except in those circumstances when you know something unusual about the relative costs of making type I and type II errors. If you know that a type II error will be extremely costly, for example, then it makes sense to consider using a 10% level of significance when you determine your critical value. Such judgments are difficult, however, so we encourage beginning researchers to adopt a 5% level of significance as standard.

If we can reject a null hypothesis at the 5% level of significance, we can summarize our results by saying that the coefficient is "statistically significant" at the 5% level. Since the 5% level is arbitrary, we shouldn't jump to conclusions about the value of a variable simply because its coefficient misses being significant by a small amount; if a different level of significance had been chosen, the result might have been different.

Some researchers produce Tables of regression results, typically without hypothesized signs for their coefficients, and then mark "significant" coefficients with asterisks. The asterisks indicate when the t -score is larger in absolute value than the two-sided 10% critical value (which merits one asterisk), the two-sided 5% critical value (**), or the two-sided 1% critical value (***). Such a use of the t -value should be regarded as a descriptive rather than a hypothesis-testing use of statistics.

6.3 Hypothesis testing in Eviews

We can use the estimated equation to perform hypothesis tests on the coefficients of the model. For example, to test the hypothesis that the coefficient on the price term is equal to 2, we will perform a Wald test. First, determine the coefficient of interest by selecting **View/Representations** from the equation toolbar:



Screenshot 6.1 Equation representation.

Note that the coefficients are assigned in the order that the variables appear in the specification so that the coefficient for the PR term is labeled C(4). To test the restriction on C(4) you should select **View/Coefficient Diagnostics/Wald Test Coefficient Restrictions...**, and enter the restriction “c(4)=2”. Eviews will report the results of the Wald test:

Wald Test:
Equation: UNTITLED
Null Hypothesis: C(4)=2

Test Statistic	Value	df	Probability
t-statistic	-4.850856	159	0.0000
F-statistic	23.53081	(1, 159)	0.0000
Chi-square	23.53081	1	0.0000

Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
-2 + C(4)	-4.572204	0.942556

Restrictions are linear in coefficients.

Screenshot 6.2 Results of Wald Test.

The low probability values indicate that the null hypothesis that $C(4)=2$ is strongly rejected. We should, however, be somewhat cautious of accepting this result without additional analysis. The low value of the Durbin-Watson statistic reported above is indicative of the presence of serial correlation in the residuals of the estimated equation. If uncorrected, serial correlation in the residuals will lead to incorrect estimates of the standard errors, and invalid statistical inference for the coefficients of the equation.

The Durbin-Watson statistic can be difficult to interpret. To perform a more general Breusch-Godfrey test for serial correlation in the residuals, select **View/Residual Diagnostics/Serial Correlation LM Test...** from the equation toolbar, and specify an order of serial correlation to test against. Entering “1” yields a test against first-order serial correlation:

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	813.0060	Prob. F(1,158)	0.0000
Obs*R-squared	136.4770	Prob. Chi-Square(1)	0.0000

Test Equation:

Dependent Variable: RESID

Method: Least Squares

Date: 08/18/09 Time: 14:10

Sample: 1952Q2 1992Q4

Included observations: 163

Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.006355	0.013031	-0.487683	0.6265
LOG(GDP)	0.000997	0.002645	0.376929	0.7067
RS	-0.000567	0.001018	-0.556748	0.5785
DLOG(PR)	0.404143	0.381676	1.058864	0.2913
RESID(-1)	0.920306	0.032276	28.51326	0.0000
R-squared	0.837282	Mean dependent var	1.23E -15	
Adjusted R-squared	0.833163	S.D. dependent var	0.054969	
S.E. of regression	0.022452	Akaike info criterion	-4.724644	
Sum squared resid	0.079649	Schwarz criterion	-4.629744	
Log likelihood	390.0585	Hannan-Quinn criter.	-4.686116	
F-statistic	203.2515	Durbin-Watson stat	1.770965	
Prob(F-statistic)	0.000000			

Screenshot 6.3 Breusch-Godfrey Test.

The top part of the output presents the test statistics and associated probability values. The test regression used to carry out the test is reported below the statistics. The statistic labeled

“Obs*R-squared” is the LM test statistic for the null hypothesis of no serial correlation. The (effectively) zero probability value strongly indicates the presence of serial correlation in the residuals.

Lets consider another example of Hypotheses testing. This exercise will estimate and test some hypotheses about the CAPM beta for several US stocks. First, **Open a new workfile** to accommodate monthly data commencing in January 2002 and ending in April 2007. Then **import the Excel file ‘capm.xls’**. The file is organised by observation and contains six columns of numbers plus the dates in the first column, so in the ‘Names for series or Number if named in file’ box, type 6. As before, do not import the dates so the data start in cell B2. The monthly stock prices of four companies (Ford, General Motors, Microsoft and Sun) will appear as objects, along with index values for the S&P500 (‘sandp’) and three-month US-Treasury bills (‘ustb3m’). **Save the Eviews workfile as ‘capm.wk1’**.

In order to estimate a CAPM equation for the Ford stock, for example, we need to first transform the price series into returns and then the excess returns over the risk free rate. To transform the series, click on the Generate button (**Genr**) in the workfile window. In the new window, type

$$RSANDP=100*\text{LOG}(SANDP/SANDP(-1))$$

This will create a new series named RSANDP that will contain the returns of the S&P500. The operator (-1) is used to instruct Eviews to use the oneperiod lagged observation of the series. To estimate percentage returns on the Ford stock, press the **Genr** button again and type

$$RFORD=100*\text{LOG}(FORD/FORD(-1))$$

This will yield a new series named RFORD that will contain the returns of the Ford stock. Eviews allows various kinds of transformations to the series. For example:

X2=X/2	creates a new variable called X2 that is half of X
XSQ=X^2	creates a new variable XSQ that is X squared
LX=LOG(X)	creates a new variable LX that is the log of X
LAGX=X(-1)	creates a new variable LAGX containing X lagged by one period creates a new variable
LAGX2=X(-2)	LAGX2 containing X lagged by two periods
Other functions include:	
d(X)	first difference of X
d(X,n)	nth order difference of X
dlog(X)	first difference of the logarithm of X
dlog(X,n)	nth order difference of the logarithm of X
abs(X)	absolute value of X

If, in the transformation, the new series is given the same name as the old series, then the old series will be overwritten. Note that the returns for the S&P index could have been constructed using a simpler command in the 'Genr' window such as

RSANDP=100*DLOG(SANDP)

Before we can transform the returns into excess returns, we need to be slightly careful because the stock returns are monthly, but the Treasury bill yields are annualised. We could run the whole analysis using monthly data or using annualized data and it should not matter which we use, but the two series must be measured consistently. So, to turn the T-bill yields into monthly Figures and to write over the original series, press the **Genr** button again and type

USTB3M=USTB3M/12

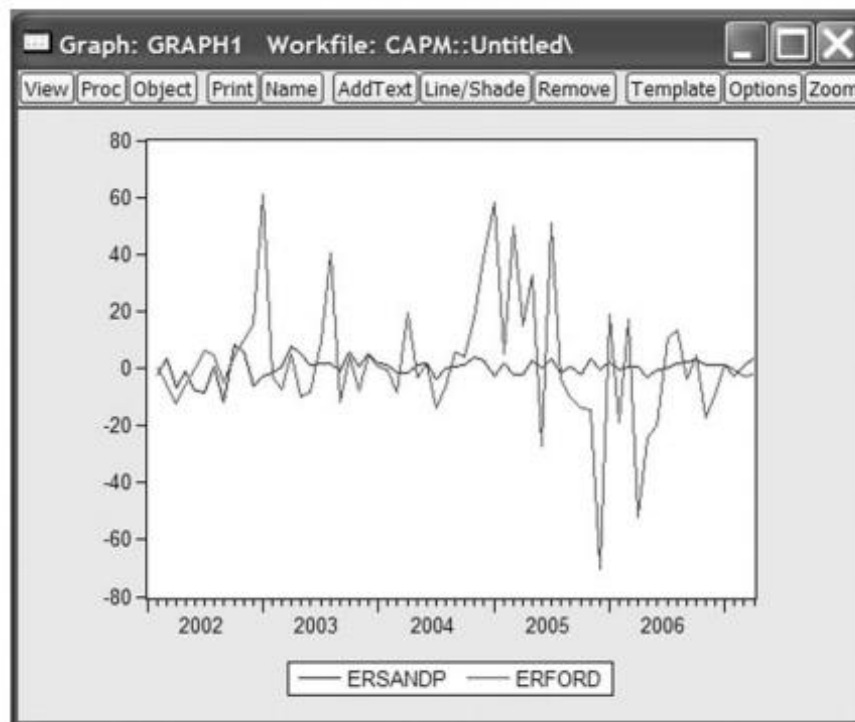
Now, to compute the excess returns, click **Genr** again and type

ERSANDP=RSANDP-USTB3M

where ‘ERSANDP’ will be used to denote the excess returns, so that the original raw returns series will remain in the workfile. The Ford returns can similarly be transformed into a set of excess returns. Now that the excess returns have been obtained for the two series, before running the regression, plot the data to examine visually whether the series appear to move together. To do this, create a new object by clicking on the **Object/New Object** menu on the menu bar. Select **Graph**, provide a name (call the graph **Graph1**) and then in the new window provide the names of the series to plot. In this new window, type

ERSANDP ERFORD

Then press **OK** and screenshot 6.4 will appear.



Screenshot 6.4 Plot of two series.

This is a time-series plot of the two variables, but a scatter plot may be more informative. To examine a scatter plot, Click **Options**, choose the **Type** tab, then select **Scatter** from the list and click **OK**. There appears to be a weak association between ERFTAS and

ERFORD. Close the window of the graph and return to the workfile window.

To estimate the CAPM equation, click on **Object/New Objects**. In the new window, select **Equation** and name the object **CAPM**. Click on **OK**. In the window, specify the regression equation. The regression equation takes the form

$$(R_{Ford} - r_f)_t = a + \beta(R_M - r_f)_t + u_t$$

Since the data have already been transformed to obtain the excess returns, in order to specify this regression equation, type in the equation window

ERFORD C ERSANDP

To use all the observations in the sample and to estimate the regression using LS -Least Squares (NLS and ARMA), click on **OK**. The results screen appears as in the screenshot below. Make sure that you **save the Workfile** again to include the transformed series and regression results!

Dependent Variable: ERFORD
Method: Least Squares
Date: 08/21/07 Time: 15:02
Sample (adjusted): 2002M02 2007M04
Included observations: 63 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
C	2.020219	2.801382	0.721151	0.4736
ERSANDP	0.359726	0.794443	0.452803	0.6523
R-squared	0.003350	Mean dependent var		2.097445
Adjusted R-squared	-0.012989	S.D. dependent var		22.05129
S.E. of regression	22.19404	Akaike info criterion		9.068756
Sum squared resid	30047.09	Schwarz criterion		9.136792
Log likelihood	-283.6658	Hannan-Quinn criter.		9.095514
F-statistic	0.205031	Durbin-Watson stat		1.785699
Prob(F-statistic)	0.652297			

Screenshot 6.5 CAPM regression results

Take a couple of minutes to examine the results of the regression. What is the slope coefficient estimate and what does it

signify? Is this coefficient statistically significant? The beta coefficient (the slope coefficient) estimate is 0.3597. The p -value of the t -ratio is 0.6523, signifying that the excess return on the market proxy has no significant explanatory power for the variability of the excess returns of Ford stock. What is the interpretation of the intercept estimate? Is it statistically significant?

In fact, there is a considerably quicker method for using transformed variables in regression equations, and that is to write the transformation directly into the equation window. In the CAPM example above, this could be done by typing

DLOG(FORD)-USTB3M C DLOG(SANDP)-USTB3M

into the equation window. As well as being quicker, an advantage of this approach is that the output will show more clearly the regression that has actually been conducted, so that any errors in making the transformations can be seen more clearly.

How could the hypothesis that the value of the population coefficient is equal to 1 be tested? The answer is to click on **View/Coefficient Tests/Wald – Coefficient Restrictions**. . . and then in the box that appears, **Type C(2)=1**. The conclusion here is that the null hypothesis that the CAPM beta of Ford stock is 1 cannot be rejected and hence the estimated beta of 0.359 is not significantly different from 1.²⁰

6.4 Examples of t -tests

Examples of one-sided t -tests

The most common use of the one-sided t -test is to determine whether a regression coefficient is significantly different from zero in the direction predicted by theory. Let's face it: if you expect a positive sign for a coefficient and you get a negative $\hat{\beta}$, it's hard to reject the possibility that the true β might be negative (or zero). On the other hand, if you expect a positive sign and get a positive $\hat{\beta}$,

²⁰ Although the value 0.359 may seem a long way from 1, considered purely from an econometric perspective, the sample size is quite small and this has led to a large parameter standard error, which explains the failure to reject both $H_0 : \beta = 0$ and $H_0 : \beta = 1$.

things get a bit tricky. If $\hat{\beta}$ is positive but fairly close to zero, then a one-sided t -test should be used to determine whether the $\hat{\beta}$ is different enough from zero to allow the rejection of the null hypothesis. Recall that in order to be able to control the amount of type I error we make, such a theory implies an alternative hypothesis of $H_1: \beta > 0$ (the expected sign) and a null hypothesis of $H_0: \beta \leq 0$. Let's look at some complete examples of these kinds of one-sided t -tests. Consider a simple model of the aggregate annual retail sales of new cars that specifies that sales of new cars (CARS) are a function of real disposable income (YD) and the average retail price of a car adjusted by the consumer price index (PRICE). Suppose you spend some time reviewing the literature on the automobile industry and are inspired to test a new theory. You decide to add a third independent variable, the number of sports utility vehicles sold (SUV) to take account of the fact that some potential new car buyers purchase SUVs instead. You therefore hypothesize the following model:

$$CARS = a + \beta_1^+ YD + \beta_2^- PRICE + \beta_3^- SUV + u \quad (6.4)$$

As you can see from the hypothesized signs above the coefficients in equation 6.4, you expect β_1 to be positive and β_2 and β_3 to be negative. This makes sense, since you'd expect higher incomes, lower prices, or lower sales of SUVs to increase new car sales, holding the other variables in the equation constant. The four steps to use when working with the t -test are:

1. Set up the null and alternative hypotheses.
2. Choose a level of significance and therefore a critical t -value.
3. Run the regression and obtain an estimated t -value (or t -score).
4. Apply the decision rule by comparing the calculated t -value with the critical t -value in order to reject or not reject the null hypothesis.

Let's look at each step in more detail.

1. *Set up the null and alternative hypotheses.*²¹ From equation 6.4, the onesided hypotheses are set up as:

1. $H_0: \beta_1 \leq 0$
 $H_1: \beta_1 > 0$
2. $H_0: \beta_2 \geq 0$
 $H_1: \beta_2 < 0$
3. $H_0: \beta_3 \geq 0$
 $H_1: \beta_3 < 0$

Remember that a t -test typically is not run on the estimate of the constant term α .

2. *Choose a level of significance and therefore a critical t -value.* Assume that you have considered the various costs involved in making type I and type II errors and have chosen 5 percent as the level of significance with which you want to test. There are 10 observations in the data set that is going to be used to test these hypotheses, and so there are $10 - 3 - 1 = 6$ degrees of freedom. At a 5-percent level of significance, the critical t -value, t_{crit} , can be found in Statistical Table B-1 to be 1.943. Note that the level of significance does not have to be the same for all the coefficients in the same regression equation. It could well be that the costs involved in an incorrectly rejected null hypothesis for one coefficient are much higher than for another, so lower levels of significance would be used. In this equation, though, for all three variables:

$$t_{crit} = 1.943$$

3. *Run the regression and obtain an estimated t -value.* You now use the data (annual from 2000 to 2009) to run the regression on your OLS computer package, getting:

²¹ The null hypothesis can be stated either as $H_0: \beta \leq 0$ or $H_0: \beta = 0$ because the value used to test $H_0: \beta \leq 0$ is the value in the null hypothesis closest to the border between the acceptance and the rejection regions. When the amount of type I error is calculated, this border value of β is the one that is used, because over the whole range of $\beta \leq 0$, the value $\beta = 0$ gives the maximum amount of type I error. The classical approach limits this maximum amount to a preassigned level - the chosen level of significance.

$$CARS_t = 1.30 + 4.91YD_t + 0.00123PRICE_t - 7.14SUV_t \quad (6.5)$$

$\begin{matrix} (2.38) & (0.00044) & (71.38) \\ t = 2.1 & 2.8 & -0.1 \end{matrix}$

where: $CARS_t$ = new car sales (in hundreds of thousands of units) in year t

YD_t = real U.S. disposable income (in hundreds of billions of dollars)

$PRICE_t$ = the average real price of a new car in year t (in dollars)

SUV_t = the number of sports utility vehicles sold in year t (in millions)

Once again, we use our standard documentation notation, so the Figures in parentheses are the estimated standard errors of the $\hat{\beta}$ s. The t -values to be used in these hypothesis tests are printed out by standard OLS programs:

$$t_k = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \quad (k=1,2,\dots,K) \quad (6.6)$$

For example, the estimated coefficient of SUV divided by its estimated standard error is $-7.14/71.38 = -0.1$. Note that since standard errors are always positive, a negative estimated coefficient implies a negative t -value.

4. *Apply the decision rule by comparing the calculated t -value with the critical t -value in order to reject or not reject the null hypothesis.* As stated in Section 6.2, the decision rule for the t -test is to

Reject H_0 if $|t_k| > t_c$ and if t_k also has the sign implied by H_1 .

Do not reject H_0 otherwise.

What would these decision rules be for the three hypotheses, given the relevant critical t -value (1.943) and the calculated t -values?

For β_1 : Reject H_0 if $|2.1| > 1.943$ and if 2.1 is positive.

In the case of disposable income, you reject the null hypothesis that $\beta_1 \leq 0$ since 2.1 is indeed greater than 1.943. The result (that is,

H1: $\beta_1 > 0$) is as you expected on the basis of theory, since the more income in the country, the more new car sales you'd expect.

For β_2 : Reject H_0 if $|2.8| > 1.943$ and if 2.8 is negative.

For prices, the t -statistic is large in absolute value (being greater than 1.943) but has a sign that is contrary to our expectations, since the alternative hypothesis implies a negative sign. Since both conditions in the decision rule must be met before we can reject H_0 , you cannot reject the null hypothesis that prices have a zero or positive effect on new car sales!

Despite your surprise,²² you stick with your contention that prices belong in the equation and that their expected impact should be negative.

Notice that the coefficient of PRICE is quite small, 0.00123, but that this size has no effect on the t -calculation other than its relationship to the standard error of the estimated coefficient.

For β_3 : Reject H_0 if $|-0.1| > 1.943$ and if -0.1 is negative.

For sales of sports utility vehicles, the coefficient $\hat{\beta}_3$ is not statistically different from zero, since $|-0.1| < 1.943$, and you cannot reject the null hypothesis that $\beta \geq 0$ even though the estimated coefficient has the sign implied by the alternative hypothesis. After thinking this model over again, you come to the conclusion that you were hasty in adding the variable to the equation.

Figure 6.6 illustrates all three of these outcomes by plotting the critical t -value and the calculated t -values for all three null hypotheses on a t -distribution that is centered around zero (the value in the null hypothesis closest to the border between the acceptance and rejection regions). Students are urged to analyze the results of tests on the estimated coefficients of Equation 6.5 assuming different numbers of observations and different levels of significance.

²² Actually, it shouldn't be a surprise to occasionally get a positive estimated coefficient for price in a demand equation, particularly in such a small sample. Supply and demand are determined simultaneously, but we didn't specify a supply equation in our model. Thus our "demand" equation might be picking up the positive impact of price on quantity from the omitted supply equation. We'll deal with the simultaneity issue in Chapter 14.

The purpose of this example is to provide practice in testing hypotheses, and the results of such a poorly thought-out equation for such a small number of observations should not be taken too seriously. Given all that, however, it's still instructive to note that you did not react the same way to your inability to reject the null hypotheses for the price and sports utility vehicle variables. That is, the failure of the sports utility vehicle variable's coefficient to be significantly negative caused you to realize that perhaps the addition of this variable was ill-advised. The failure of the price variable's coefficient to be significantly negative did not cause you to consider the possibility that price has no effect on new car sales. Put differently, estimation results should never be allowed to cause you to want to adjust theoretically sound variables or hypotheses, but if they make you realize you have made a serious mistake, then it would be foolhardy to ignore that mistake.

What to do about the positive coefficient of price, on the other hand, is what the "art" of econometrics is all about. Surely a positive

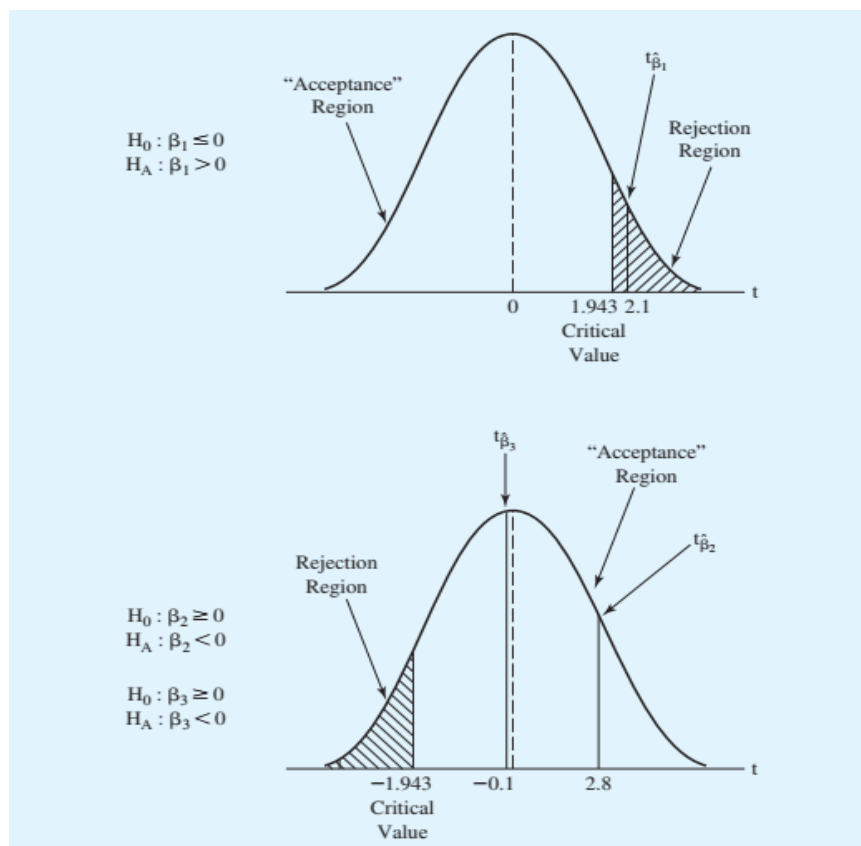


Figure 6.6 One-sided t -tests of the coefficients of the new car sales model

Given the estimates in Equation 5.6 and the critical t -value of 1.943 for a 5% level of significance, one sided, 6 degrees of freedom t -test, we can reject the null hypothesis for $\hat{\beta}_1$, but not for $\hat{\beta}_2$ or $\hat{\beta}_3$.

coefficient is unsatisfactory, but throwing the price variable out of the equation seems even more so. Possible answers to such issues are addressed more than once in the chapters that follow.

Examples of two-sided t -tests

Although most hypotheses in regression analysis should be tested with one-sided t -tests, two-sided t -tests are appropriate in particular situations. Researchers sometimes encounter hypotheses that should be rejected if estimated coefficients are significantly different from zero, or a specific nonzero value, in either direction. This situation requires a two-sided t -test. The kinds of circumstances that call for a two-sided test fall into two categories:

1. Two-sided tests of whether an estimated coefficient is significantly different from zero, and
2. Two-sided tests of whether an estimated coefficient is significantly different from a specific nonzero value.

Let's take a closer look at these categories:

1. *Testing whether a $\hat{\beta}$ is statistically different from zero.* The first case for a two-sided test of $\hat{\beta}$ arises when there are two or more conflicting hypotheses about the expected sign of a coefficient. For example, in the Rohat's restaurant equation of Section 4.2, the impact of the average income of an area on the expected number of Rohat's customers in that area is ambiguous. A high-income neighborhood might have more total customers going out to dinner, but those customers might decide to eat at a more formal restaurant than Rohat's. As a result, you might run a two-sided t -test around zero to determine whether the estimated coefficient of income is significantly different from zero in *either* direction. In other words, since there are reasonable cases to be made for either a positive or a negative coefficient, it is appropriate to test the $\hat{\beta}$ for income with a two-sided t -test:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

As Figure 6.7 illustrates, a two-sided test implies two different rejection regions (one positive and one negative) surrounding the acceptance region. A critical t -value, t_{crit} , must be increased in order to achieve the same level of significance with a two-sided test as can be achieved with a one-sided test.²³ As a result, there is an advantage to testing hypotheses with a one-sided test if the underlying theory allows because, for the same t -values, the possibility of type I error is half as much for a one-sided test as for a two-sided test. In cases where there are powerful theoretical arguments on both sides,

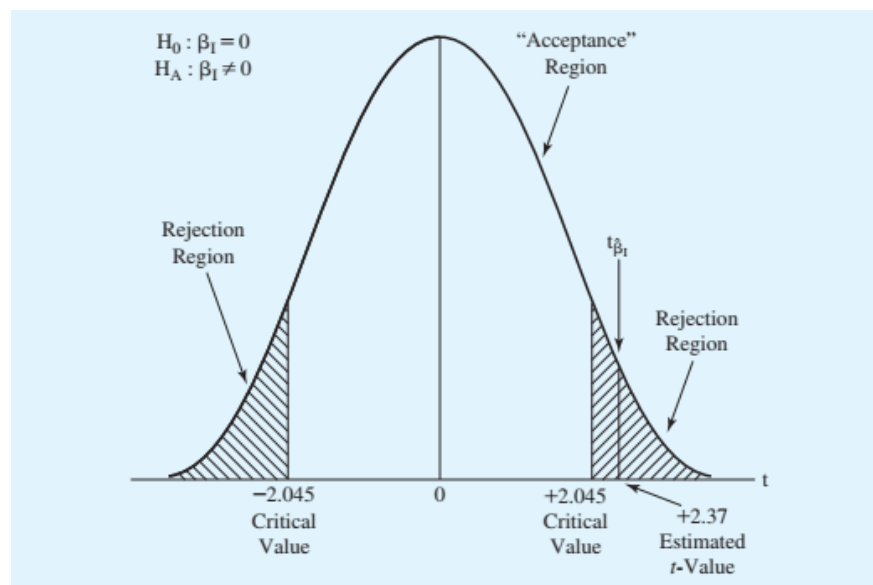


Figure 6.7 two-sided t -test of the coefficient of income in the Rohat's model

however, the researcher has no alternative to using a two-sided t -test around zero. To see how this works, let's follow through the Rohat's income variable example in more detail.

²³ See Figure 6.5. In that figure, the same critical t -value has double the level of significance for a two-sided test as for a one-sided test.

- a. *Set up the null and alternative hypotheses.*

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

- b. *Choose a level of significance and therefore a critical t -value.* You decide to keep the level of significance at 5%, but now this amount must be distributed between two rejection regions for 29 degrees of freedom. Hence, the correct critical t -value is 2.045 (found in Statistical Table B-1 for 29 degrees of freedom and a 5%, two sided test). Note that, technically, there now are two critical t -values, +2.045 and -2.045.
- c. *Run the regression and obtain an estimated t -value.* Since the value implied by the null hypothesis is still zero, the estimated t -value of +2.37 given in Equation 6.3 is applicable.
- d. *Apply the decision rule by comparing the calculated t -value with the critical t -value in order to reject or not reject the null hypothesis.* We once again use the decision rule stated in Section 6.2, but since the alternative hypothesis specifies either sign, the decision rule simplifies to:

$$\text{For } \beta_1: \text{Reject } H_0 \text{ if } |2.37| > 2.045$$

In this case, you reject the null hypothesis that β_1 equals zero because 2.37 is greater than 2.045 (see Figure 6.7). Note that the positive sign implies that, at least for Rohat's restaurants, income increases customer volume (holding constant population and competition). Given this result, we might well choose to run a one-sided t -test on the next year's Rohat's data set. For more practice with two-sided t -tests see «Econometric Analysis using Eviews: Practical tutorial».

2. *Two-sided t -tests of a specific nonzero coefficient value.* The second case for a two-sided t -test arises when there is reason to expect a specific nonzero value for an estimated

coefficient. For example, if a previous researcher has stated that the true value of some coefficient almost surely equals a particular number, β_{H0} , then that number would be the one to test by creating a two-sided t -test around the hypothesized value, β_{H0} .

In such a case, the null and alternative hypotheses become:

$$H0: \beta_k = \beta_{H0}$$

$$H1: \beta_k \neq \beta_{H0}$$

where β_{H0} is the specific nonzero value hypothesized. Since the hypothesized β value is no longer zero, the formula with which to calculate the estimated t -value is equation 6.2, repeated here:

$$\text{test statistics} = \frac{\hat{\beta} - \beta^*}{SE(\beta)} \quad (6.2)$$

This t -statistic is still distributed around zero if the null hypothesis is correct, because we have subtracted β_{H0} from the estimated regression coefficient whose expected value is supposed to be β_{H0} when $H0$ is true. Since the t -statistic is still centered around zero, the decision rule developed earlier still is applicable.

6.4 Confidence interval

Now that you've learned how to do hypothesis tests using the t -statistic and the p -value, you're probably thinking it would be fun to learn a third way. OK, maybe not! But there is indeed a third way. It's based on the concept of a confidence interval.

A **confidence interval** is a range of values that will contain the true value of β a certain percentage of the time, say 90 or 95 percent. The formula for a confidence interval is

$$\text{Confidence interval} = \hat{\beta} \pm t_{crit} \times SE(\hat{\beta}) \quad (6.7)$$

where t_{crit} is the two-sided critical value of the t -statistic for whatever significance level we choose. If you want a 90% confidence interval, you'd choose the critical value for the 10% significance level. For a 95% confidence interval, you'd use a 5% critical value.

To see how confidence intervals can be used for hypothesis tests, let's return to equation (4.4) of the Chapter 4 (Rohat's restaurant chain), and test the significance of the income coefficient:

$$\begin{aligned} \hat{Y}_i &= 102.192 - 9075N_i + 0.3547P_i + 1.288I_i \\ &\quad \quad \quad (2053) \quad \quad (0.0727) \quad \quad (0.543) \\ t &= -4.42 \quad 4.88 \quad 2.37 \\ N &= 33 \quad \quad \bar{R}^2 = 0.579 \end{aligned} \quad (6.8)$$

We'd typically expect sales at a restaurant to rise as income rises (a normal good), but Rohat's is a fairly low-priced restaurant chain, so there's a chance that sales will tail off if income gets too high (an inferior good). As a result, many econometricians would choose $\beta_I = 0$ as their null hypothesis and therefore run a two-sided test of β_I . In a situation where a two-sided test is appropriate, a confidence interval makes a lot of sense.

What would a 90% confidence interval for β_I look like? Well, $\hat{\beta}_I = 1.288$ and $SE(\hat{\beta}_I) = 0.543$, so all we need is a 10-percent two-sided critical t -value for 29 degrees of freedom. Using Statistical Table B-1, we see $t_{crit} = 1.699$. Substituting these values into equation 6.7, we get:

$$\begin{aligned} 90\% \text{ confidence interval around } \hat{\beta}_I &= 1.288 \pm 1.699 \times 0.543 \\ &= 1.288 \pm 0.923 \\ \text{and therefore } 0.365 &\leq \beta_I \leq 2.211 \end{aligned}$$

What exactly does this mean? If the Classical Assumptions hold true, the confidence interval formula produces ranges that contain the true value of β 90 percent of the time. In this case,

there's a 90 percent chance the true value of β_I is between 0.365 and 2.211. If it's not in that range, it's due to an unlucky sample.

How can we use a confidence interval for a two-tailed hypothesis test? If the null hypothesis is $\beta_I = 0$, we can reject it at the 10% level because 0 is not in the confidence interval. If the null hypothesis is that $\beta_I = 1.0$, we cannot reject it because 1.0 is in the interval. In general, if your null hypothesis border value is in the confidence interval, you cannot reject the null hypothesis.

Thus, confidence intervals can be used for two-sided tests, but they are more complicated. So why bother with them? It turns out that confidence intervals are very useful in telling us how precise a coefficient estimate is. And for many people using econometrics in the real world, this may be more important than hypothesis testing. An example will make it easier to understand.

Samira Niyozzi, a building contractor who specializes in starter homes for young families. It's a really competitive business, so to make a profit she needs to build appealing but inexpensive houses. As a result, Samira wants to know which features she can add to her houses that will increase the selling price more than they increase her costs. Since she took econometrics in college, she decides to estimate a model of starter home prices in her city using 13 independent variables (such as square footage, number of bathrooms, etc.). She hopes to use the results to decide which features might turn a profit for her on the houses that she's planning on building.

Let's focus on how much an additional bathroom might increase the sales price. Samira knows that her marginal cost of adding a bathroom is about \$8,000. She collects 100 observations on recently sold starter homes in her city and estimates her model.

The results appear in the first row of Table 6.1. The estimate of the coefficient of bathrooms is about \$21,770, well above the \$8,000 marginal cost. Sounds like a no-brainer to add a bathroom, right? Not so fast. Look at the 90-percent confidence interval in the first row of Table 6.1. It is huge, ranging from about \$187 to \$43,356. If the true value is in that interval, there's a pretty big chance it could be below \$8,000, meaning Samira would lose

money by adding a bathroom. Or, she could come out way ahead. What should she do?

Remember one of the lessons of Section 5.2 on sampling distributions: Bigger samples decrease the variance of $\hat{\beta}$. In plain English, as the sample size grows, $\hat{\beta}$ tends to get closer and closer to the true value of β . As a result, the confidence interval for β shrinks. Let's see what happens if Samira increases her sample to 1,000 observations.

The results for $\hat{\beta}_{bath}$ when $N=1,000$ appear in Table 6.1, right below those for $N = 100$. Notice that $\hat{\beta}_{bath}$ has fallen from almost \$22,000 to less than \$13,000. Does this mean Samira should not add a bathroom? Not at all! Look at the 90-percent confidence interval. The lower end has risen to \$8,346.29, a bit more than the \$8,000 marginal cost. While Samira could still lose money on the deal, it looks like a much safer bet than the small sample results suggest.

Why does a confidence interval become so much narrower with a bigger sample? Well, take a look at equation (6.7). As you can see, the width of a confidence interval depends entirely on the product of t_{crit} and $SE(\hat{\beta})$. What happens to t_{crit} and $SE(\hat{\beta})$ as the sample size rises? If you take a look at Table B-1, you'll see that as the sample size rises, t_c falls. Simultaneously, as the sample size rises, the variance of the sampling distribution falls, so the $SE(\hat{\beta})$ (the square root of the estimated variance) must fall as well. If both t_{crit} and $SE(\hat{\beta})$ fall, then their multiple must fall, and a bigger sample will indeed lead to a narrower confidence interval.

Table 6.1 Selected results from two regressions on selling prices of starter homes as a function of house characteristics

obs	variable	Coef.	Std. Err.	t	p-value	[90% Conf. Interval]
100	bath	21771.65	12981.1	1.68	0.097	187.1275 to 43356.18
1000	bath	12935.06	2787.154	4.64	0.000	8346.288 to 17523.83

These results were generated with data from Nashville, TN house sales in 2012. Because the model included 13 independent variables, the first regression has 86 degrees of freedom, and the second has 986 degrees of freedom.

This example illustrates how confidence intervals provide information on how precise an estimated coefficient is. In addition, confidence intervals also are extremely useful in forecasting, and we'll cover that topic in Chapter 15.

6.5 The F -Test

Although the t -test is invaluable for hypotheses about individual regression coefficients, it can't be used to test multiple hypotheses simultaneously. Such a limitation is unfortunate because many interesting ideas involve a number of hypotheses or involve one hypothesis about multiple coefficients. For example, suppose that you want to test the null hypothesis that there is no seasonal variation in a quarterly regression equation that has dummy variables for the seasons. To test such a hypothesis, most researchers would use the F -test.

What is the F -test?

The F -test is a formal hypothesis test that is designed to deal with a null hypothesis that contains multiple hypotheses or a single hypothesis about a group of coefficients.²⁴ Such "joint" or "compound" null hypotheses are appropriate whenever the underlying economic theory specifies values for multiple coefficients simultaneously.

The way in which the F -test works is fairly ingenious. The first step is to translate the particular null hypothesis in question into constraints that will be placed on the equation. The resulting constrained equation can be thought of as what the equation would look like if the null hypothesis were correct; you substitute the hypothesized values into the regression equation in order to see what would happen if the equation were constrained to agree with the null hypothesis. As a result, in the F -test the null hypothesis

²⁴ As you will see, the F -test works by placing constraints or restrictions on the equation to be tested. Because of this, it's equivalent to say that the F -test is for tests that involve multiple linear restrictions.

always leads to a constrained equation, even if this violates our standard practice that the alternative hypothesis contains what we expect is true.

The second step in an F -test is to estimate this constrained equation with OLS and compare the fit of this constrained equation with the fit of the unconstrained equation. If the fits of the constrained equation and the unconstrained equation are not substantially different, the null hypothesis should not be rejected. If the fit of the unconstrained equation is substantially better than that of the constrained equation, then we reject the null hypothesis. The fit of the constrained equation is never superior to the fit of the unconstrained equation, as we'll explain next. The fits of the equations are compared with the general F -statistic:

$$F = \frac{(RSS_M - RSS) / M}{RSS / (N - K - 1)} \quad (6.9)$$

where: RSS = residual sum of squares from the unconstrained equation

RSS_M = residual sum of squares from the constrained equation

M = number of constraints placed on the equation (usually equal to the number of β s eliminated from the unconstrained equation)

$(N-K-1)$ = degrees of freedom in the unconstrained equation

RSS_M is always greater than or equal to RSS . Imposing constraints on the coefficients instead of allowing OLS to select their values can never decrease the summed squared residuals. (Recall that OLS selects that combination of values of the coefficients that minimizes RSS .) At the extreme, if the unconstrained regression yields exactly the same estimated coefficients as does the constrained regression, then the RSS are equal and the F -statistic is zero. In this case, H_0 is not rejected because the data indicate that the constraints appear to be correct. As the difference between the constrained coefficients and the

unconstrained coefficients increases, the data indicate that the null hypothesis is less likely to be true. Thus, when F gets larger than the critical F -value, the hypothesized restrictions specified in the null hypothesis are rejected by the test.

The decision rule to use in the F -test is to reject the null hypothesis if the calculated F -value (F) from equation (6.9) is greater than the appropriate critical F -value (F_{crit}):

$$\begin{aligned} &\text{Reject } H_0 \text{ if } F > F_{crit} \\ &\text{Do not reject } H_0 \text{ if } F \leq F_{crit} \end{aligned}$$

The critical F -value, F_{crit} , is determined from Statistical Table B-2 or B-3, depending on the level of significance chosen by the researcher and on the degrees of freedom. The F -statistic has two types of degrees of freedom: the degrees of freedom for the numerator of equation (6.9) (M , the number of constraints implied by the null hypothesis) and the degrees of freedom for the denominator of equation (6.9) ($N-K-1$, the degrees of freedom in the unconstrained regression equation). The underlying principle here is that if the calculated F -value (or F -ratio) is greater than the critical value, then the estimated equation's fit is substantially better than the constrained equation's fit, and we can reject the null hypothesis of no effect.

EXAMPLE 6.3

Dropping the time subscripts for simplicity, suppose that the general regression is

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u \quad (6.10)$$

and that the restriction $\beta_3 + \beta_4 = 1$ is under test (there exists some hypothesis from theory which suggests that this would be an interesting hypothesis to study). The unrestricted regression is (6.10) above, but what is the restricted regression? It could be expressed as

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u \quad \text{s.t. (subject to) } \beta_3 + \beta_4 = 1 \quad (6.11)$$

The restriction ($\beta_3 + \beta_4 = 1$) is substituted into the regression so that it is automatically imposed on the data. The way that this would be achieved would be to make either β_3 or β_4 the subject of (6.11), e.g.

$$\beta_3 + \beta_4 = 1 \Rightarrow \beta_4 = 1 - \beta_3 \quad (6.12)$$

and then substitute into (6.11) for β_4

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + (1 - \beta_3) x_4 + u \quad (6.13)$$

Equation (6.13) is already a restricted form of the regression, but it is not yet in the form that is required to estimate it using a computer package. In order to be able to estimate a model using OLS, software packages usually require each RHS variable to be multiplied by one coefficient only. Therefore, a little more algebraic manipulation is required. First, expanding the brackets around $(1 - \beta_3)$

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + x_4 - \beta_3 x_4 + u \quad (6.14)$$

Then, gathering all of the terms in each β_i together and rearranging

$$(y - x_4) = \beta_1 + \beta_2 x_2 + \beta_3 (x_3 - x_4) + u \quad (6.15)$$

Note that any variables without coefficients attached (e.g. x_4 in (6.14)) are taken over to the LHS and are then combined with y . Equation (6.15) is the restricted regression. It is actually estimated by creating two new variables - call them, say, P and Q , where $P = y - x_4$ and $Q = x_3 - x_4$ - so the regression that is actually estimated is

$$P = \beta_1 + \beta_2 x_2 + \beta_3 Q + u \quad (6.16)$$

What would have happened if instead β_3 had been made the subject of (6.12) and β_3 had therefore been removed from the equation? Although the equation that would have been estimated would have been different from (6.16), the value of the residual sum of squares for these two models (both of which have imposed upon them the same restriction) would be the same.

The F-test of overall significance

Although R^2 and $\bar{R}^2$ measure the overall degree of fit of an equation, they don't provide a formal hypothesis test of that overall fit. Such a test is provided by the F -test. The null hypothesis in an F -test of overall significance is that all the slope coefficients in the equation equal zero simultaneously. For an equation with K independent variables, this means that the null and alternative hypotheses would be²⁵:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_K = 0$$

$$H_1: H_0 \text{ is not true}$$

To show that the overall fit of the estimated equation is statistically significant, we must be able to reject this null hypothesis using the F -test.

For the F -test of overall significance, equation (6.9) simplifies to:

$$F = \frac{ESS / K}{RSS / (N - K - 1)} = \frac{\sum (\hat{y}_i - \bar{y})^2 / K}{\sum e_i^2 / (N - K - 1)} \quad (6.17)$$

This is the ratio of the explained sum of squares (ESS) to the residual sum of squares (RSS), adjusted for the number of independent variables (K) and the number of observations in the sample (N). In this case, the “constrained equation” to which we’re comparing the overall fit is:

$$y_i = a + u_i \quad (6.18)$$

which is nothing more than saying $\hat{y}_i = \bar{y}$. Thus the F -test of overall significance is really testing the null hypothesis that the fit of the equation is no better than that provided by using the mean alone.

²⁵ Note that we don't hypothesize that $a = 0$. This would imply that $E(\bar{y}) = 0$. Note also that for the test of overall significance, $M = K$.

To see how this works, let's test the overall significance of the Rohat's restaurant model. Since there are three independent variables, the null and alternative hypotheses are:

$$H_0: \beta_N = \beta_P = \dots = \beta_I = 0$$

H1: H0 is not true

To decide whether to reject or not reject this null hypothesis, we need to calculate Equation 5.11 for the Woody's example. There are three constraints in the null hypothesis, so $K = 3$. If we check the Eviews output for the Rohat's equation presented in previous Chapter, we can see that $N = 33$, $RSS = 6,133,300,000$, and $ESS = 9,928,900,000$. Thus the appropriate F -ratio is:

$$F = \frac{ESS / K}{RSS / (N - K - 1)} = \frac{9,928,900,000 / 3}{6,133,300,000 / 29} = 15.65 \quad (6.19)$$

In practice, this calculation is never necessary, since virtually every computer regression package routinely provides the computed F -ratio for a test of overall significance as a matter of course. On the Woody's computer output, the value of the F -statistic can be found near the top of the right-hand column. Our decision rule tells us to reject the null hypothesis if the calculated F -value is greater than the critical F -value. To determine that critical F -value, we need to know the level of significance and the degrees of freedom. If we assume a 5% level of significance, the appropriate Table to use is Statistical Table B-2. The numerator degrees of freedom equal 3 (K), and the denominator degrees of freedom equal 29 ($N - K - 1$), so we need to look in Statistical Table B-2 for the critical F -value for 3 and 29 degrees of freedom.

As the reader can verify,²⁶ $F_{crit} = 2.93$ is well below the calculated F -value of 15.65, so we can reject the null hypothesis and

²⁶ Note that this critical F -value must be interpolated. The critical value for 30 denominator degrees of freedom is 2.92, and the critical value for 25 denominator degrees of freedom is 2.99. Since both numbers are well below the calculated F -value of 15.65, however, the interpolation isn't necessary to reject the null hypothesis. As a result, many

conclude that the Woody's equation does indeed have a significant overall fit.

Just as p -values provide an alternative approach to the t -test, so too can p -values provide an alternative approach to the F -test of overall significance. Most standard regression estimation programs report not only the F -value for the test of overall significance but also the p -value associated with that test. If the p -value is less than your chosen level of significance, you can reject the null hypothesis.

The relationship between the t - and the F -distributions.

Any hypothesis that could be tested with a t -test could also have been tested using an F -test, but not the other way around. So, single hypotheses involving one coefficient can be tested using a t - or an F test, but multiple hypotheses can be tested only using an F -test. For example, consider the hypothesis

$$H_0 : \beta_2 = 0.5$$

$$H_1 : \beta_2 \neq 0.5$$

This hypothesis could have been tested using the usual t -test

$$\text{test stat} = \frac{\hat{\beta}_2 - 0.5}{SE(\hat{\beta}_2)} \quad (6.20)$$

or it could be tested in the framework above for the F -test. Note that the two tests always give the same conclusion since the t -distribution is just a special case of the F -distribution. For example, consider any random variable Z that follows a t -distribution with $T - k$ degrees of freedom, and square it. The square of the t is equivalent to a particular form of the F -distribution

$$Z^2 \sim t^2(T - k) \text{ then also } Z^2 \sim F(1, T - k)$$

Thus the square of a t -distributed random variable with $T - k$ degrees of freedom also follows an F -distribution with 1 and $T - k$

researchers don't bother with such interpolations unless the calculated F -value is inside the range of the interpolation.

degrees of freedom. This relationship between the t and the F -distributions will always hold - take some examples from the statistical Tables and try it!

The F -distribution has only positive values and is not symmetrical. Therefore, the null is rejected only if the test statistic exceeds the critical F -value, although the test is a two-sided one in the sense that rejection will occur if $\hat{\beta}_2$ is significantly bigger or significantly smaller than 0.5.

Other cases of the F -test

There are many other uses of the F -test besides the test of overall significance. For example, let's take a look at the problem of testing the significance of seasonal dummies. **Seasonal dummies** are dummy variables that are used to account for seasonal variation in the data in time-series models. In a quarterly model, if:

$$\begin{aligned} x_{1t} &= \begin{cases} 1 & \text{in quarter 1} \\ 0 & \text{otherwise} \end{cases} \\ x_{2t} &= \begin{cases} 1 & \text{in quarter 2} \\ 0 & \text{otherwise} \end{cases} \\ x_{3t} &= \begin{cases} 1 & \text{in quarter 3} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

then:

$$Y_t = a + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + u_t \quad (6.21)$$

where x_4 is a nondummy independent variable and t is quarter. Notice that only three dummy variables are required to represent four seasons. In this formulation β_1 shows the extent to which the expected value of y in the first quarter differs from its expected value in the fourth quarter, the omitted condition. β_2 and β_3 can be interpreted similarly.

Inclusion of a set of seasonal dummies “deseasonalizes” y . This procedure may be used as long as Y and X_4 are not “seasonally adjusted” prior to estimation. Many researchers avoid the type of

seasonal adjustment done prior to estimation because they think it distorts the data in unknown and arbitrary ways, but seasonal dummies have their own limitations such as remaining constant for the entire time period. As a result, there is no unambiguously best approach to deseasonalizing data.

To test the hypothesis of significant seasonality in the data, one must test the hypothesis that all the dummies equal zero simultaneously rather than test the dummies one at a time. In other words, the appropriate test of seasonality in a regression model using seasonal dummies involves the use of the F -test instead of the t -test.

In this case, the null hypothesis is that there is *no* seasonality:

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

H_1 : H_0 is not true

The constrained equation would then be $y = a + \beta_4 x_4 + u$. To determine whether the whole set of seasonal dummies should be included, the fit of the estimated constrained equation would be compared to the fit of the estimated unconstrained equation by using the F -test in equation (6.9). Note that this example uses the F -test to test null hypotheses that include only a subset of the slope coefficients. Also note that in this case $M = 3$, because three coefficients (β_1 , β_2 , and β_3) have been eliminated from the equation. The exclusion of some seasonal dummies because their estimated coefficients have low t -scores is not recommended. Seasonal dummy coefficients should be tested with the F -test instead of with the t -test because seasonality is usually a single compound hypothesis rather than 3 individual hypotheses (or 11 with monthly data) having to do with each quarter (or month). To the extent that a hypothesis is a joint one, it should be tested with the F -test. If the hypothesis of seasonal variation can be summarized into a single dummy variable, then the use of the t -test will cause no exercises. Often, where seasonal dummies are unambiguously called for, no hypothesis testing at all is undertaken.

6.5 Sample Eviews output for multiple hypothesis tests

Reload the ‘capm.wk1’ workfile constructed in the previous chapter. As a reminder the results are included again below.

Dependent Variable: ERFORD				
Method: Least Squares				
Date: 08/21/07 Time: 15:02				
Sample (adjusted): 2002M02 2007M04				
Included observations: 63 after adjustments				
	Coefficient	Std. Error	t-Statistic	Prob.
C	2.020219	2.801382	0.721151	0.4736
ERSANDP	0.359726	0.794443	0.452803	0.6523
R-squared	0.003350	Mean dependent var		2.097445
Adjusted R-squared	−0.012989	S.D. dependent var		22.05129
S.E. of regression	22.19404	Akaike info criterion		9.068756
Sum squared resid	30047.09	Schwarz criterion		9.136792
Log likelihood	−283.6658	Hannan-Quinn criter.		9.095514
F-statistic	0.205031	Durbin-Watson stat		1.785699
Prob(F-statistic)	0.652297			

Screenshot 6.1 Results of CAPM model estimation.

If we examine the regression F -test, this also shows that the regression slope coefficient is not significantly different from zero, which in this case is exactly the same result as the t -test for the beta coefficient (since there is only one slope coefficient). Thus, in this instance, the F -test statistic is equal to the square of the slope t -ratio.

Now suppose that we wish to conduct a joint test that both the intercept and slope parameters are 1. We would perform this test exactly as for a test involving only one coefficient. Select **View/Coefficient Tests/Wald - Coefficient Restrictions**. . . and then in the box that appears, type **C(1)=1,C(2)=1**. There are two versions of the test given: an F -version and a χ^2 - version. The F -version is adjusted for small sample bias and should be used when the regression is estimated using a small sample (see chapter 4). Both statistics asymptotically yield the same result, and in this case the p -values are very similar. The conclusion is that the joint null hypothesis, $H_0 : \beta_1 = 1$ and $\beta_2 = 1$, is not rejected.

Multiple regression in Eviews using an APT-style model

In the spirit of arbitrage pricing theory (APT), the following example will examine regressions that seek to determine whether the monthly returns on Microsoft stock can be explained by reference to unexpected changes in a set of macroeconomic and financial variables. **Open a new Eviews workfile** to store the data. There are 254 monthly observations in the file ‘macro.xls’, starting in March 1996 and ending in April 2017. There are 13 series plus a column of dates. The series in the Excel file are the Microsoft stock price, the S&P500 index value, the consumer price index, an industrial production index, Treasury bill yields for the following maturities: three months, six months, one year, three years, five years and ten years, a measure of ‘narrow’ money supply, a consumer credit series, and a ‘credit spread’ series. The latter is defined as the difference in annualised average yields between a portfolio of bonds rated AAA and a portfolio of bonds rated BAA.

Import the data from the Excel file and save the resulting workfile as ‘macro.wf1’.

The first stage is to generate a set of changes or *differences* for each of the variables, since the APT posits that the stock returns can be explained by reference to the *unexpected changes* in the macroeconomic variables rather than their levels. The unexpected value of a variable can be defined as the difference between the actual (realised) value of the variable and its expected value. The question then arises about how we believe that investors might have formed their expectations, and while there are many ways to construct measures of expectations, the easiest is to assume that investors have naïve expectations that the next period value of the variable is equal to the current value. This being the case, the entire change in the variable from one period to the next is the unexpected change (because investors are assumed to expect no change).²⁷

²⁷ It is an interesting question as to whether the differences should be taken on the levels of the variables or their logarithms. If the former, we have absolute changes in the variables, whereas the latter would lead to proportionate changes. The choice between the two is essentially an empirical one, and this example assumes that the former is chosen, apart from for the stock price series themselves and the consumer price series.

Transforming the variables can be done as described above. Press **Genr** and then enter the following in the 'Enter equation' box:

dspread = baa aaa spread - baa aaa spread(-1)

Repeat these steps to conduct all of the following transformations:

dcredit = consumer credit - consumer credit(-1)

dprod = industrial production - industrial production(-1)

rmsoft = 100*dlog(microsoft)

rsandp = 100*dlog(sandp)

dmoney = m1money supply - m1money supply(-1)

inflation = 100*dlog(cpi)

term = ustb10y - ustb3m

and then click **OK**. Next, we need to apply further transformations to some of the transformed series, **so repeat the above steps** to generate

dinflation = inflation - inflation(-1)

mustb3m = ustb3m/12

rterm = term - term(-1)

ermsoft = rmsoft - mustb3m

ersandp = rsandp - mustb3m

The final two of these calculate excess returns for the stock and for the index.

We can now run the regression. So click **Object/New Object/Equation** and name the object '**msoftreg**'. Type the following variables in the Equation specification window

**ERMSOFT C ERSANDP DPROD DCREDIT DINFLATION
DMONEY DSPREAD RTERM**

and use **Least Squares** over the whole sample period. The Table of results will appear as follows.

Dependent Variable: ERMSOFT
 Method: Least Squares
 Date: 08/21/07 Time: 21:45
 Sample (adjusted): 1986M05 2007M04
 Included observations: 252 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.587603	1.457898	-0.403048	0.6873
ERSANDP	1.489434	0.203276	7.327137	0.0000
DPROD	0.289322	0.500919	0.577583	0.5641
DCREDIT	-5.58E-05	0.000160	-0.347925	0.7282
DINFLATION	4.247809	2.977342	1.426712	0.1549
DMONEY	-1.161526	0.713974	-1.626847	0.1051
DSPREAD	12.15775	13.55097	0.897187	0.3705
RTERM	6.067609	3.321363	1.826843	0.0689
R-squared	0.203545	Mean dependent var	-0.420803	
Adjusted R-squared	0.180696	S.D. dependent var	15.41135	
S.E. of regression	13.94965	Akaike info criterion	8.140017	
Sum squared resid	47480.62	Schwarz criterion	8.252062	
Log likelihood	-1017.642	Hannan-Quinn criter.	8.185102	
F-statistic	8.908218	Durbin-Watson stat	2.156221	
Prob(F-statistic)	0.000000			

Screenshot 6.2 Results of APT model estimation.

Take a few minutes to examine the main regression results. Which of the variables has a statistically significant impact on the Microsoft excess returns? Using your knowledge of the effects of the financial and macroeconomic environment on stock returns, examine whether the coefficients have their expected signs and whether the sizes of the parameters are plausible.

The regression F-statistic takes a value 8.908. Remember that this tests the null hypothesis that all of the slope parameters are jointly zero. The p-value of zero attached to the test statistic shows that this null hypothesis should be rejected.

However, there are a number of parameter estimates that are not significantly different from zero – specifically those on the DPROD, DCREDIT and DSPREAD variables. Let us test the null hypothesis that the parameters on these three variables are jointly zero using an F -test. To test this, Click on **View/Coefficient Tests/Wald – Coefficient Restrictions**. . . and in the box that appears type **C(3)=0, C(4)=0, C(7)=0** and click **OK**. The resulting F -test statistic follows an $F(3, 244)$ distribution as there are three restrictions, 252 usable observations and eight parameters to

estimate in the unrestricted regression. The F -statistic value is 0.402 with p -value 0.752, suggesting that the null hypothesis cannot be rejected. The parameters on DINLATION and DMONEY are almost significant at the 10% level and so the associated parameters are not included in this F -test and the variables are retained.

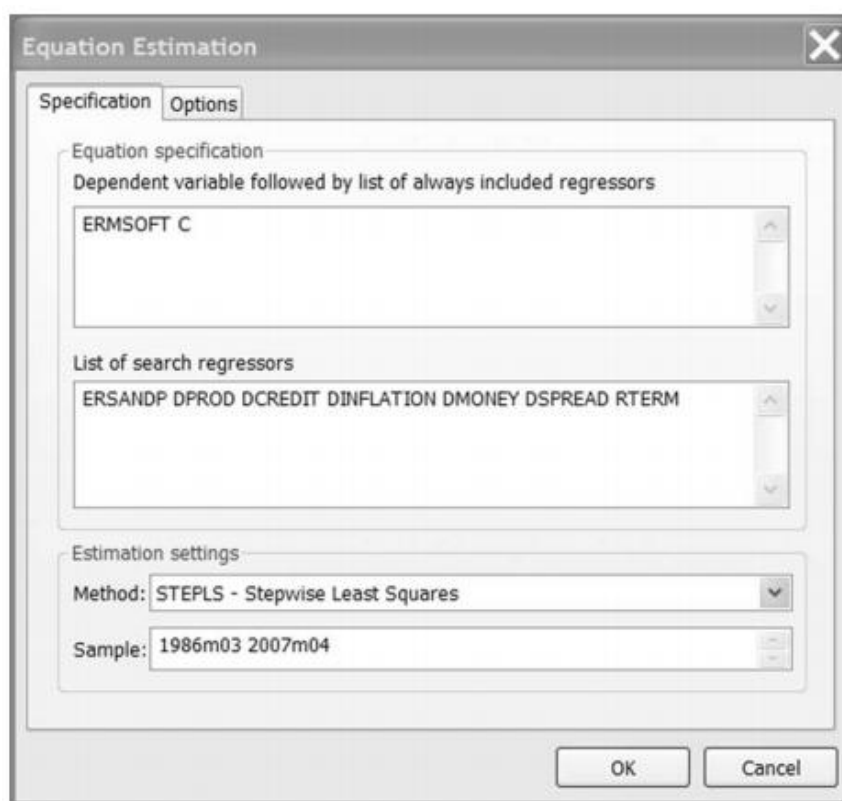
There is a procedure known as a *stepwise regression* that is now available in Eviews. Stepwise regression is an automatic variable selection procedure which chooses the jointly most ‘important’ (variously defined) explanatory variables from a set of candidate variables. There are a number of different stepwise regression procedures, but the simplest is the uni-directional forwards method. This starts with no variables in the regression (or only those variables that are always required by the researcher to be in the regression) and then it selects first the variable with the lowest p -value (largest t -ratio) if it were included, then the variable with the second lowest p -value conditional upon the first variable already being included, and so on. The procedure continues until the next lowest p -value relative to those already included variables is larger than some specified threshold value, then the selection stops, with no more variables being incorporated into the model.

To conduct a stepwise regression which will automatically select from among these variables the most important ones for explaining the variations in Microsoft stock returns, click **Proc** and then **Equation**. Name the equation **Msoftstepwise** and then in the ‘Estimation settings/Method’ box, change *LS - Least Squares (NLS and ARMA)* to **STEPLS – Stepwise Least Squares** and then in the top box that appears, ‘Dependent variable followed by list of always included regressors’, enter

ERMSOFT C

This shows that the dependent variable will be the excess returns on Microsoft stock and that an intercept will always be included in the regression. If the researcher had a strong prior view that a particular explanatory variable must always be included in the regression, it should be listed in this first box. In the second box, ‘List of search regressors’, type the list of all of the explanatory

variables used above: **ERSANDP DPROD DCREDIT DINFLATION DMONEY DSPREAD RTERM**. The window will appear as in screenshot 6.3.



Screenshot 6.3 Stepwise procedure equation estimation window.

Clicking on the ‘Options’ tab gives a number of ways to conduct the regression. For example, ‘Forwards’ will start with the list of required regressors (the intercept only in this case) and will sequentially add to them, while ‘Backwards’ will start by including all of the variables and will sequentially delete variables from the regression. The default criterion is to include variables if the p -value is less than 0.5, but this seems high and could potentially result in the inclusion of some very insignificant variables, so **modify this to 0.2** and then click **OK** to see the results. As can be seen, the excess market return, the term structure, money supply and unexpected inflation variables have all been included, while the default spread and credit variables have been omitted.

Dependent Variable: ERMSOFT
Method: Stepwise Regression
Date: 08/27/07 Time: 10:21
Sample (adjusted): 1986M05 2007M04
Included observations: 252 after adjustments
Number of always included regressors: 1
Number of search regressors: 7
Selection method: Stepwise forwards
Stopping criterion: p-value forwards/backwards = 0.2/0.2

	Coefficient	Std. Error	t-Statistic	Prob.*
C	−0.947198	0.8787	−1.077954	0.2821
ERSANDP	1.471400	0.201459	7.303725	0.0000
RTERM	6.121657	3.292863	1.859068	0.0642
DMONEY	−1.171273	0.702523	−1.667238	0.0967
DINFLATION	4.013512	2.876986	1.395040	0.1643
R-squared	0.199612	Mean dependent var		−0.420803
Adjusted R-squared	0.186650	S.D. dependent var		15.41135
S.E. of regression	13.89887	Akaike info criterion		8.121133
Sum squared resid	47715.09	Schwarz criterion		8.191162
Log likelihood	−1018.263	Hannan-Quinn criter.		8.149311
F-statistic	15.40008	Durbin-Watson stat		2.150604
Prob(F-statistic)	0.000000			
Selection Summary				
Added ERSANDP				
Added RTERM				
Added DMONEY				
Added DINFLATION				

*Note: p-values and subsequent tests do not account for stepwise selection.

Screenshot 6.4 Stepwise regression results.

Stepwise procedures have been strongly criticised by statistical purists. At the most basic level, they are sometimes argued to be no better than automated procedures for data mining, in particular if the list of potential candidate variables is long and results from a ‘fishing trip’ rather than a strong prior financial theory. More subtly, the iterative nature of the variable selection process implies that the size of the tests on parameters attached to variables in the final model will not be the nominal values (e.g. 5%) that would have applied had this model been the only one estimated. Thus the p -values for tests involving parameters in the final regression should really be modified to take into account that the model results from a sequential procedure, although they are usually not in statistical packages such as Eviews.

SUMMARY

1. Hypothesis testing makes inferences about the validity of specific economic (or other) theories from a sample of the population for which the theories are supposed to be true. The four basic steps of hypothesis testing (using a t -test as an example) are:
 - a. Set up the null and alternative hypotheses.
 - b. Choose a level of significance and, therefore, a critical t -value.
 - c. Run the regression and obtain an estimated t -value.
 - d. Apply the decision rule by comparing the calculated t -value with the critical t -value in order to reject or not reject the null hypothesis.
2. The null hypothesis states the range of values that the regression coefficient is expected to take on if the researcher's theory is not correct. The alternative hypothesis is a statement of the range of values that the regression coefficient is expected to take if the researcher's theory is correct.
3. The two kinds of errors we can make in such hypothesis testing are:

Type I: We reject a null hypothesis that is true.

Type II: We do not reject a null hypothesis that is false.

4. The t -test tests hypotheses about individual coefficients from regression equations. The form for the t -statistic is

$$t_k = \frac{(\hat{\beta}_k - \beta_{H0})}{SE(\hat{\beta}_k)} \quad k = 1, 2, \dots, K$$

In many regression applications, β_{H0} is zero. Once you have calculated a t -value and chosen a critical t -value, you reject the null hypothesis if the t -value is greater in absolute value than the critical t -value and if the t -value has the sign implied by the alternative hypothesis.

5. The t -test is easy to use for a number of reasons, but care should be taken when using the t -test to avoid confusing statistical significance with theoretical validity or empirical evidence.
6. The F -test is a formal hypothesis test designed to deal with a null hypothesis that contains multiple hypotheses or a single hypothesis about a group of coefficients. The most common use of the F -test is to test the overall significance of an estimated equation.

CHAPTER 7. SPECIFICATION: CHOOSING THE INDEPENDENT VARIABLES

Before any equation can be estimated, it must be specified. *Specifying* an econometric equation consists of three parts: choosing the correct independent variables, the correct functional form, and the correct form of the stochastic error term.

A **specification error** results when any one of these choices is made incorrectly. This chapter is concerned with only the first of these, choosing the variables; the second and third choices will be taken up in later chapters. The fact that researchers can decide which independent variables to include in regression equations is a source of both strength and weakness in econometrics. The strength is that the equations can be formulated to fit individual needs, but the weakness is that researchers can estimate many different specifications until they find the one that “proves” their point, even if many other results disprove it. A major goal of this chapter is to help you understand how to choose variables for your regressions without falling prey to the various errors that result from misusing the ability to choose.

The primary consideration in deciding whether an independent variable belongs in an equation is whether the variable is essential to the regression on the basis of theory. If the answer is an unambiguous yes, then the variable definitely should be included in the equation, even if it seems to be lacking in statistical significance. If theory is ambivalent or less emphatic, a dilemma arises. Leaving a relevant variable out of an equation is likely to bias the remaining estimates, but including an irrelevant variable leads to higher variances of the estimated coefficients. Although we’ll develop statistical tools to help us deal with this decision, it’s difficult in practice to be sure that a variable is relevant, and so the problem often remains unresolved.

We devote the fourth section of the chapter to specification searches and the pros and cons of various approaches to such

searches. For example, poorly done specification searches often cause bias or make the usual tests of significance inapplicable. Instead, we suggest trying to minimize the number of regressions estimated and relying as much as possible on theory rather than statistical fit when choosing variables. There are no pat answers, however, and so the final decisions must be left to each individual researcher.

7.1 Omitted variables

Suppose that you forget to include one of the relevant independent variables when you first specify an equation (after all, no one's perfect!). Or suppose that you can't get data for one of the variables that you *do* think of. The result in both these situations is an **omitted variable**, defined as an important explanatory variable that has been left out of a regression equation.

Whenever you have an omitted (or *left-out*) variable, the interpretation and use of your estimated equation become suspect. Leaving out a relevant variable, like price from a demand equation, not only prevents you from getting an estimate of the coefficient of price but also usually causes bias in the estimated coefficients of the variables that are in the equation.

The bias caused by leaving a variable out of an equation is called **omitted variable bias**. In an equation with more than one independent variable, the coefficient β_k represents the change in the dependent variable y caused by a one-unit increase in the independent variable x_k , holding constant the other independent variables in the equation. If a variable is omitted, then it is not included as an independent variable, and it is not held constant for the calculation and interpretation of $\hat{\beta}_k$. This omission can cause bias: It can force the expected value of the estimated coefficient away from the true value of the population coefficient.

Thus, omitting a relevant variable is usually evidence that the entire estimated equation is suspect, because of the likely bias in the coefficients of the variables that remain in the equation. Let's look at this issue in more detail.

The consequences of an omitted variable

What happens if you omit an important variable from your equation (perhaps because you can't get the data for the variable or didn't even think of the variable in the first place)? The major consequence of omitting a relevant independent variable from an equation is to cause bias in the regression coefficients that remain in the equation. Suppose that the true regression model is:

$$y_i = a + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \quad (7.1)$$

where u_i is a classical error term. If you omit x_2 from the equation, then the equation becomes:

$$y_i = a^* + \beta_1^* x_{1i} + u_i^* \quad (7.2)$$

where u_i^* equals:

$$u_i^* = u_i + \beta_2 x_{2i} \quad (7.3)$$

because the stochastic error term includes the effects of any omitted variables, as mentioned in Section 1.2. Why does equation (7.2) include a^* and β_1^* instead of a and β_1 ? The answer lies in the meaning of a regression coefficient. β_1 is the impact of a one-unit increase in x_1 on y , *holding x_2 constant*, but x_2 isn't in equation (7.2), so OLS can't hold it constant. As a result, β_1^* is the impact of a one-unit increase in x_1 on y , *not holding x_2 constant*.

From equations (7.2) and (7.3), it might seem as though we could get unbiased estimates even if we left x_2 out of the equation. Unfortunately, this is not the case,²⁸ because the included coefficients almost surely pick up some of the effect of the omitted variable and therefore will change, causing bias. To see why, take another look at equations (7.2) and (7.3). Most pairs of variables are correlated to some degree, so x_1 and x_2 almost surely are correlated.

²⁸ To avoid bias, x_1 and x_2 must be perfectly uncorrelated in the sample - an extremely unlikely result.

When x_2 is omitted from the equation, the impact of x_2 goes into u^* , so u^* and x_2 are correlated. Thus if x_2 is omitted from the equation and x_1 and x_2 are correlated, both x_1 and u^* will change when x_2 changes, and the error term will no longer be independent of the explanatory variable. That violates Classical assumption 3!

In other words, if we leave an important variable out of an equation, we violate Classical assumption 3 (that the explanatory variables are independent of the error term), unless the omitted variable is uncorrelated with all the included independent variables (which is extremely unlikely). In general, when there is a violation of one of the classical assumptions, the Gauss - Markov theorem does not hold, and the OLS estimates are not BLUE. Given linear estimators, this means that the estimated coefficients are no longer unbiased or are no longer minimum variance (for all linear unbiased estimators), or both. In such a circumstance, econometricians first determine the exact property (unbiasedness or minimum variance) that no longer holds and then suggest an alternative estimation technique that might be better than OLS.

An omitted variable causes Classical assumption 3 to be violated in a way that causes bias. Estimating equation (7.2) when equation (7.1) is the truth will cause bias. This means that:

$$E(\hat{\beta}_1^*) \neq \beta_1 \quad (7.4)$$

Instead of having an expected value equal to the true β_1 , the estimate will compensate for the fact that x_2 is missing from the equation. If x_1 and x_2 are correlated and x_2 is omitted from the equation, then the OLS estimation procedure will attribute to x_1 variations in y that are actually caused by x_2 , and a biased estimate of β_1 will result. To see how an omitted variable can cause bias, let's look at an extremely early application of regression analysis. During World War II, the Allies were interested in improving the accuracy of their bombers, so they estimated an equation where the dependent variable was bomber accuracy and the independent variables included such things as the speed and altitude of the bombing group and the amount of enemy fighter opposition. As

expected, the estimated coefficients supported the hypotheses that higher speeds and higher altitudes led to larger aiming errors, but the researchers were shocked to discover that more enemy fighter opposition appeared to improve the accuracy of the pilot and bombardier! What was going on? The answer is omitted variable bias. It turns out that the equation didn't include a variable for cloud cover over the target, and cloud cover typically prevented enemy fighters from flying. When it was cloudy, the bombers couldn't see the ground and made large errors, but OLS attributed these errors to the lack of enemy fighter opposition because there was no variable for cloud cover in the equation and because few fighters could fly when it was cloudy. Put differently, the coefficient of enemy fighters picked up the impact of the omitted variable of cloud cover because the two variables were highly correlated. This is omitted variable bias!

To generalize for a model with two independent variables, the expected value of the coefficient of an included variable (x_1) when a relevant variable (x_2) is omitted from the equation equals:

$$E(\hat{\beta}_1^*) = \beta_1 + \beta_2 \times \hat{a}_1 \quad (7.5)$$

where $\hat{a}_1$ is an estimate of the slope coefficient of the secondary regression that relates x_2 to x_1 :

$$\hat{x}_{2i} = \hat{a}_0 + \hat{a}_1 x_{1i} \quad (7.6)$$

If x_1 and x_2 are positively correlated, $\hat{a}_1$ will be positive. If x_1 and x_2 are negatively correlated, $\hat{a}_1$ will be negative. If x_1 and x_2 are uncorrelated, $\hat{a}_1$ will be zero.

Let's take a look at equation (7.5). It states that the expected value of the included variable's coefficient is equal to its true value plus the omitted variable's true coefficient times a function of the correlation between the included and omitted variables.²⁹ Since the

²⁹ Equations (7.5) and (7.7) hold when there are exactly two independent variables, but the more general equations are quite similar.

expected value of an unbiased estimate equals the true value, the right-hand term in equation (7.5) measures the omitted variable bias in the equation:

$$Bias = \beta_2 \hat{a}_1 \quad (7.7)$$

In general terms, the bias thus equals the coefficient of the omitted variable times a function of the correlation between the included and omitted variables. This bias exists unless:

1. the true coefficient equals zero, or
2. the included and omitted variables are uncorrelated in the sample.

The term $\beta_2 \hat{a}_1$ is the amount of **expected bias** introduced into the estimate of the coefficient of the included variable by leaving out the omitted variable. Although it's true that there is no bias if the included and excluded variables are uncorrelated, there almost always is some correlation between any two variables in the real world, and so bias is almost always caused by the omission of a relevant variable.

EXAMPLE 7.1

For an example of omitted variable bias, let's go back to equation 6.8, the Rohat's restaurants model that we first studied in Section 4.2:

$$\begin{array}{rcccl} \hat{Y}_i = 102.192 - 9075N_i + 0.3547P_i + 1.288I_i & & & & \\ \quad \quad \quad (2053) \quad \quad (0.0727) \quad \quad (0.543) & & & & \\ t = -4.42 \quad 4.88 \quad 2.37 & & & & \\ N = 33 \quad \quad \bar{R}^2 = 0.579 & & & & (6.8) \end{array}$$

where Y = customers (check volume), N = the number of competitive restaurants nearby, P = the population nearby, and I = the average household income nearby.

Let's take a look at what happens if we drop population (P) from the equation:

$$\begin{aligned}\hat{Y}_i &= 84.439 - 1487N_i + 2.322I_i \\ &\quad (1778) \quad (0.664) \\ t &= -0.84 \quad +3.50 \\ N &= 33 \quad \bar{R}^2 = 0.258\end{aligned}\tag{7.8}$$

Stop for a minute and compare Equations 5.4 and 6.8. The most noticeable difference is that $\bar{R}^2$ has fallen from 0.579 to 0.258 because we've omitted population. However, check out the estimated coefficient and t -score for competition (N). The coefficient has changed from -9075 to -1487, and the t -score has changed from -4.42 to -0.84. What a disaster! The coefficient of N now is insignificantly different from zero! How could this have happened?

The answer is omitted variable bias. Population and competition are understandably quite correlated; the more people there are in an area, the more restaurants you'd expect to find. As a result, when population is dropped from the equation, OLS attributes the impact of the omitted variable to the included variables to the extent that they're correlated with the omitted variable. Was this positive or negative bias? Well, $\hat{\beta}_N$ increased from a large negative number to a smaller negative number, so the bias is positive. The positive impact of population almost completely offset the negative impact of competition, resulting in a coefficient not far from zero. Note that we could have predicted that the bias was going to be positive by using our expected bias equation,³⁰ equation (7.7). Because the expected sign of β_P is positive and because we'd expect $\hat{a}_1$ (related to the correlation between population and competition) to be positive, the expected bias in $\hat{\beta}_N$ is positive:

$$\text{Expected bias in } \hat{\beta}_N = \beta_P \times \hat{a}_1 = (+)(+) = +\tag{7.9}$$

Just as we would have predicted, omitting N caused positive bias. Leaving population out of the equation caused the coefficient of competition to pick up the impact of population to the extent that the two variables were correlated.

³⁰ It is important to note the distinction between expected bias and any actual observed differences between coefficient estimates. Because of the random nature of the error term (and hence the $\hat{\beta}$ s), the change in an estimated coefficient brought about by dropping a relevant variable from the equation will not necessarily be in the expected direction. Biasedness refers to the central tendency of the sampling distribution of the β N s, not to every single drawing from that distribution. However, we usually (and justifiably) rely on these general tendencies.

To sum, if a relevant variable is left out of a regression equation,

1. there is no longer an estimate of the coefficient of that variable in the equation, and
2. the coefficients of the remaining variables are likely to be biased.

Although the amount of the bias might not be very large in some cases (when, for instance, there is little correlation between the included and excluded variables), it is extremely likely that at least a small amount of omitted variable bias will be present in all such situations.

Correcting for an omitted variable

In theory, the solution to a problem of omitted variable bias seems easy: Add the omitted variable to the equation! Unfortunately, that's easier said than done, for a couple of reasons.

First, omitted variable bias is hard to detect. The amount of bias introduced can be small and not immediately detectable. This is especially true when there is no reason to believe that you have misspecified the model. Some indications of specification bias are obvious (such as an estimated coefficient that is significant in the direction opposite from that expected), but others are not so clear. The best indicators of an omitted relevant variable are the theoretical underpinnings of the model itself. What variables *must* be included? What signs do you expect? Do you have any notions about the range into which the coefficient values should fall? The best way to avoid omitting an important variable is to invest the time in thinking carefully through the equation before the data are entered into the computer.

A second source of complexity is the problem of choosing which variable to add to an equation once you decide that it is suffering from omitted variable bias. Some beginning researchers, when faced with this dilemma, will add all the possible relevant variables to the equation at once, but this process leads to less precise estimates, as will be discussed in the next section. Other beginning researchers will test a number of different variables and keep the one in the equation that does the best statistical job of appearing to reduce the bias (by giving plausible signs and satisfactory *t*-values). This technique, adding a "left-out" variable to "fix" a strange-looking regression result, is invalid because the

variable that best corrects a case of specification bias might do so only by chance rather than by being the true solution to the problem. In such an instance, the “fixed” equation may give superb statistical results for the sample at hand but then do terribly when applied to other samples because it does not describe the characteristics of the true population.

Dropping a variable will not help cure omitted variable bias. If the sign of an estimated coefficient is different from expected, it cannot be changed to the expected direction by dropping a variable that has a t -score lower (in absolute value) than the t -score of the coefficient estimate that has the unexpected sign. Furthermore, the sign in general will not likely change even if the variable to be deleted has a large t -score.³¹ If an unexpected result leads you to believe that you have an omitted variable, one way to decide which variable to add to the equation is to use expected bias analysis. If the sign of the expected bias (using equation (7.7)) is the same as the sign of your unexpected result, then the variable might be the source of the apparent bias. If the sign of the expected bias is *not* the same as the sign of your unexpected result, however, then the variable is extremely unlikely to have caused your unexpected result. Expected bias analysis should be used only when you’re choosing between theoretically sound potential variables.

Although you can never actually observe bias (since you don’t know the true β), the use of this technique to screen potential causes of specification bias should reduce the number of regressions run and increase the validity of the results.

A brief warning: It may be tempting to conduct what might be called “residual analysis” by examining a plot of the residuals in an attempt to find patterns that suggest variables that have been accidentally omitted. A major problem with this approach is that the coefficients of the estimated equation will possibly have some of the effects of the left-out variable already altering their estimated values. Thus, residuals may show a pattern that only vaguely resembles the pattern of the actual omitted variable. The chances are high that the pattern shown in the residuals may lead to the selection of an incorrect variable. In addition, care should be taken to use

³¹ Ignazio Visco, “On obtaining the right sign of a coefficient estimate by omitting a variable from the regression,” *Journal of Econometrics*, Vol. 7, No. 1, pp. 115–117.

residual analysis only to choose between theoretically sound candidate variables rather than to generate those candidates.

7.2 Irrelevant variables

What happens if you include a variable in an equation that doesn't belong there? This case, **irrelevant variables**, is the converse of omitted variables and can be analyzed using the model we developed in section 7.1. The addition of a variable to an equation where it doesn't belong does not cause bias, but it does increase the variances of the estimated coefficients of the included variables.

Impact of Irrelevant variables

If the true regression specification is:

$$y_i = a + \beta_1 x_{1i} + u_i \quad (7.10)$$

but the researcher for some reason includes an extra variable,

$$y_i = a + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i^{**} \quad (7.11)$$

the misspecified equation's error term can be seen to be:

$$u_i^{**} = u_i - \beta_2 x_{2i} \quad (7.12)$$

Such a mistake won't cause bias if the true coefficient of the irrelevant variable is zero. That is, an estimate of β_1 in equation (7.11) is unbiased when $\beta_2 = 0$. However, the inclusion of an irrelevant variable will increase the variance of the estimated coefficients, and this increased variance will tend to decrease the absolute magnitude of their t -scores. Also, an irrelevant variable usually will decrease the $\bar{R}^2$ (but not the R^2).

Thus, although the irrelevant variable causes no bias, it causes exercises for the regression because it reduces the t -scores and $\bar{R}^2$.

EXAMPLE 7.2

Let's return to the Rohat's equation and see what happens when we add an irrelevant variable to the model. The original equation was:

$$\begin{aligned} \hat{Y}_i &= 102.192 - 9075N_i + 0.3547P_i + 1.288I_i \\ &\quad \begin{matrix} (2053) & (0.0727) & (0.543) \end{matrix} \\ t &= -4.42 \quad 4.88 \quad 2.37 \\ N &= 33 \quad \bar{R}^2 = 0.579 \end{aligned} \tag{6.8}$$

where Y = customers (check volume), N = the number of competitive restaurants nearby, P = the population nearby, and I = the average household income nearby.

What's the most irrelevant variable that you can think of? How about A_i = the last three digits of the street address of the i th Rohat's restaurant? That's pretty random! If we add A_i to equation (6.8), we get:

$$\begin{aligned} \hat{Y}_i &= 98.125 - 8975N_i + 0.360P_i + 1.301I_i + 58.07A_i \\ &\quad \begin{matrix} (2082) & (0.074) & (0.550) & (95.21) \end{matrix} \\ t &= -4.33 \quad + 4.88 \quad + 2.37 \quad + 0.61 \\ N &= 33 \quad \bar{R}^2 = 0.569 \end{aligned} \tag{7.13}$$

A comparison of equations (6.8) and (7.13) will make the theory in section 7.2 come to life. First of all, $\bar{R}^2$ has fallen slightly, indicating the reduction in fit adjusted for degrees of freedom. Second, none of the regression coefficients from the original equation changed very much; compare these results with the larger differences between equations (6.8) and (7.9). Further, the standard errors of the estimated coefficients increased. Finally, the t -score for the potential variable (A) is small, indicating that it is not significantly different from zero.

Given the theoretical shakiness of the new variable, these results indicate that it is irrelevant and never should have been included in the regression.

Four important specification criteria

We have now discussed at least four valid criteria to help decide whether a given variable belongs in the equation. We think these criteria are so important that we urge beginning researchers to work through them every time a variable is added or subtracted.

1. *Theory*: Is the variable's place in the equation unambiguous and theoretically sound?
2. *t-Test*: Is the variable's estimated coefficient significant in the expected direction?
3. $\bar{R}^2$: Does the overall fit of the equation (adjusted for degrees of freedom) improve when the variable is added to the equation?
4. *Bias*: Do other variables' coefficients change significantly when the variable is added to the equation?

If all these conditions hold, the variable belongs in the equation; if none of them do, the variable is irrelevant and can be safely excluded from the equation. When a typical omitted relevant variable is included in the equation, its inclusion probably will increase $\bar{R}^2$ and change at least one other coefficient.

If an irrelevant variable, on the other hand, is included, it will reduce $\bar{R}^2$, have an insignificant *t*-score, and have little impact on the other variables' coefficients. In many cases, all four criteria do not agree. It is possible for a variable to have an insignificant *t*-score that is greater than one, for example. In such a case, it can be shown that $\bar{R}^2$ will go up when the variable is added to the equation and yet the *t*-score still will be insignificant.

Whenever our four specification criteria disagree, the econometrician must use careful judgment and should not rely on a single criterion like $\bar{R}^2$ to determine the specification. Researchers should not misuse this freedom by testing various combinations of variables until they find the results that appear to statistically support the point they want to make. All such decisions are a bit easier when you realize that the single most important determinant of a variable's relevance is its theoretical justification. No amount of statistical evidence should make a theoretical necessity into an "irrelevant" variable. Once in a while, a researcher is forced to leave a theoretically important variable out of an equation for lack of data; in such cases, the usefulness of the equation is limited.

7.3 An illustration of the misuse of specification criteria

At times, the four specification criteria outlined in the previous section will lead the researcher to an incorrect conclusion if those criteria are applied to a problem without proper concern for economic principles or common sense. In particular, a t -score can often be insignificant for reasons other than the presence of an irrelevant variable. Since economic theory is the most important test for including a variable, an example of why a variable should not be dropped from an equation simply because it has an insignificant t -score is in order.

Suppose you believe that the demand for Brazilian coffee in the Tajikistan is a negative function of the real price of Brazilian coffee (P_{bc}) and a positive function of both the real price of tea (P_t) and real disposable income in the Tajikistan (Y_d). Suppose further that you obtain the data, run the implied regression, and observe the following results:

$$\begin{aligned} COFF\hat{E}E &= 9.1 + 7.8P_{bc} + 2.4P_t + 0.0035Y_d \\ &\quad \begin{matrix} (15.6) & (1.2) & (0.0010) \\ t = 0.5 & 2.0 & 3.5 \end{matrix} \\ \bar{R}^2 &= 0.60 \quad N = 25 \end{aligned} \tag{7.14}$$

The coefficients of the second and third variables, P_t and Y_d , appear to be fairly significant in the direction you hypothesized, but the first variable, P_{bc} , appears to have an insignificant coefficient with an unexpected sign. If you think there is a possibility that the demand for Brazilian coffee is price-inelastic (that is, its coefficient is zero), you might decide to run the same equation without the price variable, obtaining:

$$\begin{aligned} COFF\hat{E}E &= 9.3 + 2.6P_t + 0.0036Y_d \\ &\quad \begin{matrix} (1.0) & (0.0009) \\ t = 2.6 & 4.0 \end{matrix} \\ \bar{R}^2 &= 0.61 \quad N = 25 \end{aligned} \tag{7.15}$$

By comparing Equations 7.14 and 7.15, we can apply our four specification criteria for the inclusion of a variable in an equation that were outlined in the previous section:

1. *Theory*: If it's possible that the demand for coffee could be price-inelastic, the theory behind dropping the variable seems plausible.
2. *t-Test*: The *t*-score of the possibly irrelevant variable is 0.5, insignificant at any level.
3. $\bar{R}^2$: $\bar{R}^2$ increases when the variable is dropped, indicating that the variable is irrelevant.
4. *Bias*: The remaining coefficients change only a small amount when P_{bc} is dropped, suggesting that there is little - if any - bias caused by excluding the variable.

Based upon this analysis, you might conclude that the demand for Brazilian coffee is indeed price-inelastic and that the variable is therefore irrelevant and should be dropped from the model. As it turns out, this conclusion would be unwarranted. Although the elasticity of demand for coffee in general might be fairly low (actually, the evidence suggests that it is inelastic only over a particular range of prices), it is hard to believe that Brazilian coffee is immune to price competition from other kinds of coffee. Indeed, one would expect quite a bit of sensitivity in the demand for Brazilian coffee with respect to the price of, for example, Colombian coffee. To test this hypothesis, the price of Colombian coffee (P_{cc}) should be added to the original equation (7.14):

$$\begin{aligned}
 \hat{COFFEE} = & 10.0 + \underset{(4.0)}{8.0 P_{cc}} + \underset{(2.0)}{5.6 P_{bc}} + \underset{(1.3)}{2.6 P_t} + \underset{(0.0010)}{0.0030 Y_d} \\
 & t = 2.0 \quad -2.8 \quad 2.0 \quad 3.0 \\
 & \bar{R}^2 = 0.65 \quad N = 25
 \end{aligned}
 \tag{7.16}$$

By comparing equations (7.14) and (7.16), we can once again apply our four specification criteria:

1. *Theory*: Both prices should always have been included in the model; their logical justification is quite strong.

2. *t*-Test: The *t*-score of the new variable, the price of Colombian coffee, is 2.0, significant at most levels.
3. $\bar{R}^2$: $\bar{R}^2$ increases with the addition of the variable, indicating that the variable was an omitted variable.
4. *Bias*: Although two of the coefficients remain virtually unchanged, indicating that the correlations between these variables and the price of Colombian coffee variable are low, the coefficient for the price of Brazilian coffee does change significantly, indicating bias in the original result.

The moral to be drawn is that theoretical considerations never should be discarded, even in the face of statistical insignificance. If a variable known to be extremely important from a theoretical point of view turns out to be statistically insignificant in a particular sample, that variable should be left in the equation despite the fact that it makes the results look bad.

Don't conclude that the particular path outlined in this example is the correct way to specify an equation. Trying a long string of possible variables until you get the particular one that makes the coefficient of Pbc turn negative and significant is not the way to obtain a result that will stand up well to other samples or alternative hypotheses. The original equation should never have been run without the Colombian coffee price variable. Instead, the problem should have been analyzed enough so that such errors of omission were unlikely before any regressions were attempted at all. The more thinking that's done before the first regression is run, and the fewer alternative specifications that are estimated, the better the regression results are likely to be.

7.4.Specification Searches

One of the weaknesses of econometrics is that a researcher potentially can manipulate a data set to produce almost *any* result by specifying different regressions until estimates with the desired properties are obtained. Because the integrity of all empirical work is thus open to question, the subject of how to search for the best specification is quite controversial among econometricians. Our goal in this section isn't to summarize or settle this controversy;

instead, we hope to provide some guidance and insight for beginning researchers.

Best practices in specification searches

The issue of how best to choose a specification from among alternative possibilities is a difficult one, but our experience leads us to make the following recommendations:

1. Rely on theory rather than statistical fit as much as possible when choosing variables, functional forms, and the like.
2. Minimize the number of equations estimated (except for sensitivity analysis, to be discussed later in this section).
3. Reveal, in a footnote or appendix, all alternative specifications estimated.

If theory, not $\bar{R}^2$ or t -scores, is the most important criterion for the inclusion of a variable in a regression equation, then it follows that most of the work of specifying a model should be done before you attempt to estimate the equation. Since it's unreasonable to expect researchers to be perfect, there will be times when additional specifications must be estimated. However, these new estimates should be few in number and should be thoroughly grounded in theory. In addition, they should be explicitly taken into account when testing for significance and/or summarizing results. In this way, the danger of misleading the reader about the statistical properties of the final equation will be reduced.

Sequential specification searches

Most econometricians tend to specify equations by estimating an initial equation and then sequentially dropping or adding variables (or changing functional forms) until a plausible equation is found with “good statistics.” Faced with knowing that a few variables are relevant (on the basis of theory) but not knowing whether other additional variables also are relevant, the generally accepted practice appears to be inspecting $\bar{R}^2$ and t -tests for all variables for each specification. Indeed, casual reading of the previous section might make it seem as if such a sequential specification search is the best way to go about finding the “truth.”

Instead, as we shall see, there is a vast difference between a sequential specification search and our recommended approach.

The **sequential specification search** technique allows a researcher to estimate an undisclosed number of regressions and then present a final choice (which is based upon an unspecified set of expectations about the signs and significance of the coefficients) as if it were the only specification estimated. Such a method misstates the statistical validity of the regression results for two reasons:

1. The statistical significance of the results is overestimated because the estimations of the previous regressions are ignored.
2. The expectations used by the researcher to choose between various regression results rarely, if ever, are disclosed. Thus the reader has no way of knowing whether all the other regression results had opposite signs or insignificant coefficients for the important variables.

Unfortunately, there is no universally accepted way of conducting sequential searches, primarily because the appropriate test at one stage in the procedure depends on which tests previously were conducted, and also because the tests have been very difficult to invent.

Instead we recommend trying to keep the number of regressions estimated as low as possible; to focus on theoretical considerations when choosing variables or functional forms; and to document all the various specifications investigated. That is, we recommend combining parsimony (using theory and analysis to limit the number of specifications estimated) with disclosure (reporting all the equations estimated).

Not everyone agrees with our advice. Some researchers feel that the true model will show through if given the chance and that the best statistical results (including signs of coefficients, etc.) are most likely to have come from the true specification. In addition, reasonable people often disagree as to what the “true” model should look like. As a result, different researchers can look at the same data set and come up with very different “best” equations. Because this

can happen, the distinction between good and bad econometrics is not always as clear-cut as is implied by the previous paragraphs. As long as researchers have a healthy respect for the dangers inherent in specification searches, they are very likely to proceed in a reasonable way.

Bias caused by relying on the t -test or $\bar{R}^2$ to choose variables

In the previous section, we stated that sequential specification searches are likely to mislead researchers about the statistical properties of their results. In particular, the practice of dropping a potential independent variable simply because its coefficient has a low t -score or because it lowers $\bar{R}^2$ will cause systematic bias in the estimated coefficients (and their t -scores) of the remaining variables.

Let's say the hypothesized model is:

$$y_i = a + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \quad (7.17)$$

Assume further that, on the basis of theory, we are certain that x_1 belongs in the equation but that we are not as certain that x_2 belongs. Many inexperienced researchers use only the t -test on $\hat{\beta}_2$ to determine whether x_2 should be included. If this preliminary t -test indicates that $\hat{\beta}_2$ is significantly different from zero, then these researchers leave x_2 in the equation. If, however, the t -test does *not* indicate that $\hat{\beta}_2$ is significantly different from zero, then such researchers drop x_2 from the equation and consider y to be a function of x_1 .

Two kinds of mistakes can be made using such a system. First, x_2 sometimes can be left in the equation when it does not belong there, but such a mistake does not change the expected value of $\hat{\beta}_1$.

Second, x_2 sometimes can be dropped from the equation when it belongs. In this second case, the estimated coefficient of x_1 will be biased. In other words, $\hat{\beta}_1$ will be biased every time x_2 belongs in the equation and is left out, and x_2 will be left out every time that

its estimated coefficient is not significantly different from zero. We will have systematic bias in our equation!

To summarize, the t -test is biased by sequential specification searches. Since most researchers consider a number of different variables before settling on the final model, someone who relies on the t -test or $\bar{R}^2$ is likely to encounter this problem systematically.

Data mining

Data mining involves estimating a wide variety of alternative specifications *before* a “best” equation is chosen. Readers of this text will not be surprised to hear that we urge extreme caution when data mining. Improperly done data mining is worse than doing nothing at all.

Done properly, data mining involves exploring a data set not for the purpose of testing hypotheses or finding a specification, but for the purpose of uncovering empirical regularities that can inform economic theory.³² After all, we can’t expect economic theorists to think of everything! *Be careful, however!* If you develop a hypothesis using data mining techniques, you must test that hypothesis on a *different* data set (or in a different context) than the one you used to develop the hypothesis. A new data set must be used because our typical statistical tests have little meaning if the new hypothesis is tested on the data set that was used to generate it. After all, the researcher already knows ahead of time what the results will be! The use of dual data sets is easiest when there is a plethora of data. This sometimes is the case in cross-sectional research projects but rarely is the case for time series research.

Data mining without using dual data sets is almost surely the worst way to choose a specification. In such a situation, a researcher could estimate virtually every possible combination of the various alternative independent variables, could choose the results that “look” the best, and then could report the “best” equation as if no data mining had been done. This improper use of data mining

³² For an excellent presentation of this approach, see Lawrence H. Summers, “The Scientific Illusion in Empirical Macroeconomics,” *Scandinavian Journal of Economics*, Vol. 93, No. 2, pp. 129–148.

ignores the fact that a number of specifications have been examined before the final one is reported.

In addition, data mining will cause you to choose specifications that reflect the peculiarities of your particular data set. How does this happen? Suppose you have 100 true null hypotheses and you run 100 tests of these hypotheses. At the 5% level of significance, you'd expect to reject about five true null hypotheses and thus make about five type I errors. By looking for high t -values, a data mining search procedure will find these type I errors and incorporate them into your specification. As a result, the reported t -scores will overstate the statistical significance of the estimated coefficients. In essence, improper data mining to obtain desired statistics for the final regression equation is a potentially unethical empirical research method. Whether the improper data mining is accomplished by estimating one equation at a time or by estimating batches of equations or by techniques like stepwise regression procedures,⁸ the conclusion is the same. Hypotheses developed by data mining should always be tested on a data set different from the one that was used to develop the hypothesis. Otherwise, the researcher hasn't found any scientific evidence to support the hypothesis; rather, a specification has been chosen in a way that is essentially misleading. As put by one econometrician, "if you torture the data long enough, they will confess."³³

Sensitivity analysis

Throughout this text, we've encouraged you to estimate as few specifications as possible and to avoid depending on fit alone to choose between those specifications. If you read the current economics literature, however, it won't take you long to find well-known researchers who have estimated five or more specifications and then have listed all their results in an academic journal article. What's going on?

³³ Thomas Mayer, "Economics as a Hard Science: Realistic Goal or Wishful Thinking?" *Economic Inquiry*, Vol. 18, No. 2, p. 175. (This quote also has been attributed to Ronald Coase.)

In almost every case, these authors have employed a technique called sensitivity analysis. **Sensitivity analysis** consists of purposely running a number of alternative specifications to determine whether particular results are *robust* (not statistical flukes). In essence, we're trying to determine how sensitive a potential "best" equation is to a change in specification because the true specification isn't known. Researchers who use sensitivity analysis run (and report on) a number of different reasonable specifications and tend to discount a result that appears significant in some specifications and insignificant in others. Indeed, the whole purpose of sensitivity analysis is to gain confidence that a particular result is significant in a variety of alternative specifications, functional forms, variable definitions, and/or subsets of the data.

EXAMPLE 7.3

An example of choosing independent variables

It's time to get some experience choosing independent variables. After all, every equation so far in the text has come with the specification already determined, but once you've finished this course you'll have to make all such specification decisions on your own. In future chapters, we'll use a technique called "interactive regression learning exercises" to allow you to make your own actual specification choices and get feedback on your choices. To start, though, let's work through a specification together.

To keep things as simple as possible, we'll begin with a topic near and dear to your heart - your GPA! Suppose a friend who attends a small liberal arts college surveys all 25 members of her econometrics class, obtains data on the variables listed here, and asks for your help in choosing a specification:

GPA_i = the cumulative college grade point average of the i th student on a four-point scale

$HGPA_i$ = the cumulative high school grade point average of the i th student on a four-point scale

$MSAT_i$ = the highest score earned by the i th student on the math section of the SAT test (800 maximum)

$VSAT_i$ = the highest score earned by the i th student on the verbal section of the SAT test (800 maximum)

$SAT_i = MSAT_i + VSAT_i$

$GREK_i$ = a dummy variable equal to 1 if the i th student is a member of a fraternity or sorority, 0 otherwise

HRS_i = the i th student's estimate of the average number of hours spent studying per course per week in college

$PRIV_i$ = a dummy variable equal to 1 if the i th student graduated from a private high school, 0 otherwise

$JOCK_i$ = a dummy variable equal to 1 if the i th student is or was a member of a varsity intercollegiate athletic team for at least one season, 0 otherwise

$\ln EX_i$ = the natural log of the number of full courses that the i th student has completed in college

Assuming that GPA_i is the dependent variable, which independent variables would you choose? Before you answer, think through the possibilities carefully. What does the literature tell us on this subject? (Is there literature?) What are the expected signs of each of the coefficients? How strong is the theory behind each variable? Which variables seem obviously important? Which variables seem potentially irrelevant or redundant? Are there any other variables that you wish your friend had collected?

To get the most out of this example, you should take the time to *write down* the exact specification that you would run:

$$GPA_i = f(?, ?, ?, ?) + u$$

It's hard for most beginning econometricians to avoid the temptation of including *all* of these variables in a GPA equation and then dropping any variables that have insignificant t -scores. Even though we mentioned in the previous section that such a specification search procedure will result in biased coefficient estimates, most beginners don't trust their own judgment and tend to include too many variables. With this warning in mind, do you want to make any changes in your proposed specification? No? OK, let's compare notes. We believe that grades are a function of a student's ability, how hard the student works, and the student's experience taking college courses. Consequently, our specification would be:

$$GPA_i = a + \beta_1^+ HGPA_i + \beta_2^+ HRS_i + \beta_3^+ \ln EX_i + u_i$$

We can already hear you complaining! What about SATs, you say? Everyone knows they're important. How about jocks and Greeks? Don't they have lower GPAs? Don't prep schools grade harder and prepare students better than public high schools?

Before we answer, it's important to note that we think of specification choice as choosing which variables to *include*, not which variables to *exclude*. That is, we don't assume automatically that a given variable should be

SPECIFICATION: CHOOSING THE INDEPENDENT VARIABLES

included in an equation simply because we can't think of a good reason for dropping it.

Given that, however, why did we choose the variables we did? First, we think that the best predictor of a student's college *GPA* is his or her high school *GPA*. We have a hunch that once you know *HGPA*, *SATs* are redundant, at least at a liberal arts college³⁴ where there are few multiple choice tests. In addition, we're concerned that possible racial and gender bias in the SAT test makes it a questionable measure of academic potential, but we recognize that we could be wrong on this issue.

As for the other variables, we're more confident. For example, we feel that once we know how many hours a week a student spends studying, we couldn't care less what that student does with the rest of his or her time, so *JOCK* and *GREK* are superfluous once *HRS* is included. In addition, the higher *lnEX* is, the better student study habits are and the more likely students are to be taking courses in their major. Finally, while we recognize that some private schools are superb and that some public schools are not, we'd guess that *PRIV* is irrelevant; it probably has only a minor effect. If we estimate this specification on the 25 students, we obtain:

$$\begin{aligned} \hat{GPA}_i &= -0.26 + 0.49 \underset{(0.21)}{HGPA}_i + 0.06 \underset{(0.02)}{HRS}_i + 0.42 \underset{(0.14)}{\ln EX}_i \\ t &= 2.33 \qquad 3.00 \qquad 3.00 \\ N &= 25 \qquad \bar{R}^2 = 0.585 \end{aligned} \tag{7.18}$$

Since we prefer this specification on theoretical grounds, since the overall fit seems reasonable, and since each coefficient meets our expectations in terms of sign, size, and significance, we consider this an acceptable equation. The only circumstance under which we'd consider estimating a second specification would be if we had theoretical reasons to believe that we had omitted a relevant variable. The only variable that might meet this description is *SAT_i* (which we prefer to the individual *MSAT* and *VSAT*):

$$\begin{aligned} \hat{GPA}_i &= -0.92 + 0.47 \underset{(0.22)}{HGPA}_i + 0.05 \underset{(0.02)}{HRS}_i + 0.44 \underset{(0.14)}{\ln EX}_i + 0.00060 \underset{(0.00064)}{SAT}_i \\ t &= 2.12 \qquad 2.50 \qquad 3.12 \qquad 0.93 \\ N &= 25 \qquad \bar{R}^2 = 0.583 \end{aligned} \tag{7.19}$$

³⁴ In contrast, SATs tend to have a statistically significant effect on GPAs at large research universities. For example, see Andrew Barkley and Jerry Forst, "The Determinants of First-Year Academic Performance in the College of Agriculture at Kansas State University, 1990–1999," *Journal of Agricultural and Applied Economics*, Vol. 36, No 2, pp. 437–448.

Let's use our four specification criteria to compare equations (7.18) and (7.19):

1. *Theory*: As discussed previously, the theoretical validity of *SAT* tests is a matter of some academic controversy, but they still are one of the most-cited measures of academic potential in this country.
2. *t-Test*: The coefficient of *SAT* is positive, as we'd expect, but it's not significantly different from zero.
3. $\bar{R}^2$: As you'd expect (since *SAT*'s *t*-score is under 1), $\bar{R}^2$ falls slightly when *SAT* is added.
4. *Bias*: None of the estimated slope coefficients changes substantially when *SAT* is added, though some of the *t*-scores do change because of the increase in the $SE(\hat{\beta})$ s caused by the addition of *SAT*.

Thus, the statistical criteria do not convincingly contradict our theoretical contention that *SAT* is irrelevant. Finally, it's important to recognize that different researchers could come up with different final equations on this topic. A researcher whose prior expectation was that *SAT* unambiguously belonged in the equation would have estimated equation (7.19) and accepted that equation without bothering to estimate equation (7.18). Other researchers, in the spirit of sensitivity analysis, would report both equations.

SUMMARY

1. The omission of a variable from an equation will cause bias in the estimates of the remaining coefficients to the extent that the omitted variable is correlated with included variables.
2. The bias to be expected from leaving a variable out of an equation equals the coefficient of the excluded variable times a function of the simple correlation coefficient between the excluded variable and the included variable in question.
3. Including a variable in an equation in which it is actually irrelevant does not cause bias, but it will usually increase the variances of the included variables' estimated coefficients, thus lowering their *t*-values, widening their confidence intervals, and lowering $\bar{R}^2$.
4. Four useful criteria for the inclusion of a variable in an equation are:
 - a. Theory
 - b. *t*-test

- c. $\bar{R}^2$
- d. Bias

5. Theory, not statistical fit, should be the most important criterion for the inclusion of a variable in a regression equation. To do otherwise runs the risk of producing incorrect and/or disbelieved results.

APPENDIX: 7

7A. Additional specification criteria

So far in this chapter, we've suggested four criteria for choosing the independent variables (economic theory, $\bar{R}^2$, the t -test, and possible bias in the coefficients). Sometimes, however, these criteria don't provide enough information for a researcher to feel confident that a given specification is best. For instance, there can be two different specifications that both have excellent theoretical underpinnings. In such a situation, many econometricians use additional, often more formal, specification criteria to provide comparisons of the properties of the alternative estimated equations.

The use of formal specification criteria is not without exercises, however. First, no test, no matter how sophisticated, can "prove" that a particular specification is the true one. The use of specification criteria, therefore, must be tempered with a healthy dose of economic theory and common sense. A second problem is that more than 20 such criteria have been proposed. How do we decide which one(s) to use? Because many of these criteria overlap with one another or have varying levels of complexity, a choice between the alternatives is a matter of personal preference.

In this section, we'll describe the use of three of the most popular specification criteria, J. B. Ramsey's RESET test, Akaike's Information Criterion, and the Bayesian Information Criterion. Our inclusion of just these techniques does not imply that other tests and criteria are not appropriate or useful. Indeed, the reader will find that most other formal specification criteria have quite a bit in common with at least one of the techniques that we include. We

think that you'll be better able to use and understand other formal specification criteria³⁵ once you've mastered these three.

Ramsey's regression specification error test (RESET)

One of the most-used formal specification criteria other than $\bar{R}^2$ is the Ramsey regression specification error test (RESET).³⁶ The **Ramsey RESET test** is a general test that determines the likelihood of specification error by measuring whether the fit of a given equation can be significantly improved by the addition of $\hat{y}^2$, $\hat{y}^3$, and $\hat{y}^4$ terms. What's the intuition behind RESET? The additional terms act as proxies for any possible (unknown) omitted variables or incorrect functional forms. If the proxies can be shown by the F -test to have improved the overall fit of the original equation, then we have evidence that there is some sort of specification error in our equation. The $\hat{y}^2$, $\hat{y}^3$, and $\hat{y}^4$ terms form a *polynomial* functional form. Such a polynomial is a powerful curve-fitting device that has a good chance of acting as a proxy for a specification error if one exists. If there is no specification error, then we'd expect the coefficients of the added terms to be insignificantly different from zero because there is nothing for them to act as a proxy for.

The Ramsey RESET test involves three steps:

1. Estimate the equation to be tested using OLS:

$$\hat{Y}_i = \hat{a} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} \quad (7.22)$$

2. Take the $\hat{Y}_i$ values from equation (7.22) and create $\hat{Y}_i^2$, $\hat{Y}_i^3$, and $\hat{Y}_i^4$ terms. Then add these terms to the original equation as additional explanatory variables and estimate the new equation with OLS:

$$\hat{Y}_i = \hat{a} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \beta_3 \hat{Y}_i^2 + \beta_4 \hat{Y}_i^3 + \beta_5 \hat{Y}_i^4 + u_i \quad (7.23)$$

³⁵ For example, the likelihood ratio test can be used as a specification test. For an introductory level summary of six other specification criteria, see Ramu Ramanathan, *Introductory Econometrics* (Fort Worth: Harcourt Brace Jovanovich, 1998, pp. 164–166).

³⁶ J. B. Ramsey, "Tests for Specification Errors in Classical Linear Squares Regression Analysis," *Journal of the Royal Statistical Society*, Vol. 31, No. 2, pp. 350–371.

3. Compare the fits of equations (7.22) and (7.23), using the F -test. Specifically, test the hypothesis that the coefficients of all three of the added terms are equal to zero:

$$H_0: \beta_3 = \beta_4 = \beta_5 = 0$$

H_1 : otherwise

If the two equations are significantly different in overall fit, we can conclude that it's likely that equation (7.22) is misspecified.

The appropriate F -statistic to use is equation (7.9) from Section 7.4:

$$F = \frac{(RSS_M - RSS) / M}{RSS / (N - K - 1)} \quad (7.9)$$

where RSS_M is the residual sum of squares from the constrained equation (equation 7.22), RSS is the residual sum of squares from the unconstrained equation³⁷ (equation 7.23), M is the number of constraints (3 in this case), and $(N - K - 1)$ is the number of degrees of freedom in the unconstrained equation. If F is greater than the critical F -value with M and $(N - K - 1)$ degrees of freedom, then we can reject the null hypothesis and conclude that there is a specification error in equation (7.22). Many econometric software programs, including Eviews, have a command that will automatically run equation (7.23) and calculate the F -statistic using equation (7.9). While the Ramsey RESET test is fairly easy to use, it does little more than signal *when* a major specification error might exist. If you encounter a significant Ramsey RESET test, then you face the daunting task of figuring out exactly *what* the error is! Thus, the test often ends up being more useful in “supporting” (technically, not refuting) a researcher’s contention that a given

³⁷ Because of the obvious correlation between the three YN values, equation 7.23 (with most RESET equations) suffers from extreme multicollinearity. Since the purpose of the RESET equation is to see whether the overall fit can be improved by adding in proxies for an omitted variable (or other specification error), this extreme multicollinearity is not a concern.

specification has no major specification errors than it is in helping find an otherwise undiscovered flaw.³⁸

Akaike's information criterion and the bayesian information criterion

A second category of formal specification criteria involves adjusting the summed squared residuals (RSS) by one factor or another to create an index of the fit of an equation. The most popular criterion of this type is $\bar{R}^2$, but a number of interesting alternatives have been proposed. **Akaike's Information Criterion (AIC)** and the **Bayesian Information Criterion (BIC)** are methods of comparing alternative specifications by adjusting RSS for the sample size (N) and the number of independent variables (K). These criteria can be used to augment our four basic specification criteria when we try to decide if the improved fit caused by an additional variable is worth the decreased degrees of freedom and increased complexity caused by the addition. Their equations are:

$$AIC = \text{Log}(RSS/N) + 2(K + 1)/N \quad (7.24)$$

$$BIC = \text{Log}(RSS/N) + \text{Log}(N)(K + 1)/N \quad (7.25)$$

In practice, these calculations may not be necessary because AIC and BIC are automatically calculated by some regression software packages, including Eviews. To use AIC and BIC , estimate two alternative specifications and calculate AIC and BIC for each equation. The lower AIC or BIC are, the better the specification. Note that even though the two criteria were developed

³⁸ The particular version of the Ramsey RESET test we describe in this section is only one of a number of possible formulations of the test. For example, some researchers delete the $\hat{Y}^4$ term from equation (7.23). At present, there is a mild split among econometricians about RESET. Jeff Wooldridge, "Score Diagnostics for Linear Models Estimated by Two Stage Least Squares," in G.S. Maddala, P.C.B. Phillips, and T.N. Srinivasan (eds.), *Advances in Econometrics and Quantitative Economics* (Oxford: Blackwell, 1995), pp. 66–87, argues that RESET is primarily a functional form test. However, many applied econometricians continue to rely on RESET for a variety of specification tests, some even going so far as to use RESET in an attempt to distinguish between pure and impure serial correlation and heteroskedasticity (to be discussed in Chapters 10 and 11).

independently to maximize different object functions, their equations are quite similar. Both criteria tend to penalize the addition of another explanatory variable more than $\bar{R}^2$ does. As a result, *AIC* and *BIC* will quite often³⁹ be minimized by an equation with fewer independent variables than the ones that maximize $\bar{R}^2$. *AIC* and *BIC* require the researcher to come up with a particular alternative specification, whereas Ramsey's RESET does not. Such a distinction makes RESET easier to use, but it makes *AIC* and *BIC* more informative if a specification error is found. Thus our additional specification criteria serve different purposes. RESET is useful as a general test of the existence of a specification error, whereas *AIC* and *BIC* are useful as means of comparing two or more alternative specifications.

³⁹ Using a Monte Carlo study, Judge et al. showed that (given specific simplifying assumptions) a specification chosen by maximizing R^2 is more than 50 percent more likely to include an irrelevant variable than is one chosen by minimizing *AIC* or *BIC*. See George C. Judge, R. Carter Hill, W. E. Griffiths, Helmut Lutkepohl, and Tsoung-Chao Lee, *Introduction to the Theory and Practice of Econometrics* (New York: Wiley, 1988), pp. 849–850. At the same time, minimizing *AIC* or *BIC* will omit a relevant variable more frequently than will maximizing $\bar{R}^2$

CHAPTER 8. SPECIFICATION: CHOOSING A FUNCTIONAL FORM

Even after you've chosen your independent variables, the job of specifying the equation is not over. The next step is to choose the functional form of the relationship between each independent variable and the dependent variable. Should the equation go through the origin? Do you expect a curve instead of a straight line? Does the effect of a variable peak at some point and then start to decline? An affirmative answer to any of these questions suggests that an equation other than the standard linear model of the previous chapters might be appropriate. Such alternative specifications are important for two reasons: a correct explanatory variable may well appear to be insignificant or to have an unexpected sign if an inappropriate functional form is used, and the consequences of an incorrect functional form for interpretation and forecasting can be severe.

Theoretical considerations usually dictate the form of a regression model. The basic technique involved in deciding on a functional form is to choose the shape that best exemplifies the expected underlying economic or business principles and then to use the mathematical form that produces that shape. To help with that choice, this chapter contains plots of the most commonly used functional forms along with the mathematical equations that correspond to each.

The chapter begins with a brief discussion of the constant term. In particular, we suggest that the constant term should be retained in equations and that estimates of the constant term should not be relied on for inference or analysis. The chapter concludes with a discussion of slope dummy variables.

8.1 The use and interpretation of the constant term

In the linear regression model, a is the intercept or constant term. It is the expected value of y when all the explanatory variables (and the error term) equal zero. An estimate of a has at least three components:

1. the true a ,
2. the constant impact of any specification errors (an omitted variable, for example), and
3. the mean of u for the correctly specified equation (if not equal to zero).

Unfortunately, these components can't be distinguished from one another because we can observe only $\hat{a}$, the sum of the three components. The result is that we have to analyze $\hat{a}$ differently from the way we analyze the other coefficients in the equation.

At times, a is of theoretical importance. Consider, for example, the following cost equation:

$$c_i = a + \beta_1 Q_i + u_i$$

where c_i is the total cost of producing output Q_i . The term $\beta_1 Q_i$ represents the total variable cost associated with output level Q_i , and a represents then total fixed cost, defined as the cost when output $Q_i = 0$. Thus, a regression equation might seem useful to a researcher who wanted to determine the relative magnitudes of fixed and variable costs. This would be an example of relying on the constant term for inference.

On the other hand, the product involved might be one for which it is known that there are few - if any - fixed costs. In such a case, a researcher might want to eliminate the constant term; to do so would conform to the notion of zero fixed costs and would conserve a degree of freedom (which would presumably make the estimate of β_1 more precise). This would be an example of suppressing the constant term.

Neither suppressing the constant term nor relying on it for inference is advisable, however, and the reasons for these conclusions are explained in the following sections.

Do not suppress the constant term

In most cases, suppressing the constant term leads to a violation of the classical assumptions, because it's very rare that economic theory implies that the true intercept, a , equals zero. If you omit the constant term, then the impact of the constant is forced into the estimates of the other coefficients, causing potential bias. This is demonstrated in Figure 8.1. Given the pattern of the x and y observations, estimating a regression equation with a constant term would likely produce an estimated regression line very similar to the true regression line, which has a constant term (a) quite different from zero. The slope of this estimated line is very low, and the t -score of the estimated slope coefficient may be very close to zero.

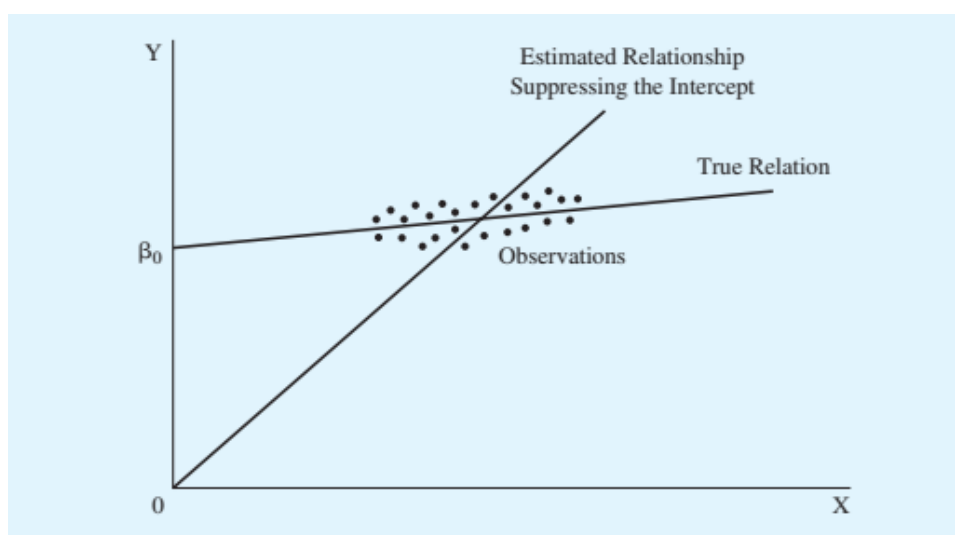


Figure 8.1 The harmful effect of suppressing the constant term.

If the constant (or intercept) term is suppressed, the estimated regression will go through the origin. Such an effect potentially biases the $\hat{\beta}$ s and inflates their t -scores. In this particular example, the true slope is close to zero in the range of the sample, but forcing the regression through the origin makes the slope appear to be significantly positive.

However, if the researcher were to suppress the constant term, which implies that the estimated regression line must pass through the origin, then the estimated regression line shown in Figure 8.1 would result. The slope coefficient is biased upward compared with the true slope coefficient. The t -score is biased upward as well, and

it may very well be large enough to indicate that the estimated slope coefficient is statistically significantly positive. Such a conclusion would be incorrect.

Thus, even though some regression packages allow the constant term to be suppressed (set to zero), the general rule is: *Don't!*

Do not rely on estimates of the constant term

It would seem logical that if it's a bad idea to suppress the constant term, then the constant term must be an important analytical tool to use in evaluating the results of the regression. Unfortunately, there are at least two reasons that suggest that the intercept should *not* be relied on for purposes of analysis or inference.

First, the error term is generated, in part, by the omission of a number of marginal independent variables, the mean effect of which is placed in the constant term. The constant term acts as a garbage collector, with an unknown amount of this mean effect being dumped into it. The constant term's estimated coefficient may be different from what it would have been without performing this task, which is done for the sake of the equation as a whole. As a result, it's meaningless to run a t -test on $\hat{a}$.

Second, the constant term is the value of the dependent variable when all the independent variables and the error term are zero, but the variables used for economic analysis are usually positive. Thus, the origin often lies *outside* the range of sample observations (as can be seen in Figure 8.1). Since the constant term is an estimate of y when the x s are outside the range of the sample observations, estimates of it are tenuous.

8.2 Alternative functional forms

The choice of a functional form for an equation is a vital part of the specification of that equation. Before we can talk about those functional forms, however, we need to make a distinction between an equation that is linear in the coefficients and one that is linear in the variables. An equation is **linear in the variables** if plotting the

function in terms of x and y generates a straight line. For example, equation (8.1):

$$y = a + \beta_1 x + u \quad (8.1)$$

is linear in the variables, but equation (8.2):

$$y = a + \beta_1 x^2 + u \quad (8.2)$$

is not linear in the variables, because if you were to plot equation (8.2) it would be a quadratic, not a straight line. An equation is **linear in the coefficients** only if the coefficients (the β s) appear in their simplest form - they are not raised to any powers (other than one), are not multiplied or divided by other coefficients, and do not themselves include some sort of function (like logs or exponents). For example, equation (8.1) is linear in the coefficients, but equation (8.3):

$$y = a + x^{\beta_1} \quad (8.3)$$

is not linear in the coefficients a and β_1 . Equation (8.3) is not linear because there is no rearrangement of the equation that will make it linear in the β s of original interest, a and β_1 . In fact, of all possible equations for a single explanatory variable, *only* functions of the general form:

$$f(y) = a + \beta_1 f(x) \quad (8.4)$$

are linear in the coefficients a and β_1 . Linear regression analysis can be applied to an equation that is nonlinear in the variables as long as the equation is linear in the coefficients. Indeed, when econometricians use the phrase “linear regression” (for example, in the classical assumptions), they usually mean “regression that is linear in the coefficients.”

The use of OLS requires that the equation be linear in the coefficients, but there is a wide variety of functional forms that are linear in the coefficients while being nonlinear in the variables.

Indeed, in previous chapters we've already used several equations that are linear in the coefficients and nonlinear in the variables, but we've said little about when to use such nonlinear equations. The purpose of the current section is to present the details of the most frequently used functional forms to help the reader develop the ability to choose the correct one when specifying an equation.

The choice of a functional form almost always should be based on the underlying theory and only rarely on which form provides the best fit. The logical form of the relationship between the dependent variable and the independent variable in question should be compared with the properties of various functional forms, and the one that comes closest to that underlying theory should be chosen. To allow such a comparison, the paragraphs that follow characterize the most frequently used forms in terms of graphs, equations, and examples. In some cases, more than one functional form will be applicable, but usually a choice between alternative functional forms can be made on the basis of the information we'll present.

Linear form

The linear regression model, used almost exclusively in this text thus far, is based on the assumption that the slope of the relationship between the independent variable and the dependent variable is constant:⁴⁰

$$\frac{\Delta y}{\Delta x_k} = \beta_k \quad k = 1, 2, \dots, K$$

If the hypothesized relationship between y and x is such that the slope of the relationship can be expected to be constant, then the linear functional form should be used.

⁴⁰ Throughout this section, the "delta" notation (Δ) will be used instead of the calculus notation to make for easier reading. The specific definition of Δ is "change" and it implies a small change in the variable it is attached to. For example, the term Δx should be read as "change in x ." Since a regression coefficient represents the change in the expected value of y brought about by a one-unit increase in x_k (holding constant all other variables in the equation), then $\beta_k = \Delta y / \Delta x_k$. Those comfortable with calculus should substitute partial derivative signs for Δ s.

Since the slope is constant, the **elasticity** of y with respect to x (the percentage change in the dependent variable caused by a 1% increase in the independent variable, holding the other variables in the equation constant) can be calculated fairly easily:

$$elasticity_{y,x_k} = \frac{\Delta y / y}{\Delta x_k / x_k} = \frac{\Delta y}{\Delta x_k} \times \frac{x_k}{y} = \beta_k \frac{x_k}{y}$$

Unless theory, common sense, or experience justifies using some other functional form, you should use the linear model. Because, in effect, it's being used by default, the linear model is sometimes referred to as the *default* functional form.

Double-log form

The double-log form is the most common functional form that is nonlinear in the variables while still being linear in the coefficients. Indeed, the double-log form is so popular that some researchers use it as their default functional form instead of the linear form. In a **double-log functional form**, the natural log of y is the dependent variable and the natural log of x is the independent variable:

$$\ln y = a + \beta_1 \ln x_1 + \beta_2 \ln x_2 + u \quad (8.5)$$

where $\ln y$ refers to the natural log of y , and so on.

The double-log form, sometimes called the log-log form, often is used because a researcher has specified that the elasticities of the model are constant and the slopes are not. This is in contrast to the linear model, in which the slopes are constant but the elasticities are not. In a double-log equation, an individual regression coefficient can be interpreted as an elasticity because:

$$\beta_k = \frac{\Delta(\ln y)}{\Delta(\ln x_k)} = \frac{\Delta y / y}{\Delta x_k / x_k} = elasticity_{y,x_k} \quad (8.6)$$

Since regression coefficients are constant, the condition that the model have a constant elasticity is met by the double-log equation.

SPECIFICATION: CHOOSING A FUNCTIONAL FORM

The way to interpret β_k in a double-log equation is that if x_k increases by 1% while the other x s are held constant, then y will change by β_k percent. Since elasticities are constant, the slopes are now no longer constant.

Figure 8.2 is a graph of the double-log function (ignoring the error term). The panel on the left shows the economic concept of an isoquant or an indifference curve. Isoquants from production functions show the different combinations of factors x_1 and x_2 , probably capital and labor, that can be used to produce a given level of output y . The panel on the right of Figure 8.2 shows the relationship between y and x_1 that would exist if x_2 were held

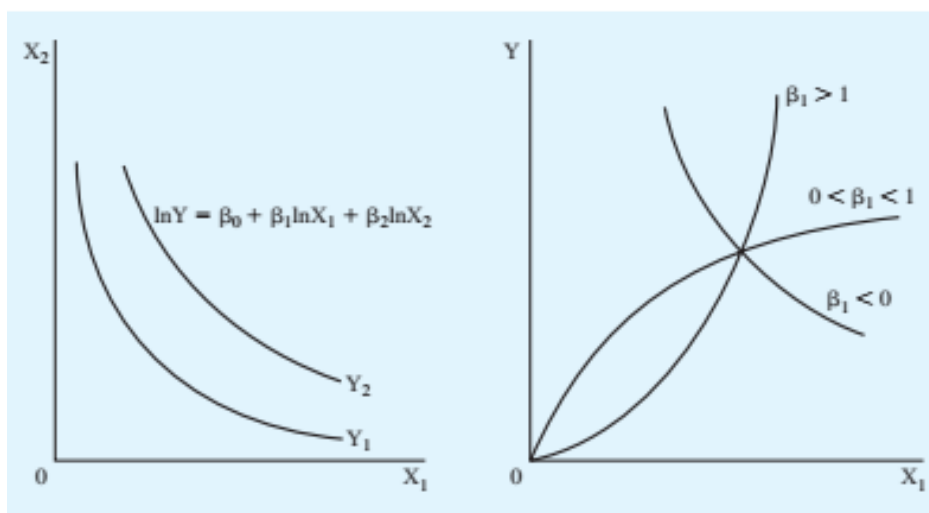


Figure 8.2 Double-log functions

Depending on the values of the regression coefficients, the double-log functional form can take on a number of shapes. The left panel shows the use of a double-log function to depict a shape useful in describing the economic concept of an isoquant or an indifference curve. The right panel shows various shapes that can be achieved with a double-log function if x_2 is held constant or is not included in the equation.

constant or were not included in the model. Note that the shape of the curve depends on the sign and magnitude of coefficient β_1 . If β_1 is negative, a double-log functional form can be used to model a typical demand curve.

Double-log models should be run only when the logged variables take on positive values. Dummy variables, which can take on the value of zero, should not be logged.

Box 8.1

If e (a constant equal to 2.71828) to the “ b th power” produces x , then b is the log of x :

b is the log of x to the base e if: $e^b = x$

Thus, a **log** (or logarithm) is the exponent to which a given base must be taken in order to produce a specific number. While logs come in more than one variety, we’ll use only **natural logs** (logs to the base e) in this text. The symbol for a natural log is “ $\ln$ ” so $\ln(x) = b$ means that $(2.71828)^b = x$ or, more simply,

$\ln(x) = b$ means that $e^b = x$

For example, since $e^2 = 2.7182822 = 7.389$, we can state that:

$\ln(7.389) = 2$

Thus, the natural log of 7.389 is 2! Two is the power of e that produces 7.389. Let’s look at some other natural log calculations:

$\ln(100) = 4.605$
 $\ln(1000) = 6.908$
 $\ln(10000) = 9.210$
 $\ln(100000) = 11.513$
 $\ln(1000000) = 13.816$

Note that as a number goes from 100 to 1000000 its natural log goes from 4.605 to only 13.816! Since logs are exponents, even a small change in a log can mean a big change in impact. As a result, logs can be used in econometrics if a researcher wants to reduce the absolute size of the numbers associated with the same actual meaning.

One useful property of natural logs in econometrics is that they make it easier to Figure out impacts in percentage terms. If you run a double-log regression, the meaning of a slope coefficient is the percentage change in the

dependent variable caused by a one percentage point increase in the independent variable, holding the other independent variables in the equation constant.¹ It's because of this percentage change property that the slope coefficients in a double-log equation are elasticities.

¹ This is because the derivative of a natural log of X equals dX/X (or ΔX/X), which is the same as percentage change.

Semilog form

The **semilog functional form** is a variant of the double-log equation in which some but not all of the variables (dependent and independent) are expressed in terms of their natural logs. For example, you might choose to use the logarithm of one of the original independent variables, as in:

$$y_i = a + \beta_1 \ln x_{1i} + \beta_2 x_{2i} + u_i \quad (8.7)$$

In this case, the economic meanings of the two slope coefficients are different, since x_2 is linearly related to y while x_1 is nonlinearly related to y .

The right-hand side of Figure 8.3 shows the relationship between y and x_1 in this kind of semilog equation when x_2 is held constant. Note that if β_1 is greater than zero, the impact of changes in x_1 on y decreases as x_1 gets bigger. Thus, the semilog functional form should be used when the relationship between x_1 and y is hypothesized to have this “increasing at a decreasing rate” form.⁴¹ Applications of the semilog form are quite frequent. For example, most consumption functions tend to increase at a decreasing rate past some level of income. These *Engel curves* tend to flatten out

⁴¹ Another functional form that can be used when you anticipate that the relationship between x and y has an “increasing at a decreasing rate” shape is the inverse functional form. This form expresses Y as a function of the reciprocal (or inverse) of one or more of the independent variables (in this case x_1): $y_i = a + \beta_1(1/x_{1i}) + \beta_2 x_{2i} + u_i$. The inverse functional form should be used when the impact of a particular independent variable is expected to approach zero as that independent variable approaches infinity. To see this, note that as x_1 gets larger, its impact on y decreases. Be careful, however, because x_1 cannot equal zero, since if x_1 equaled zero, dividing it into anything would result in infinite or undefined values.

because as incomes get higher, a smaller percentage of income goes to consumption and a greater percentage goes to saving. Consumption thus increases at a decreasing rate. If y is the consumption of an item and x_1 is disposable income (with x_2 standing for all the other independent variables), then the use of the semilog functional form is justified whenever the item's consumption can be expected to increase at a decreasing rate as income increases.

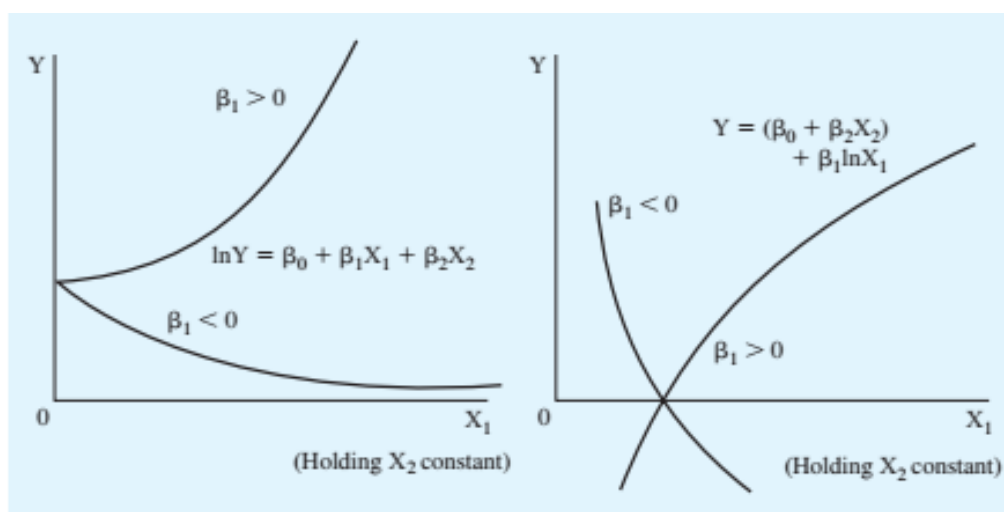


Figure 8.3 Semilog functions

The semilog functional form on the right ($\ln x$) can be used to depict a situation in which the impact of x_1 on y is expected to increase at a decreasing rate as x_1 gets bigger as long as β_1 is greater than zero (holding x_2 constant). The semilog functional form on the left ($\ln y$) can be used to depict a situation in which an increase in x_1 causes y to increase at an increasing rate.

As an example, let's consider the following annual model of the per capita consumption of beef in the Tajikistan:

$$\begin{aligned} \hat{CB}_t &= 37.54 - 0.88P_t + 11.9Yd_t \\ &\quad \quad \quad (0.16) \quad \quad \quad (1.76) \\ t &= -5.36 \quad \quad 6.75 \\ N &= 28 \quad \quad \bar{R}^2 = 0.631 \end{aligned} \tag{8.8}$$

where: CB = per capita consumption of beef

P = the price of beef in somoni

Yd = disposable income in somoni

If we substitute the log of disposable income ($\ln Yd_t$) for disposable income in equation 8.8, we get:

$$\begin{aligned} \hat{CB}_t &= -71.75 - \underset{(0.13)}{0.87} P_t + \underset{(11.11)}{98.87} Yd_t \\ t &= -6.93 \quad 8.90 \\ N &= 28 \quad \bar{R}^2 = 0.750 \end{aligned} \tag{8.9}$$

In equation (8.9), the independent variables include the price of beef and the *log* of disposable income. Equation (8.9) would be appropriate if we hypothesize that as income rises, consumption will increase at a decreasing rate. For other products, perhaps like yachts or summer homes, no such decreasing rate could be hypothesized, and the semilog function would not be appropriate.

Not all semilog functions have the log on the right-hand side of the equation, as in equation (8.7). The alternative semilog form is to have the log on the left-hand side of the equation. This would mean that the natural log of y would be a function of unlogged values of the x s, as in:

$$\ln y = a + \beta_1 x_1 + \beta_2 x_2 + u \tag{8.10}$$

This model has neither a constant slope nor a constant elasticity, but the coefficients do have a very useful interpretation. If x_1 increases by one *unit*, then y will change in *percentage* terms. Specifically, y will change by $\beta_1 \times 100\%$ for every unit that x_1 increases, holding x_2 constant. The left-hand side of Figure 8.3 shows such a semilog function.

This fact means that the $\ln y$ semilog function of equation (8.10) is perfect for any model in which the dependent variable adjusts in percentage terms to a unit change in an independent variable. The most common economic and business application of equation (8.10) is in a model of the earnings of individuals, where firms often give annual raises in percentage terms. In such a model, y would be the salary or wage of the i th employee, and x_1 would be the experience of the i th worker. Each year x_1 would increase by one, and β_1 would measure the percentage raises given by the firm.

Note that we now have two different kinds of semilog functional forms, creating possible confusion. As a result, many econometricians use phrases like “right-side semilog” or “lin-log

functional form” to refer to equation (8.7) while using “left-side semilog” or “log-lin functional form” to refer to equation (8.10).

Polynomial form

In most average cost functions, the slope of the cost curve changes sign as output changes. If the slopes of a relationship are expected to depend on the level of the variable itself, then a polynomial model should be considered. **Polynomial functional forms** express y as a function of independent variables, some of which are raised to powers other than 1. For example, in a second-degree polynomial (also called a quadratic) equation, at least one independent variable is squared:

$$y_i = a + \beta_1 x_{1i} + \beta_2 (x_{1i})^2 + \beta_3 x_{2i} + u_i \quad (8.11)$$

Such a model can indeed produce slopes that change sign as the independent variables change. The slope of y with respect to x_1 in equation (8.11) is:

$$\frac{\Delta y}{\Delta x_1} = \beta_1 + 2\beta_2 x_1 \quad (8.12)$$

Note that the slope depends on the level of x_1 . For small values of x_1 , β_1 might dominate, but for large values of x_1 , β_2 will always dominate. If this were a cost function, with y being the average cost of production and x_1 being the level of output of the firm, then we would expect β_1 to be negative and β_2 to be positive if the firm has the typical U-shaped cost curve depicted in the left half of Figure 8.4.

For another example, consider a model of annual employee earnings as a function of the age of each employee and a number of other measures of productivity such as education. What is the expected impact of age on earnings? As a young worker gets older, his or her earnings will typically increase.

Beyond some point, however, an increase in age will not increase earnings by very much at all, and around retirement we’d expect earnings to start to fall abruptly with age. As a result, a logical relationship between earnings and age might look something like the right half of Figure 8.4; earnings would rise, level off, and then fall as age increased. Such a theoretical relationship could be modeled with a quadratic equation:

$$\text{earnings}_i = a + \beta_1 \text{age}_i + \beta_2 \text{age}_i^2 + \dots + u_i \quad (8.13)$$

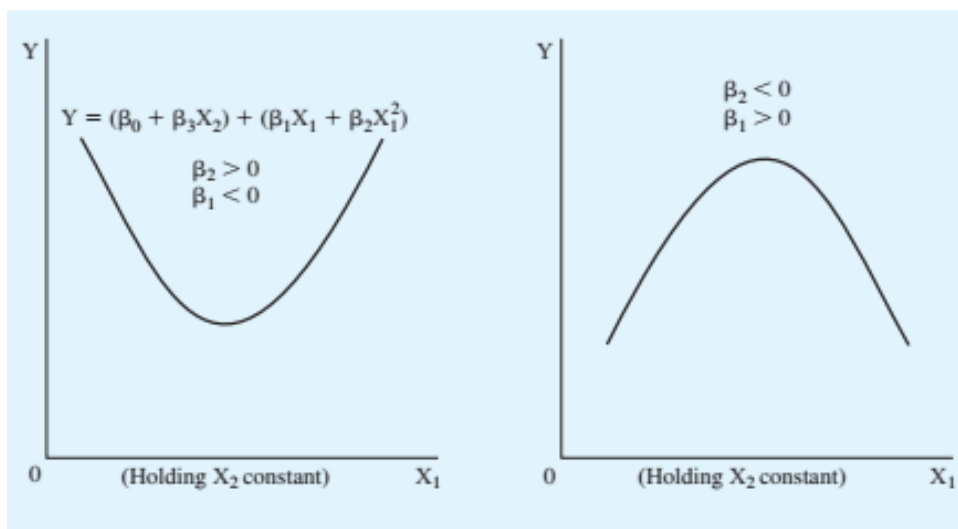


Figure 8.4 Polinomial functions

Quadratic functional forms (polynomials with squared terms) take on U or inverted U-shapes, depending on the values of the coefficients (holding x_2 constant). The left panel shows the shape of a quadratic function that could be used to show a typical cost curve; the right panel allows the description of an impact that rises and then falls (like the impact of age on earnings).

What would the expected signs of $\hat{\beta}_1$ and $\hat{\beta}_2$ be? Since you expect the impact of age to rise and fall, you'd thus expect $\hat{\beta}_1$ to be positive and $\hat{\beta}_2$ to be negative (all else being equal). In fact, this is exactly what many researchers in labor economics have observed.

With polynomial regressions, the interpretation of the individual regression coefficients becomes difficult, and the equation may produce unwanted results for particular ranges of x . Great care must be taken when using a polynomial regression equation to ensure that the functional form will achieve what is intended by the researcher and no more.

Choosing a functional form

The best way to choose a functional form for a regression model is to select the specification that best matches the underlying theory of the equation. In a majority of cases, the linear form will be adequate, and for most of the rest, common sense will point out a fairly easy choice from the alternatives outlined above. Table 8.1

contains a summary of the properties of the various alternative functional forms.

Table 8.1 Summary of alternative functional forms

Functional Form	Equation (one X only)	The Change in Y when X Changes
Linear	$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$	If X increases by one unit, Y will change by β_1 units.
Double-log	$\ln Y_i = \beta_0 + \beta_1 \ln X_i + \epsilon_i$	If X increases by one percent, Y will change by β_1 percent. (Thus β_1 is the elasticity of Y with respect to X.)
Semilog (lnX)	$Y_i = \beta_0 + \beta_1 \ln X_i + \epsilon_i$	If X increases by one percent, Y will change by $\beta_1/100$ units.
Semilog (lnY)	$\ln Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$	If X increases by one unit, Y will change by roughly $100\beta_1$ percent.
Polynomial	$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$	If X increases by one unit, Y will change by $(\beta_1 + 2\beta_2 X)$ units.

8.3 Lagged independent variables

Virtually all the regressions we've studied so far have been “instantaneous” in nature. In other words, they have included independent and dependent variables from the same time period, as in:

$$y_t = a + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t \quad (8.14)$$

where the subscript t is used to refer to a particular point in time. If all variables have the same subscript, then the equation is instantaneous.

However, not all economic or business situations imply such instantaneous relationships between the dependent and independent variables. In many cases time elapses between a change in the independent variable and the resulting change in the dependent variable. The period of time between the cause (the change in x) and the effect (the change in y) is called a **lag**. Time periods can be measured in days, months, years, etc. Many econometric equations include one or more lagged independent variables like $x1_{t-1}$, where the subscript $t - 1$ indicates that the observation of $x1$ is from the time period previous to time period t , as in the following equation:

$$y_t = a + \beta_1 x_{1t-1} + \beta_2 x_{2t} + u_t \quad (8.15)$$

In this equation, x_1 has been lagged by one time period, but the relationship between y and x_2 is still instantaneous. While this one-time-period lag is the most frequent lag in economics, lags of two or more time periods can be used when justified by the underlying theory.

For an example of a lagged independent variable, think about the process by which the supply of an agricultural product is determined. Since agricultural goods take time to grow, decisions on how many acres to plant or how many eggs to let hatch into egg-producing hens (instead of selling them immediately) must be made months, if not years, before the product is actually supplied to the consumer. Any change in an agricultural market, such as an increase in the price that the farmer can earn for providing cotton, has a lagged effect on the supply of that product:

$$C_t = a + \beta_1^+ PC_{t-1} + \beta_2^- PF_t + u_t \quad (8.16)$$

where C_t = the quantity of cotton supplied in year t

PC_{t-1} = the price of cotton in year $t - 1$

PF_t = the price of farm labor in year t

Note that this equation hypothesizes a lag between the price of cotton and the production of cotton, but not between the price of farm labor and the production of cotton. It's reasonable to think that if cotton prices change, farmers won't be able to react immediately because it takes a while for cotton to be planted and to grow.

The meaning of the regression coefficient of a lagged variable is not the same as the meaning of the coefficient of an unlagged variable. The estimated coefficient of a lagged x measures the change in *this year's* y attributed to a one-unit increase in *last year's* x (holding constant the other x s in the equation). Thus β_1 in equation (8.16) measures the extra number of units of cotton that

would be produced this year as a result of a one-unit increase in last year's price of cotton, holding this year's price of farm labor constant. If the lag structure is hypothesized to take place over more than one time period, or if a lagged dependent variable is included on the right-hand side of an equation, the question becomes significantly more complex. Such cases, called *distributed lags*, will be dealt with in Chapter 12.

8.4 Slope dummy variables

In section 4.3 we introduced the concept of a dummy variable, which we defined as one that takes on the values of 0 or 1, depending on a qualitative attribute such as gender. In that section our sole focus was on the use of a dummy variable as an **intercept dummy**, which changes the constant or intercept term, depending on whether the qualitative condition is met. These take the general form:

$$y_i = a + \beta_1 x_i + \beta_2 d_i + u_i \quad (8.17)$$

$$\text{where } d_i = \begin{cases} 1 & \text{if the } i\text{th teacher has a graduate degree} \\ 0 & \text{otherwise} \end{cases}$$

Until now, every independent variable in this text has been multiplied by exactly one other item: the slope coefficient. To see this, take another look at equation (8.17). As you can see, x is multiplied only by β_1 , and d is multiplied only by β_2 , and there are no other factors involved.

This restriction does not apply to a new kind of variable called an interaction term. An **interaction term** is an independent variable in a regression equation that is the *multiple* of two or more other independent variables. Each interaction term has its own regression coefficient, so the end result is that the interaction term has three or more components, as in $\beta_3 x_i d_i$. Such interaction terms are used when the change in y with respect to one independent variable (in

this case x) depends on the level of another independent variable (in this case d).

Interaction terms can involve two quantitative variables ($\beta_3 x_1 x_2$) or two dummy variables ($\beta_3 d_1 d_2$), but the most frequent application of interaction terms involves one quantitative variable and one dummy variable ($\beta_3 x_1 d_1$), a combination that is typically called a *slope dummy*. **Slope dummy variables** allow the slope of the relationship between the dependent variable and an independent variable to be different depending on whether the condition specified by a dummy variable is met. This is in contrast to an intercept dummy variable, which changes the intercept, but does not change the slope, when a particular condition is met.

In general, a slope dummy is introduced by adding to the equation a variable that is the multiple of the independent variable that has a slope you want to change and the dummy variable that you want to cause the changed slope. The general form of a slope dummy equation is:

$$y_i = a + \beta_1 x_i + \beta_2 d_i + \beta_3 x_i d_i + u_i \quad (8.18)$$

Note that equation (8.18) is the same as equation (8.17), except that we have added an interaction term in which the dummy variable is multiplied by an independent variable ($\beta_3 x_i d_i$). Let's check to make sure that the slope of y with respect to x does indeed change if d changes:

when $d = 0$, $\Delta y / \Delta x = \beta_1$

when $d = 1$, $\Delta y / \Delta x = (\beta_1 + \beta_3)$

In essence, the coefficient of x *changes* when the condition specified by d is met. To see this, substitute $d = 0$ and $d = 1$, respectively, into equation (8.17) and factor out x . Note that equation (8.17) includes both a slope dummy and an intercept dummy. It turns out that whenever a slope dummy is used, it's vital to also have $\beta_1 x_i$ and $\beta_2 d$ in the equation to avoid bias in the estimate of the coefficient of the slope dummy term. If there are other x s in an equation, they should not be multiplied by d unless you hypothesize that their slopes change with respect to d as well.

Take a look at Figure 8.5, which has both a slope dummy and an intercept dummy. In Figure 8.5 the intercept will be a when $d = 0$ and $a + \beta_2$ when $d = 1$. In addition, the slope of y with respect to x will be β_1 when $d = 0$ and $\beta_1 + \beta_3$ when $d = 1$. As a result, there really are two equations:

$$\begin{aligned} y_i &= a & \beta_1 x_i + u_i & \text{when } [d = 0] \\ y_i &= (a + \beta_2) + (\beta_1 + \beta_3)x_i + u_i & & \text{when } [d = 1] \end{aligned}$$

In practice, slope dummies have many realistic uses. For example, consider the question of earnings differentials between men and women. Although there is little argument that these differentials exist, there is quite a bit of controversy over the extent to which these differentials are caused by sexual discrimination (as opposed to other factors). Suppose you decide to build a model of earnings to get a better view of this controversy. If you hypothesized that men earn more than women on average, then you would want to use an intercept dummy variable for gender in an earnings equation that

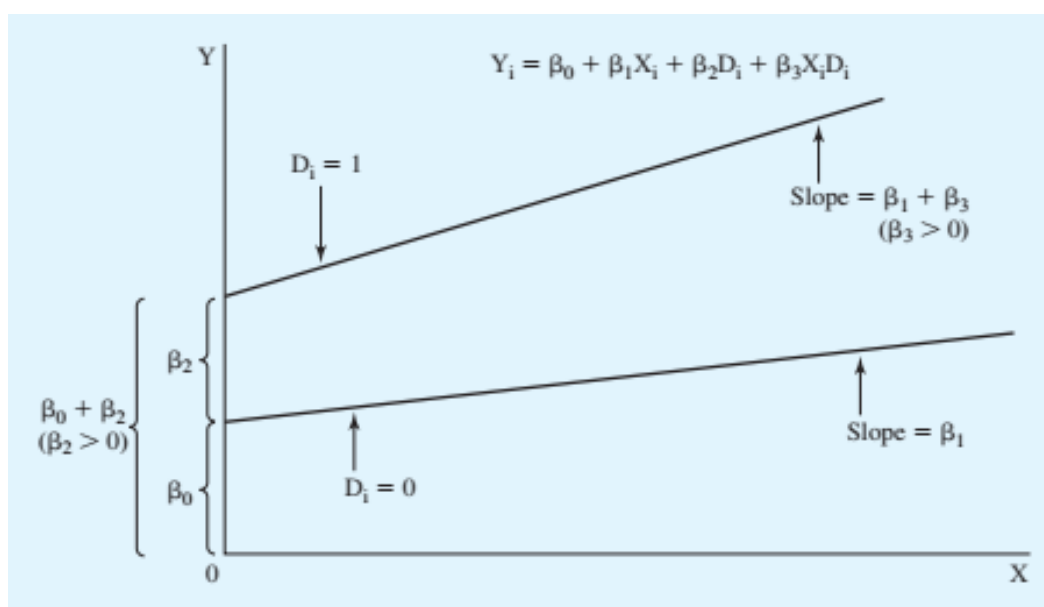


Figure 8.5 Slope and intercept dummies

If slope dummy ($\beta_3 x_i d_i$) and intercept dummy ($\beta_2 d_i$) terms are added to an equation, a graph of the equation will have different intercepts *and* different slopes depending on the value of the qualitative condition specified

SPECIFICATION: CHOOSING A FUNCTIONAL FORM

by the dummy variable. The difference between the two intercepts is β_2 , whereas the difference between the two slopes is β_3 .

included measures of experience, special skills, education, and so on, as independent variables:

$$\ln(\text{earnings}_i) = a + \beta_1 d_i + \beta_2 \exp_i + \dots + u_i \quad (8.19)$$

where $d_i = 1$ if the i th worker is male and 0 otherwise
 $\exp_i$ = the years experience of the i th worker
 e_i = a classical error term

In equation (8.19), $\hat{\beta}_1$ would be an estimate of the average difference in earnings between males and females, holding constant their experience and the other factors in the equation. Equation (8.19) also forces the impact of increases in experience (and the other factors in the equation) to have the same effect for females as for males because the slopes are the same for both genders.

If you hypothesized that men also increase their earnings more per year of experience than women, then you would include a slope dummy as well as an intercept dummy in such a model:

$$\ln(\text{earnings}_i) = a + \beta_1 d_i + \beta_2 \exp_i + \beta_3 d_i \exp_i + \dots + u_i \quad (8.20)$$

In equation (8.19), $\hat{\beta}_3$ would be an estimate of the differential impact of an extra year of experience on earnings between men and women. We could test the possibility of a positive true β_3 by running a one-tailed t -test on $\hat{\beta}_3$. If $\hat{\beta}_3$ were significantly different from zero in a positive direction, then we could reject the null hypothesis of no difference due to gender in the impact of experience on earnings, holding constant the other variables in the equation.

8.5 Exercises with incorrect functional forms

Once in a while a circumstance will arise in which the model is logically nonlinear in the variables, but the exact form of this nonlinearity is hard to specify. In such a case, the linear form is not correct, and yet a choice between the various nonlinear forms cannot be made on the basis of economic theory.

Even in these cases, however, it still pays (in terms of understanding the true relationships) to avoid choosing a functional form on the basis of fit alone.

If functional forms are similar, and if theory does not specify exactly which form to use, why should we try to avoid using goodness of fit over the sample to determine which equation to use? This section will highlight two answers to this question:

1. $\bar{R}^2$ s are difficult to compare if the dependent variable is transformed.
2. An incorrect functional form may provide a reasonable fit within the sample but have the potential to make large forecast errors when used outside the range of the sample.

$\bar{R}^2$ s are difficult to compare when y is transformed

When the dependent variable is transformed from its linear version, the overall measure of fit, the $\bar{R}^2$, cannot be used for comparing the fit of the nonlinear equation with the original linear one.⁴² This problem is not especially important in most cases because the emphasis in applied regression analysis is usually on the coefficient estimates. However, if $\bar{R}^2$ s are ever used to compare the fit of two different functional forms, then it becomes crucial that this lack of comparability be remembered. For example, suppose you were trying to compare a linear equation:

$$y = a + \beta_1 x_1 + \beta_2 x_2 + u \quad (8.21)$$

⁴² This warning also applies to other measures of overall fit, for example Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) of the appendix of the chapter 7.

with a semilog version of the same equation (using the version of a semilog function that takes the log of the dependent variable):

$$\ln y = a + \beta_1 x_1 + \beta_2 x_2 + u \quad (8.22)$$

Notice that the only difference between equations (8.21) and (8.22) is the functional form of the dependent variable. The reason that the $\bar{R}^2$ s of the respective equations cannot be used to compare overall fits of the two equations is that the total sum of squares (TSS) of the dependent variable around its mean is different in the two formulations. That is, the $\bar{R}^2$ s are not comparable because the dependent variables are different. There is no reason to expect that different dependent variables will have the identical (or easily comparable) degrees of dispersion around their means.

Incorrect functional forms outside the range of the sample

If an incorrect functional form is used, then the probability of mistaken inferences about the true population parameters will increase. Using an incorrect functional form is a kind of specification error that is similar to the omitted variable bias discussed in section 7.1. Even if an incorrect functional form provides good statistics within a sample, large residuals almost surely will arise when the misspecified equation is used on data that were not part of the sample used to estimate the coefficients.

In general, the extrapolation of a regression equation to data that are outside the range over which the equation was estimated runs increased risks of large forecasting errors and incorrect conclusions about population values. This risk is heightened if the regression uses a functional form that is inappropriate for the particular variables being studied.

Two functional forms that behave similarly over the range of the sample may behave quite differently outside that range. If the functional form is chosen on the basis of theory, then the researcher can take into account how the equation would act over any range of values, even if some of those values are outside the range of the

sample. If functional forms are chosen on the basis of fit, then extrapolating outside the sample becomes tenuous.

Figure 8.6 contains a number of hypothetical examples. As can be seen, some functional forms have the potential to fit quite poorly outside the sample range.

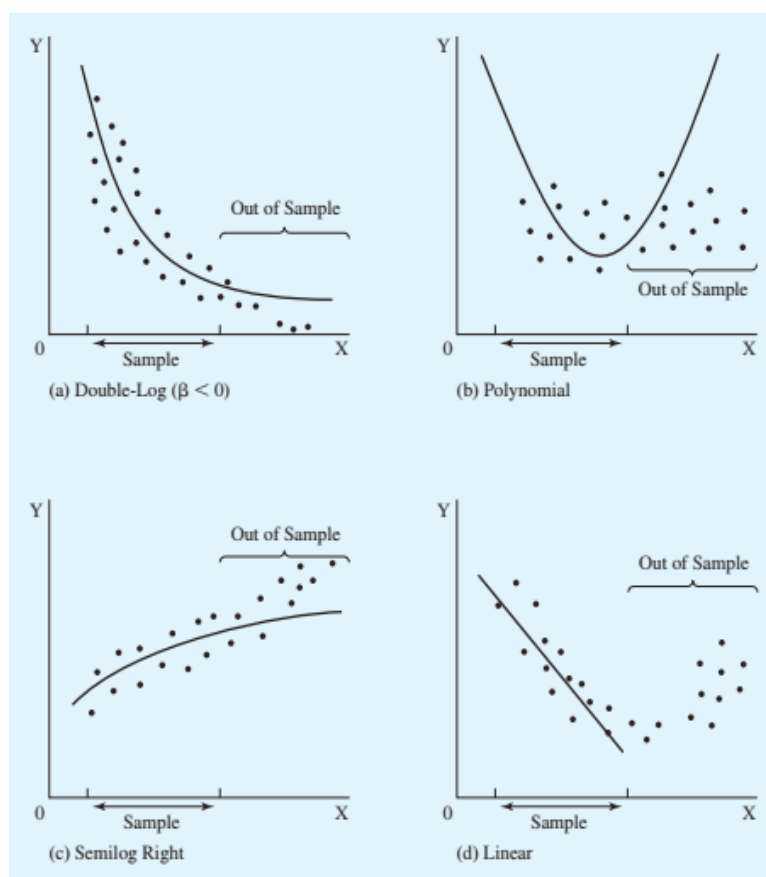


Figure 8.6 Incorrect functional forms outside the sample range

If an incorrect form is applied to data outside the range of the sample on which it was estimated, the probability of large mistakes increases. In particular, note how the polynomial functional form can change slope rapidly outside the sample range (panel b) and that even a linear form can cause mistakes if the true functional form is nonlinear (panel d).

Such graphs are meant as examples of what could happen, not as statements of what necessarily will happen, when incorrect functional forms are pushed outside the range of the sample over which they were estimated. Do not conclude from these diagrams that nonlinear functions should be avoided. If the true relationship is nonlinear, then the *linear* functional form will make large

forecasting errors outside the sample. Instead, the researcher must take the time to think through how the equation will act for values both inside and outside the sample before choosing a functional form to use to estimate the equation. If the theoretically appropriate nonlinear equation appears to work well over the relevant range of possible values, then it should be used without concern over this issue.

SUMMARY

1. Do not suppress the constant term. On the other hand, don't rely on estimates of the constant term for inference even if it appears to be statistically significant.
2. The choice of a functional form should be based on the underlying economic theory to the extent that theory suggests a shape similar to that provided by a particular functional form. A form that is linear in the variables should be used unless a specific hypothesis suggests otherwise.
3. Functional forms that are nonlinear in the variables include the double-log form, the semilog form, and the polynomial form. The double-log form is especially useful if the elasticities involved are expected to be constant. The semilog form has the advantage of allowing the effect of an independent variable to tail off as that variable increases. The polynomial form is useful if the slopes are expected to change sign, depending on the level of an independent variable.
4. A slope dummy is a dummy variable that is multiplied by an independent variable to allow the slope of the relationship between the dependent variable and the particular independent variable to change, depending on whether a particular condition is met.
5. The use of nonlinear functional forms has a number of potential exercises. In particular, the $\bar{R}^2$ s are difficult to compare if y has been transformed, and the residuals are potentially large if an incorrect functional form is used for forecasting outside the range of the sample.

CHAPTER 9. MULTICOLLINEARITY

The next three chapters deal with violations of the classical assumptions and remedies for those violations. This chapter addresses multicollinearity, and the next two chapters address serial correlation and heteroskedasticity. For each of these three exercises, we will attempt to answer the following questions:

1. What is the nature of the problem?
2. What are the consequences of the problem?
3. How is the problem diagnosed?
4. What remedies for the problem are available?

Strictly speaking, *perfect multicollinearity* is the violation of Classical assumption 6 - that no independent variable is a perfect linear function of one or more other independent variables. Perfect multicollinearity is rare, but severe imperfect multicollinearity, although not violating Classical assumption 6, still causes substantial exercises. Recall that the coefficient β_k can be thought of as the impact on the dependent variable of a one - unit increase in the independent variable x_k , holding constant the other independent variables in the equation. If two explanatory variables are significantly related, then the OLS computer program will find it difficult to distinguish the effects of one variable from the effects of the other.

In essence, the more highly correlated two (or more) independent variables are, the more difficult it becomes to accurately estimate the coefficients of the true model. If two variables move identically, then there is no hope of distinguishing between their impacts, but if the variables are only roughly correlated, then we still might be able to estimate the two effects accurately enough for most purposes.

9.1 Perfect versus imperfect multicollinearity

Perfect multicollinearity

Perfect multicollinearity⁴³ violates Classical assumption 6, which specifies that no explanatory variable is a perfect linear function of any other explanatory variable. The word *perfect* in this context implies that the variation in one explanatory variable can be *completely* explained by movements in another explanatory variable. Such a perfect linear function between two independent variables would be:

$$x_{1i} = \alpha_0 + \alpha_1 x_{2i} \quad (9.1)$$

where the α s are constants and the x s are independent variables in:

$$y_i = a + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \quad (9.2)$$

Notice that there is no error term in equation (9.1). This implies that x_1 can be exactly calculated given x_2 and the equation. Typical equations for such perfect linear relationships would be:

$$x_{1i} = 3x_{2i} \quad (9.3)$$

$$x_{1i} = 2 + 4x_{2i} \quad (9.4)$$

What are some real-world examples of perfect multicollinearity? The simplest examples involve the same variable measured in different units. Think about the distance between two cities measured in miles with x_1 and in kilometers with x_2 . The data for the variables look quite different, but they're perfectly correlated! A more subtle example is when the two variables always add up to the same amount, for instance P_1 , the percent of voters who voted in favor of a proposition, and P_2 , the percent who voted

⁴³ The word *collinearity* describes a linear correlation between two independent variables, and *multicollinearity* indicates that more than two independent variables are involved. In common usage, multicollinearity is used to apply to both cases, and so we'll typically use that term in this text even though many of the examples and techniques discussed relate, strictly speaking, to collinearity.

against it (assuming no abstentions), would always add up to 100% and therefore would be perfectly (negatively) correlated. Figure 9.1 shows a graph of explanatory variables that are perfectly correlated. As can be seen in Figure 9.1, a perfect linear function has all data points on the same straight line. There is none of the variation that accompanies the data from a typical regression.

What happens to the estimation of an econometric equation where there is perfect multicollinearity? OLS is incapable of generating estimates of the regression coefficients, and most OLS computer programs will print out an error message in such a situation. Using equation (9.2) as an example, we theoretically would obtain the following estimated coefficients and standard errors:

$$\hat{\beta}_1 = \text{indeterminate} \quad SE(\hat{\beta}_1) = \infty \quad (9.5)$$

$$\hat{\beta}_2 = \text{indeterminate} \quad SE(\hat{\beta}_2) = \infty \quad (9.6)$$

Perfect multicollinearity ruins our ability to estimate the coefficients because the two variables cannot be distinguished. You cannot “hold all the other independent variables in the equation constant” if every time one variable changes, another changes in an identical manner.

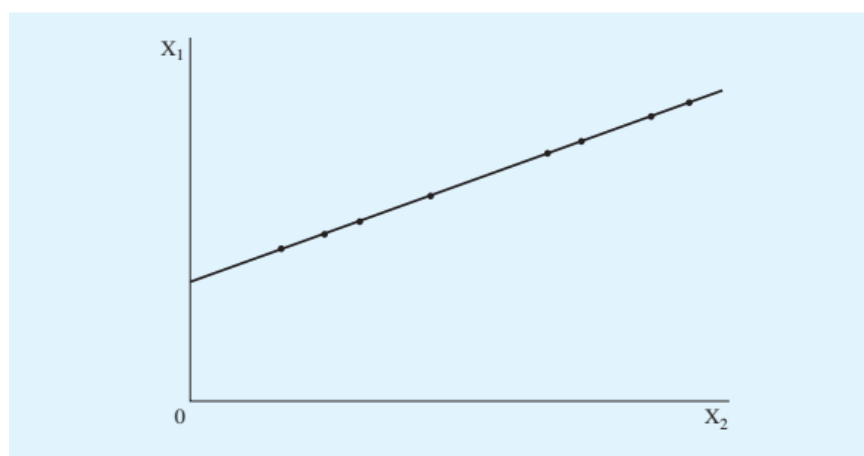


Figure 9.1 Perfect multicollinearity

With perfect multicollinearity, an independent variable can be completely explained by the movements of one or more other independent variables. Perfect multicollinearity can usually be avoided by careful screening of the independent variables before a regression is run.

Fortunately, instances in which one explanatory variable is a perfect linear function of another are rare. More important, perfect multicollinearity should be fairly easy to discover before a regression is run. You can detect perfect multicollinearity by asking whether one variable equals a multiple of another or if one variable can be derived by adding a constant to another or if a variable equals the sum of two other variables. If so, then one of the variables should be dropped because there is no essential difference between the two.

A special case related to perfect multicollinearity occurs when a variable that is definitionally related to the dependent variable is included as an independent variable in a regression equation. Such a **dominant variable** is by definition so highly correlated with the dependent variable that it completely masks the effects of all other independent variables in the equation. In a sense, this is a case of perfect collinearity between the dependent variable and an independent variable.

For example, if you include a variable measuring the amount of raw materials used by the shoe industry in a production function for that industry, the raw materials variable would have an extremely high t -score, but otherwise important variables like labor and capital would have quite insignificant t -scores. Why? In essence, if you knew how much leather was used by a shoe factory, you could predict the number of pairs of shoes produced without knowing *anything* about labor or capital. The relationship is definitional, and the dominant variable should be dropped from the equation to get reasonable estimates of the coefficients of the other variables.

Be careful, though! Dominant variables shouldn't be confused with highly significant or important explanatory variables. Instead, they should be recognized as being virtually identical to the dependent variable. While the fit between the two is superb, knowledge of that fit could have been obtained from the definitions of the variables without any econometric estimation.

Imperfect multicollinearity

Since perfect multicollinearity is fairly easy to avoid, econometricians rarely talk about it. Instead, when we use the word multicollinearity, we really are talking about severe imperfect multicollinearity. **Imperfect multicollinearity** can be defined as a

linear functional relationship between two or more independent variables that is so strong that it can significantly affect the estimation of the coefficients of the variables. In other words, imperfect multicollinearity occurs when two (or more) explanatory variables are imperfectly linearly related, as in:

$$x_{1i} = \alpha_0 + \alpha_1 x_{2i} + u_i \quad (9.7)$$

Compare equation (9.7) to equation ((9.1); notice that equation (9.7) includes u_i , a stochastic error term. This implies that although the relationship between x_1 and x_2 might be fairly strong, it is not strong enough to allow x_1 to be completely explained by x_2 ; some unexplained variation still remains. Figure (9.2) shows the graph of two explanatory variables that might be considered imperfectly multicollinear. Notice that although all the observations in the sample are fairly close to the straight line, there is still some variation in x_1 that cannot be explained by x_2 .

Imperfect multicollinearity is a strong linear relationship between the explanatory variables. The stronger the relationship is between the two (or more) explanatory variables, the more likely it is that they'll be considered significantly multicollinear. Two variables that might be only slightly related in one sample might be so strongly related in another that they could be considered to be imperfectly

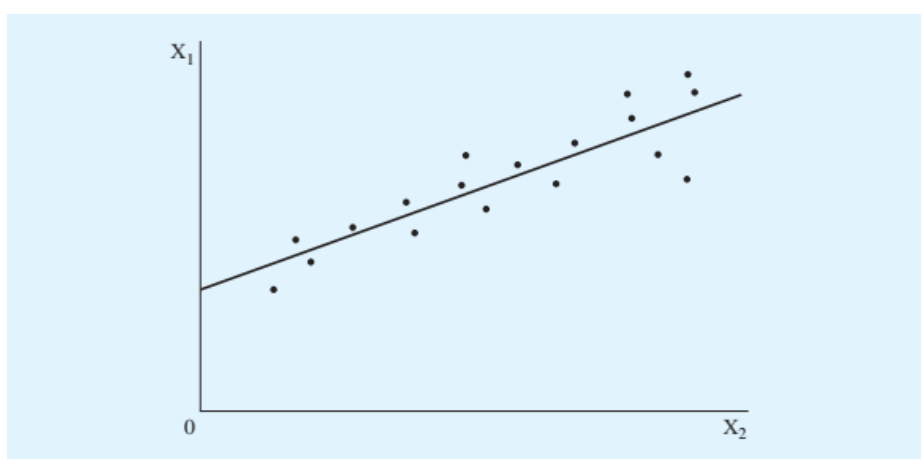


Figure 9.2 Imperfect multicollinearity

With imperfect multicollinearity, an independent variable is a strong but not perfect linear function of one or more other independent variables. Imperfect multicollinearity varies in degree from sample to sample.

multicollinear. In this sense, it is fair to say that multicollinearity is a sample phenomenon as well as a theoretical one. This contrasts with perfect multicollinearity because two variables that are perfectly related probably can be detected on a logical basis. The detection of multicollinearity will be discussed in more detail in section 9.3.

9.2 The consequences of multicollinearity

If the multicollinearity in a particular sample is severe, what will happen to estimates calculated from that sample? The purpose of this section is to explain the consequences of multicollinearity and then to explore some examples of such consequences.

Recall the properties of OLS estimators that might be affected by this or some other econometric problem. In chapter 5, we stated that the OLS estimators are BLUE (or MvLUE) if the classical assumptions hold. This means that OLS estimates can be thought of as being unbiased and having the minimum variance possible for unbiased linear estimators.

What are the consequences of multicollinearity?

The major consequences of multicollinearity are:

1. *Estimates will remain unbiased.* Even if an equation has significant multicollinearity, the estimates of the β s still will be centered around the true population β s if the first six Classical Assumptions are met for a correctly specified equation.
2. *The variances and standard errors of the estimates will increase.* This is the principal consequence of multicollinearity. Since two or more of the explanatory variables are significantly related, it becomes difficult to precisely identify the separate effects of the multicollinear variables. When it becomes hard to distinguish the effect of one variable from the effect of another, we're much more likely to make large errors in estimating the β s than we were before we encountered multicollinearity. As a result, the estimated coefficients, although still unbiased, now come

from distributions with much larger variances and, therefore, larger standard errors.

Even though the variances and standard errors are larger with multicollinearity than they are without it, OLS is still BLUE when multicollinearity exists. That is, no other linear unbiased estimation technique can get lower variances than OLS even in the presence of multicollinearity. Thus, although the effect of multicollinearity is to increase the variance of the estimated coefficients, OLS still has the property of minimum variance. These “minimum variances” are just fairly large.

Figure 8.3 compares a distribution of $\hat{\beta}$ s from a sample with severe multicollinearity to one with virtually no correlation between any of the independent variables. Notice that the two distributions have the same mean, indicating that multicollinearity does not cause bias. Also note how much wider the distribution of $\hat{\beta}$ N becomes when multicollinearity is severe; this is the result of the increase in the standard error of $\hat{\beta}$ that is caused by multicollinearity. Because of this larger variance, multicollinearity increases the likelihood of obtaining an unexpected sign⁴⁴ for a coefficient even though, as mentioned earlier, multicollinearity causes no bias.

3. *The computed t-scores will fall.* Multicollinearity tends to decrease the *t*-scores of the estimated coefficients mainly because of the formula for the *t*-statistic:

⁴⁴ These unexpected signs generally occur because the distribution of the $\hat{\beta}$ s with multicollinearity is wider than it would be without it, increasing the chance that a particular observed $\hat{\beta}$ will be on the other side of zero from the true β (have an unexpected sign).

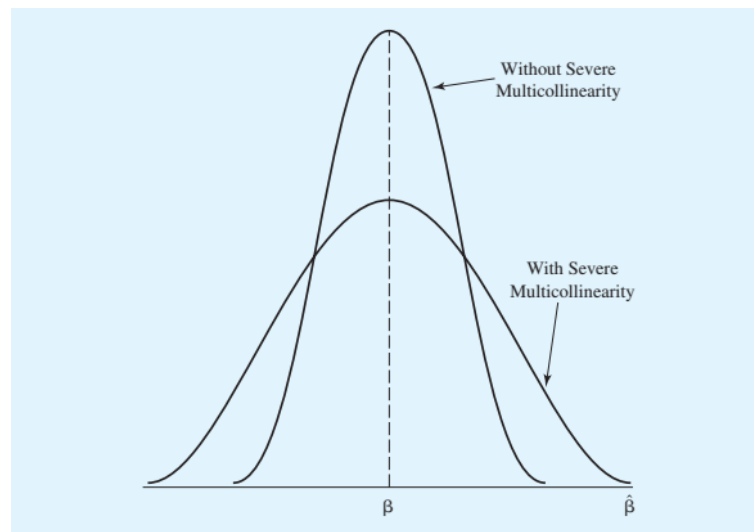


Figure 9.3 Severe multicollinearity increases the variances of the $\hat{\beta}$ s

Severe multicollinearity produces a distribution of the $\hat{\beta}$ s that is centered around the true β but that has a much wider variance. Thus, the distribution of $\hat{\beta}$ s with multicollinearity is much wider than otherwise.

$$t_k = \frac{(\hat{\beta}_k - \beta_{H_0})}{SE(\hat{\beta}_k)} \quad (9.8)$$

Notice that this equation is divided by the standard error of the estimated coefficient. Multicollinearity increases the standard error of the estimated coefficient, and if the standard error increases, then the t -score must fall, as can be seen from equation (9.8). Not surprisingly, it's quite common to observe low t -scores in equations with severe multicollinearity.

Similarly, the computed confidence intervals will widen. Because multicollinearity increases the standard error of the estimated coefficient, it makes the confidence interval wider (see equation 7.9). Put differently, since $\hat{\beta}$ is likely to be farther from the true β , the confidence interval is forced to increase.

4. *Estimates will become very sensitive to changes in specification.* The addition or deletion of an explanatory variable or of a few observations will often cause major

changes in the values of the $\hat{\beta}$ s when significant multicollinearity exists. If you drop a variable, even one that appears to be statistically insignificant, the coefficients of the remaining variables in the equation sometimes will change dramatically.

These large changes occur because OLS estimation is sometimes forced to emphasize small differences between variables in order to distinguish the effect of one multicollinear variable from that of another. If two variables are virtually identical throughout most of the sample, the estimation procedure relies on the observations in which the variables move differently in order to distinguish between them. As a result, a specification change that drops a variable that has an unusual value for one of these crucial observations can cause the estimated coefficients of the multicollinear variables to change dramatically.

5. *The overall fit of the equation and the estimation of the coefficients of nonmulticollinear variables will be largely unaffected.* Even though the individual t -scores are often quite low in a multicollinear equation, the overall fit of the equation, as measured by $\bar{R}^2$, will not fall much, if at all, in the face of significant multicollinearity. Given this, one of the first indications of severe multicollinearity is the combination of a high $\bar{R}^2$ with no statistically significant individual regression coefficients. Similarly, if an explanatory variable in an equation is not multicollinear with the other variables, then the estimation of its coefficient and standard error usually will not be affected.

Because the overall fit is largely unchanged, it's possible for the F -test of overall significance to reject the null hypothesis even though none of the t -tests on individual coefficients can do so. Such a result is a clear indication of severe imperfect multicollinearity.

Finally, since multicollinearity has little effect on the overall fit of the equation, it also will have little effect on the use of that equation for prediction or forecasting, as long as the independent variables maintain the same pattern of

multicollinearity in the forecast period that they demonstrated in the sample.

Two Examples of the Consequences of Multicollinearity

To see what severe multicollinearity does to an estimated equation, let's look at a hypothetical example. Suppose you decide to estimate a "student consumption function." After the appropriate preliminary work, you come up with the following hypothesized equation:

$$CO_i = a + \beta_1^+ Yd_i + \beta_2^- LA_i + u_i \quad (9.9)$$

where: CO_i = the annual consumption expenditures of the i th student on items other than tuition and room and board

Yd_i = the annual disposable income (including gifts) of that student

LA_i = the liquid assets (savings, etc.) of the i th student

u_i = a stochastic error term

You then collect a small amount of data from people who are sitting near you in class:

Student	CO_i	Yd_i	LA_i
Mary	\$2000	\$2500	\$25000
Robby	2300	3000	31000
Bevin	2800	3500	33000
Lesley	3800	4000	39000
Brandon	3500	4500	48000
Bruce	5000	5000	54000
Harwood	4500	5500	55000

If you run an OLS regression on your data set for equation (9.9), you obtain:

$$\begin{aligned} \hat{CO}_i &= -367.83 + 0.5113Yd_i + 0.0427LA_i \\ &\quad \quad \quad (1.0307) \quad \quad \quad (0.0942) \\ &\quad \quad \quad t = 0.496 \quad \quad 0.453 \\ \bar{R}^2 &= 0.835 \end{aligned} \quad (9.10)$$

On the other hand, if you had consumption as a function of disposable income alone, then you would have obtained:

$$\begin{aligned} C\hat{O}_i &= -471.43 + 0.9714Yd_i \\ &\quad \quad \quad (0.157) \\ t &= 6.187 \\ \bar{R}^2 &= 0.861 \end{aligned} \tag{9.11}$$

Notice from equations (9.10) and (9.11) that the t -score for disposable income increases more than tenfold when the liquid assets variable is dropped from the equation. Why does this happen? First of all, the correlation between Yd and LA is quite high. This high degree of correlation causes the standard errors of the estimated coefficients to be very high when both variables are included. In the case of $\hat{\beta}_{Yd}$, the standard error goes from 0.157 to 1.03 with the inclusion of LA ! In addition, the coefficient estimate itself changes somewhat. Further, note that the $\bar{R}^2$ s of the two equations are quite similar despite the large differences in the significance of the explanatory variables in the two equations. It's quite common for $\bar{R}^2$ to stay virtually unchanged when multicollinear variables are dropped. All of these results are typical of equations with multicollinearity.

Which equation is better? If the liquid assets variable theoretically belongs in the equation, then to drop it will run the risk of omitted variable bias, but to include the variable will mean certain multicollinearity. There is no automatic answer when dealing with multicollinearity. We'll discuss this issue in more detail in sections 9.4 and 9.5.

A second example of the consequences of multicollinearity is based on actual rather than hypothetical data. Suppose you've decided to build a cross-sectional model of the demand for gasoline by state:

$$PCON_i = a + \beta_1^+ UHM_i + \beta_2^- TAX_i + \beta_3^+ REG_i + u_i \tag{9.12}$$

where $PCON_i$ = petroleum consumption in the i th state (trillions of BTUs)

UHM_i = urban highway miles within the i th state

TAX_i = the gasoline tax rate in the i th state (cents per gallon)

REG_i = motor vehicle registrations in the i th state (thousands)

Given the definitions, let's move on to the estimation of equation (9.12) using a linear functional form (assuming a classical error term):

$$\begin{aligned}
 PC\hat{O}N_i &= 389.6 + 60.8UHM_i - 36.5TAX_i - 0.061REG_i \\
 &\quad \quad \quad (10.3) \quad \quad \quad (13.2) \quad \quad \quad (0.043) \\
 &\quad \quad \quad t = 5.92 \quad \quad \quad -2.77 \quad \quad \quad -1.43 \\
 N &= 50 \quad \quad \quad \bar{R}^2 = 0.919
 \end{aligned} \tag{9.13}$$

What's wrong with this equation? The motor vehicle registrations variable has an insignificant coefficient with an unexpected sign, but it's hard to believe that the variable is irrelevant. Is an omitted variable causing bias? It's possible, but adding a variable is unlikely to fix things. Does it help to know that the correlation between REG and UHM is extremely high? Given that, it seems fair to say that one of the two variables is redundant; both variables are really measuring the *size* of the state, so we have multicollinearity.

Notice the impact of the multicollinearity on the equation. The coefficient of a variable such as motor vehicle registrations, which has a very strong theoretical relationship to petroleum consumption, is insignificant and has a sign contrary to our expectations. This is mainly because the multicollinearity has increased the variance of the distribution of the estimated $\hat{\beta}$ s.

What would happen if we were to drop one of the multicollinear variables?

$$\begin{aligned}
 PC\hat{O}N_i &= 551.7 - 53.6TAX_i + 0.186REG_i \\
 &\quad \quad \quad (16.9) \quad \quad \quad (0.012) \\
 &\quad \quad \quad t = -3.18 \quad \quad \quad 15.88 \\
 N &= 50 \quad \quad \quad \bar{R}^2 = 0.861
 \end{aligned} \tag{9.14}$$

Dropping UHM has made REG extremely significant. Why did this occur? The answer is that the standard error of the coefficient of

REG has fallen substantially (from 0.043 to 0.012) now that the multicollinearity has been removed from the equation. Also note that the sign of the estimated coefficient has now become positive as hypothesized. The reason is that *REG* and *UHM* are virtually indistinguishable from an empirical point of view, and so the OLS program latched on to minor differences between the variables to explain the movements of *PCON*. Once the multicollinearity was removed, the direct positive relationship between *REG* and *PCON* was obvious. Either *UHM* or *REG* could have been dropped with similar results because the two variables are, in a quantitative sense, virtually identical. In fact, our guess is that a majority of experienced econometricians, when faced with the results in equation (9.13) and the high correlation between *REG* and *UHM*, would have dropped *REG* and kept *UHM*. Why did we do the opposite? Our opinion is that because *UHM* is an urban variable and *REG* is a statewide variable, *REG* is preferable from a theoretical standpoint if we're trying to understand statewide petroleum consumption. Since the two are identical quantitatively and *REG* is preferable theoretically, we'd keep *REG*, but we recognize that others could look at the same results and come to a different conclusion. Even though $\bar{R}^2$ fell when *UHM* was dropped, equation (9.14) should be considered superior to equation (9.13). This is an example of the point, first made in chapter 3, that the fit of the equation is not the most important criterion to be used in determining its overall quality.

9.3 The Detection of multicollinearity

How do we decide whether an equation has a severe multicollinearity problem? A first step is to recognize that some multicollinearity exists in every equation. It's virtually impossible in a real-world example to find a set of explanatory variables that are totally uncorrelated with each other (except for designed experiments). Our main purpose in this section will be to learn to determine *how much* multicollinearity exists in an equation, not *whether* any multicollinearity exists.

A second key point is that the severity of multicollinearity in a given equation can change from sample to sample depending on the characteristics of the sample. As a result, the theoretical underpinnings of the equation are not quite as important in the detection of multicollinearity as they are in the detection of an omitted variable or an incorrect functional form. Instead, we tend to rely more on data-oriented techniques to determine the severity of the multicollinearity in a given sample. Of course, we can never ignore the theory behind an equation. The trick is to find variables that are theoretically relevant (for meaningful interpretation) and that are also statistically nonmulticollinear (for meaningful inference).

Because multicollinearity is a sample phenomenon and the level of damage of its impact is a matter of degree, many of the methods used to detect it are informal tests without critical values or levels of significance. Indeed, there are no generally accepted, true statistical tests for multicollinearity. Most researchers develop a general feeling for the severity of multicollinearity in an estimated equation by looking at a number of the characteristics of that equation. Let's examine two of the most frequently used characteristics.

High simple correlation coefficients

One way to detect severe multicollinearity is to examine the simple correlation coefficients between the explanatory variables. The **simple correlation coefficient**, r , is a measure of the strength and direction of the linear relationship between two variables.⁴⁵ The range of r is from +1 to -1, and the sign of r indicates the direction of the correlation between the two variables. The closer the absolute

⁴⁵ The equation for r_{12} , the simple correlation coefficient between x_1 and x_2 , is:

$$r_{12} = \frac{\sum [(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)]}{\sqrt{\sum (x_{1i} - \bar{x}_1)^2 \sum (x_{2i} - \bar{x}_2)^2}}$$

Interestingly, it turns out that r and R^2 are related if the estimated equation has exactly one independent variable. The square of r equals R^2 for a regression where one of the two variables is the dependent variable and the other is the only independent variable.

value of r is to 1, the stronger is the correlation between the two variables. Thus:

If two variables are perfectly positively correlated, then $r = +1$

If two variables are perfectly negatively correlated, then $r = -1$

If two variables are totally uncorrelated, then $r = 0$

If an r is high in absolute value, then we know that these two particular x s are quite correlated and that multicollinearity is a potential problem. For example, in equation (9.10), the simple correlation coefficient between disposable income and liquid assets is 0.986. A simple correlation coefficient this high, especially in an equation with only two independent variables, is a certain indication of severe multicollinearity.

How high is high? Some researchers pick an arbitrary number, such as 0.80, and become concerned about multicollinearity any time the absolute value of a simple correlation coefficient exceeds 0.80. A better answer might be that r is high if it causes unacceptably large variances in the coefficient estimates in which we're interested.

Be careful: The use of simple correlation coefficients as an indication of the extent of multicollinearity involves a major limitation if there are more than two explanatory variables. It is quite possible for groups of independent variables, acting together, to cause multicollinearity without any single simple correlation coefficient being high enough to indicate that multicollinearity is in fact severe. As a result, simple correlation coefficients must be considered to be sufficient but not necessary tests for multicollinearity. Although a high r does indeed indicate the probability of severe multicollinearity, a low r by no means proves otherwise.⁴⁶

⁴⁶ Most authors criticize the use of simple correlation coefficients to detect multicollinearity in equations with large numbers of explanatory variables, but many researchers continue to do so because a scan of the simple correlation coefficients is a "quick and dirty" way to get a feel for the degree of multicollinearity in an equation.

High variance inflation factors (VIFs)

The use of tests to give an indication of the severity of multicollinearity in a particular sample is controversial. Some econometricians reject even the simple correlation coefficient, mainly because of the limitations cited. Others tend to use a variety of more formal tests.

One measure of the severity of multicollinearity that is easy to use and that is gaining in popularity is the variance inflation factor. The **variance inflation factor (VIF)** is a method of detecting the severity of multicollinearity by looking at the extent to which a given explanatory variable can be explained by all the other explanatory variables in the equation. There is a VIF for each explanatory variable in an equation. The VIF is an index of how much multicollinearity has increased the variance of an estimated coefficient. A high VIF indicates that multicollinearity has increased the estimated variance of the estimated coefficient by quite a bit, yielding a decreased *t*-score. Suppose you want to use the VIF to attempt to detect multicollinearity in an original equation with *K* independent variables:

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$$

Doing so requires calculating *K* different VIFs, one for each x_i . Calculating the VIF for a given x_i involves two steps:

1. *Run an OLS regression that has x_i as a function of all the other explanatory variables in the equation. For $i = 1$, this equation would be:*

$$x_1 = \alpha_1 + \alpha_2 x_2 + \alpha_3 x_3 + \dots + \alpha_k x_k + e \quad (9.15)$$

where *e* is a classical stochastic error term. Note that x_1 is not included on the right-hand side of equation (9.15), which is referred to as an auxiliary or secondary regression. Thus there are *K* auxiliary regressions, one for each independent variable in the original equation.

2. Calculate the variance inflation factor for $\hat{\beta}_i$:

$$VIF(\hat{\beta}_i) = \frac{1}{(1 - R_i^2)} \quad (9.16)$$

where R_i^2 is the coefficient of determination (the unadjusted R^2) of the auxiliary regression in step one. Since there is a separate auxiliary regression for each independent variable in the original equation, there also is an R_i^2 and a $VIF(\hat{\beta}_i)$ for each x_i . The higher the VIF, the more severe the effects of multicollinearity.

How high is high? An R_i^2 of 1, indicating perfect multicollinearity, produces a VIF of infinity, whereas an R_i^2 of 0, indicating no multicollinearity at all, produces a VIF of 1. While there is no Table of formal critical VIF values, a common rule of thumb is that if $VIF(\beta_i) > 5$, the multicollinearity is severe. As the number of independent variables increases, it makes sense to increase this number slightly.

For example, let's return to equation (9.10) and calculate the VIFs for both independent variables. Both VIFs equal 36, confirming the quite severe multicollinearity we already know exists. It's no coincidence that the VIFs for the two variables are equal. In an equation with exactly two independent variables, the two auxiliary equations will have identical R_i^2 s, leading to equal VIFs.⁴⁷

Some authors and statistical software programs replace the VIF with its reciprocal, $(1 - R_i^2)$, called *tolerance*, or TOL. Whether we calculate VIF or TOL is a matter of personal preference, but either way, the general approach is the most comprehensive multicollinearity detection technique we've discussed in this text.

Unfortunately, there are a couple of exercises with using VIFs. First, as mentioned, there is no hard-and-fast VIF decision rule.

⁴⁷ Another use for the R^2 s of these auxiliary equations is to compare them with the overall equation's R^2 . If an auxiliary equation's R^2 is higher, it's yet another sign of multicollinearity.

Second, it's possible to have multicollinear effects in an equation that has no large VIFs. For instance, if the simple correlation coefficient between x_1 and x_2 is 0.88, multicollinear effects are quite likely, and yet the VIF for the equation (assuming no other x s) is only 4.4.

In essence, then, the VIF is a sufficient but not necessary test for multicollinearity, just like the other test described in this section. Indeed, as is probably obvious to the reader by now, there is no test that allows a researcher to reject the possibility of multicollinearity with any real certainty.

9.4 Remedies for multicollinearity

What can be done to minimize the consequences of severe multicollinearity? There is no automatic answer to this question because multicollinearity is a phenomenon that could change from sample to sample even for the same specification of a regression equation. The purpose of this section is to outline a number of alternative remedies for multicollinearity that might be appropriate under certain circumstances.

The first step to take once severe multicollinearity has been diagnosed is to decide whether anything should be done at all. As we'll see, it turns out that every remedy for multicollinearity has a drawback of some sort, and so it often happens that doing nothing is the correct course of action. One reason for doing nothing is that multicollinearity in an equation will not always reduce the t -scores enough to make them insignificant or change the $\hat{\beta}$ s enough to make them differ from expectations. In other words, the mere existence of multicollinearity does not necessarily mean anything. A remedy for multicollinearity should be considered only if the consequences cause insignificant t -scores or unreliable estimated coefficients. For example, it's possible to observe a simple correlation coefficient of 0.97 between two explanatory variables and yet have each individual t -score be significant. It makes no sense to consider remedial action in such a case, as long as both variables belong in the equation on theoretical grounds, because any remedy for multicollinearity would probably cause other exercises for the equation. In a sense, multicollinearity is similar to a non-life-

threatening human disease that requires general anesthesia to operate on the patient: The risk of the operation should be undertaken only if the disease is causing a significant problem.

A second reason for doing nothing is that the deletion of a multicollinear variable that belongs in an equation will cause specification bias. If we drop a theoretically important variable, then we are *purposely* creating bias. Given all the effort typically spent avoiding omitted variables, it seems foolhardy to consider running that risk on purpose. As a result, experienced econometricians often will leave multicollinear variables in equations despite low t -scores.

The final reason for considering doing nothing to offset multicollinearity is that every time a regression is rerun, we risk encountering a specification that fits because it accidentally works for the particular data set involved, not because it is the truth. The larger the number of experiments, the greater the chances of finding the accidental result. To make things worse, when there is significant multicollinearity in the sample, the odds of strange results increase rapidly because of the sensitivity of the coefficient estimates to slight specification changes.

To sum, it is often best to leave an equation unadjusted in the face of all but extreme multicollinearity. Such advice might be difficult for beginning researchers to take, however, if they think that it's embarrassing to report that their final regression is one with insignificant t -scores. Compared to the alternatives of possible omitted variable bias or accidentally significant regression results, the low t -scores seem like a minor problem. For an example of "doing nothing" in the face of severe multicollinearity, see section 9.5.

Drop a redundant variable

On occasion, the simple solution of dropping one of the multicollinear variables is a good one. For example, some inexperienced researchers include too many variables in their regressions, not wanting to face omitted variable bias.

As a result, they often have two or more variables in their equations that are measuring essentially the same thing. In such a case the multicollinear variables are not irrelevant, since any one of them is quite probably theoretically and statistically sound. Instead, the variables might be called **redundant**; only one of them is

needed to represent the effect on the dependent variable that all of them currently represent. For example, in an aggregate demand function, it would not make sense to include disposable income and GDP because both are measuring the same thing: income. A bit more subtle is the inference that population and disposable income should not both be included in the same aggregate demand function because, once again, they really are measuring the same thing: the size of the aggregate market. As population rises, so too will income. Dropping these kinds of redundant multicollinear variables is doing nothing more than making up for a specification error; the variables should never have been included in the first place.

EXAMPLE 9.1

To see how this solution would work, let's return to the student consumption function example of equation (9.10):

$$\begin{aligned} \hat{CO}_i &= -367.83 + 0.5113Yd_i + 0.0427LA_i \\ &\quad \quad \quad (1.0307) \quad \quad \quad (0.0942) \\ &\quad \quad \quad t = 0.496 \quad \quad \quad 0.453 \\ &\quad \quad \quad \bar{R}^2 = 0.835 \end{aligned} \tag{9.10}$$

where CO = consumption, Yd = disposable income, and LA = liquid assets. When we first discussed this example, we compared this result to the same equation without the liquid assets variable:

$$\begin{aligned} \hat{CO}_i &= -471.43 + 0.9714Yd_i \\ &\quad \quad \quad (0.157) \\ &\quad \quad \quad t = 6.187 \\ &\quad \quad \quad \bar{R}^2 = 0.861 \end{aligned} \tag{9.11}$$

If we had instead dropped the disposable income variable, we would have obtained:

$$\begin{aligned} \hat{CO}_i &= -199.44 + 0.08876LA_i \\ &\quad \quad \quad (0.01443) \\ &\quad \quad \quad t = 6.153 \\ &\quad \quad \quad \bar{R}^2 = 0.860 \end{aligned} \tag{9.17}$$

Note that dropping one of the multicollinear variables has eliminated both the multicollinearity between the two explanatory variables and the low t -score of the coefficient of the remaining variable. By dropping Y_d , we were able to increase t_{LA} from 0.453 to 6.153. Since dropping a variable changes the meaning of the remaining coefficient (because the dropped variable is no longer being held constant), such dramatic changes are not unusual. The coefficient of the remaining included variable also now measures almost all of the joint impact on the dependent variable of the multicollinear explanatory variables.

Assuming you want to drop a variable, how do you decide which variable to drop? In cases of severe multicollinearity, it makes no statistical difference which variable is dropped. As a result, it doesn't make sense to pick the variable to be dropped on the basis of which one gives superior fit or which one is more significant (or has the expected sign) in the original equation. Instead, the theoretical underpinnings of the model should be the basis for such a decision. In the example of the student consumption function, there is more theoretical support for the hypothesis that disposable income determines consumption than there is for the liquid assets hypothesis. Therefore, equation (9.11) should be preferred to equation (9.17).

Increase the size of the sample

Another way to deal with multicollinearity is to attempt to increase the size of the sample to reduce the degree of multicollinearity. Although such an increase may be impossible, it's a useful alternative to be considered when feasible.

The idea behind increasing the size of the sample is that a larger data set (often requiring new data collection) will allow more accurate estimates than a small one, since the larger sample normally will reduce the variance of the estimated coefficients, diminishing the impact of the multicollinearity.

For most time series data sets, however, this solution isn't feasible. After all, samples typically are drawn by getting all the available data that seem similar. As a result, new data are generally impossible or quite expensive to find. Going out and generating new data is much easier with a cross-sectional or experimental data set than it is when the observations must be generated by the passage of time.

9.5 An example of why multicollinearity often is best left unadjusted

Let's look at an example of the idea that multicollinearity often should be left unadjusted. Suppose you work in the marketing department of a hypothetical soft drink company and you build a model of the impact on sales of your firm's advertising:

$$\begin{aligned} \hat{S}_t &= 3080 - 75.000 P_t + 4.23 A_t - 1.04 B_t \\ &\quad \begin{matrix} (25.000) & (1.06) & (0.51) \\ t = -3.00 & 3.99 & -2.04 \end{matrix} \\ \bar{R}^2 &= 0.825 \quad N = 28 \end{aligned} \tag{9.18}$$

where: S_t = sales of the soft drink in year t

P_t = average relative price of the drink in year t

A_t = advertising expenditures for the company in year t

B_t = advertising expenditures for the company's main competitor in year t

Assume that there are no omitted variables. All variables are measured in real dollars; that is, the nominal values are divided, or deflated, by a price index. On the face of it, this is a reasonable-looking result. Estimated coefficients are significant in the directions implied by the underlying theory, and both the overall fit and the size of the coefficients seem acceptable. Suppose you now were told that advertising in the soft drink industry is cut-throat in nature and that firms tend to match their main competitor's advertising expenditures. This would lead you to suspect that significant multicollinearity was possible. Further suppose that the simple correlation coefficient between the two advertising variables is 0.974 and that their respective VIFs are well over 5.

Such a correlation coefficient is evidence that there is severe multicollinearity in the equation, but there is no reason even to consider doing anything about it, because the coefficients are so powerful that their t -scores remain significant, even in the face of severe multicollinearity. Unless multicollinearity causes exercises in the equation, it should be left unadjusted. To change the specification might give us better-looking results, but the adjustment would decrease our chances of obtaining the best

possible estimates of the true coefficients. Although it's certainly lucky that there were no major exercises due to multicollinearity in this example, that luck is no reason to try to fix something that isn't broken.

When a variable is dropped from an equation, its effect will be absorbed by the other explanatory variables to the extent that they are correlated with the newly omitted variable. It's likely that the remaining multicollinear variable(s) will absorb virtually all the bias, since the variables are highly correlated. This bias may destroy whatever usefulness the estimates had before the variable was dropped. For example, if a variable, say B , is dropped from the equation to fix the multicollinearity, then the following might occur:

$$\begin{aligned}\hat{S}_t &= 2586 - 78.000 P_t + 0.52 A_t \\ &\quad \quad \quad (24.000) \quad \quad (4.32) \\ t &= -3.25 \quad \quad 0.12 \\ \bar{R}^2 &= 0.531 \quad \quad N = 28\end{aligned}\tag{9.19}$$

What's going on here? The company's advertising coefficient becomes less significant instead of more significant when one of the multicollinear variables is dropped. To see why, first note that the expected bias on $\hat{\beta}_A$ is negative because the product of the expected sign of the coefficient of B and of the correlation between A and B is negative:

$$\text{Bias} = \beta_B \times \hat{\alpha}_1 = (-)(+) = -\tag{9.20}$$

Second, this negative bias is strong enough to decrease the estimated coefficient of A until it is insignificant. Although this problem could have been avoided by using a relative advertising variable (A divided by B , for instance), that formulation would have forced identical absolute coefficients on A and $1/B$. Such identical coefficients will sometimes be theoretically expected or empirically reasonable but, in most cases, these kinds of constraints will force bias onto an equation that previously had none.

This example is simplistic, but its results are typical in cases in which equations are adjusted for multicollinearity by dropping a variable without regard to the effect that the deletion is going to have. The point here is that it's quite often theoretically or operationally unwise to drop a variable from an equation and that multicollinearity in such cases is best left unadjusted.

SUMMARY

1. Perfect multicollinearity is the violation of the assumption that no explanatory variable is a perfect linear function of other explanatory variable(s). Perfect multicollinearity results in indeterminate estimates of the regression coefficients and infinite standard errors of those estimates, making OLS estimation impossible.
2. Imperfect multicollinearity, which is what is typically meant when the word "multicollinearity" is used, is a linear relationship between two or more independent variables that is strong enough to significantly affect the estimation of the equation. Multicollinearity is a sample phenomenon as well as a theoretical one. Different samples can exhibit different degrees of multicollinearity.
3. The major consequence of severe multicollinearity is to increase the variances of the estimated regression coefficients and therefore decrease the calculated *t*-scores of those coefficients and expand the confidence intervals. Multicollinearity causes no bias in the estimated coefficients, and it has little effect on the overall significance of the regression or on the estimates of the coefficients of any nonmulticollinear explanatory variables.
4. Since multicollinearity exists, to one degree or another, in virtually every data set, the question to be asked in detection is how severe the multicollinearity in a particular sample is.
5. Two useful methods for the detection of severe multicollinearity are:

- a. Are the simple correlation coefficients between the explanatory variables high?
 - b. Are the variance inflation factors high?
If either of these answers is yes, then multicollinearity certainly exists, but multicollinearity can also exist even if the answers are no.
6. The three most common remedies for multicollinearity are:
- a. Do nothing (and thus avoid specification bias).
 - b. Drop a redundant variable.
 - c. Increase the size of the sample.
7. Quite often, doing nothing is the best remedy for multicollinearity. If the multicollinearity has not decreased t -scores to the point of insignificance, then no remedy should even be considered as long as the variables are theoretically strong. Even if the t -scores are insignificant, remedies should be undertaken cautiously, because all impose costs on the estimation that may be greater than the potential benefit of ridding the equation of multicollinearity.

CHAPTER 10. SERIAL CORRELATION

In the next two chapters we'll investigate the final component of the specification of a regression equation—choosing the correct form of the stochastic error term. Our first topic, serial correlation, is the violation of Classical assumption 6 that different observations of the error term are uncorrelated with each other. Serial correlation, also called autocorrelation, can exist in any research study in which the order of the observations has some meaning and occurs most frequently in time-series data sets. In essence, serial correlation implies that the value of the error term from one time period depends in some systematic way on the value of the error term in other time periods. Since time-series data are used in many applications of econometrics, it's important to understand serial correlation and its consequences for OLS estimators.

The approach of this chapter to the problem of serial correlation will be similar to that used in the previous chapter. We'll attempt to answer the same four questions:

1. What is the nature of the problem?
2. What are the consequences of the problem?
3. How is the problem diagnosed?
4. What remedies for the problem are available?

10.1 Time series

Virtually every equation in the text so far has been cross-sectional in nature, but that's going to change dramatically in this chapter. As a result, it's probably worthwhile to talk about some of the characteristics of time-series equations.

Time-series data involve a single entity (like a person, corporation, or state) over multiple points in time. Such a time-series approach allows researchers to investigate analytical issues that can't be examined very easily with a crosssectional regression.

For example, macroeconomic models and supply-and-demand models are best studied using time-series, not cross-sectional, data.

The notation for a time-series study is different from that for a cross-sectional one. Our familiar cross-sectional notation (for one time period and N different entities) is:

$$y_i = a + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

where i goes from 1 to N .

A time-series regression has one entity and T different time periods, however, so we'll switch to this notation:

$$y_t = a + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$$

where t goes from 1 to N .

Thus:

$y_1 = a + \beta_1 x_{11} + \beta_2 x_{21} + \beta_3 x_{31} + u_1$	refers to observations from the first time period
$y_2 = a + \beta_1 x_{12} + \beta_2 x_{22} + \beta_3 x_{32} + u_2$	refers to observations from the second time period
$y_T = a + \beta_1 x_{1T} + \beta_2 x_{2T} + \beta_3 x_{3T} + u_T$	refers to observations from the T th time period

What's so tough about that, you say? All we've done is change from i to t and change from N to T . Well, it turns out that time-series studies have some characteristics that make them more difficult to deal with than cross-sections:

1. *The order of observations in a time series is fixed.* With a cross-sectional data set, you can enter the observations in any order you want, but with timeseries data, you must keep the observations in chronological order.
2. *Time-series samples tend to be much smaller than cross-sectional ones.* Most time-series populations have many

fewer potential observations than do cross-sectional ones, and these smaller data sets make statistical inference more difficult. In addition, it's much harder to generate a time-series observation than a cross-sectional one. After all, it takes a year to get one more observation in an annual time series!

3. *The theory underlying time-series analysis can be quite complex.* In part because of the exercises mentioned above, time-series econometrics includes a number of complex topics that require advanced estimation techniques. We'll tackle these topics in chapters 12, 14, and 15.
4. *The stochastic error term in a time-series equation is often affected by events that took place in a previous time period.* This is serial correlation, the topic of our chapter, so let's get started!

10.2 Pure versus impure serial correlation

Pure serial correlation

Pure serial correlation occurs when Classical assumption 4, which assumes uncorrelated observations of the error term, is violated in a correctly specified equation. If there is correlation between observations of the error term, then the error term is said to be serially correlated. When econometricians use the term serial correlation without any modifier, they are referring to pure serial correlation.

The most commonly assumed kind of serial correlation is **first-order serial correlation**, in which the current value of the error term is a function of the previous value of the error term:

$$u_t = \rho u_{t-1} + \varepsilon_t \quad (10.1)$$

where: u = the error term of the equation in question

ρ = the first-order autocorrelation coefficient

ε = a classical (not serially correlated) error term

The functional form in equation (10.1) is called a first-order Markov scheme. The new symbol, ρ (rho, pronounced “row”), called the **first-order autocorrelation coefficient**, measures the functional relationship between the value of an observation of the error term and the value of the previous observation of the error term.

The magnitude of ρ indicates the strength of the serial correlation in an equation. If ρ is zero, then there is no serial correlation (because u would equal ε , a classical error term). As ρ approaches 1 in absolute value, the value of the previous observation of the error term becomes more important in determining the current value of u_t , and a high degree of serial correlation exists. For ρ to be greater than 1 in absolute value is unreasonable because it implies that the error term has a tendency to continually increase in absolute value over time (“explode”). As a result of this, we can state that:

$$-1 < \rho < +1 \quad (10.2)$$

The sign of ρ indicates the nature of the serial correlation in an equation. A positive value for ρ implies that the error term tends to have the same sign from one time period to the next; this is called **positive serial correlation**. Such a tendency means that if u_t happens by chance to take on a large value in one time period, subsequent observations would tend to retain a portion of this original large value and would have the same sign as the original. For example, in time-series models, the effects of a large external shock to an economy (like an earthquake) in one period may linger for several time periods. The error term will tend to be positive for a number of observations, then negative for several more, and then back positive again.

Figure 10.1 shows two different examples of positive serial correlation. The error term observations plotted in Figure 9.1 are arranged in chronological order, with the first observation being the first period for which data are available, the second being the second, and so on. To see the difference between error terms with

and without positive serial correlation, compare the patterns in Figure 10.1 with the depiction of no serial correlation ($\rho = 0$) in Figure 10.2.

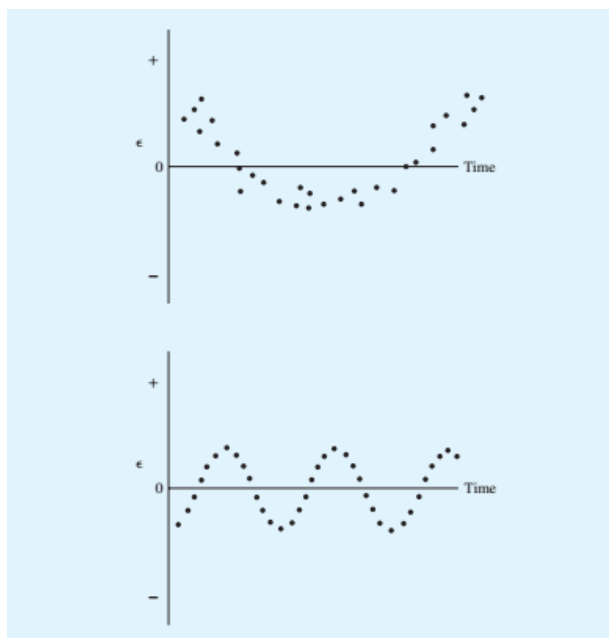


Figure 10.1 Positive serial correlation

With positive first-order serial correlation, the current observation of the error term tends to have the same sign as the previous observation of the error term. An example of positive serial correlation would be external shocks to an economy that take more than one time period to completely work through the system.

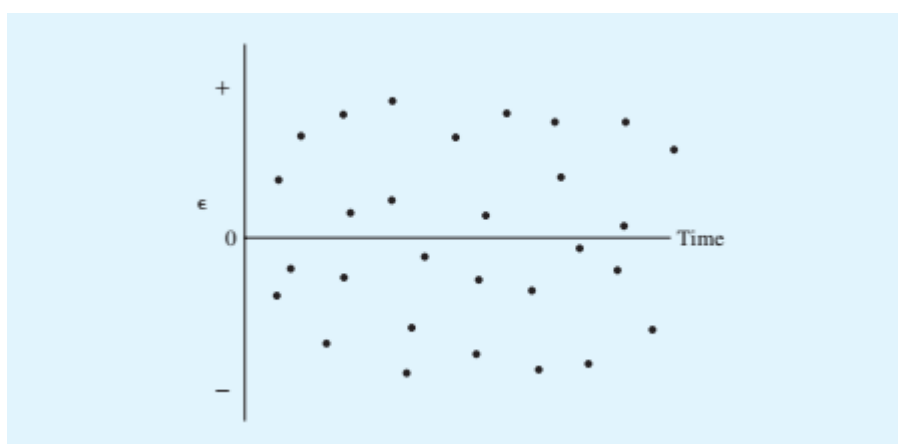


Figure 10.2 No serial correlation

With no serial correlation, different observations of the error term are completely uncorrelated with each other. Such error terms would conform to Classical assumption 4.

A negative value of ρ implies that the error term has a tendency to switch signs from negative to positive and back again in consecutive observations; this is called **negative serial correlation**. It implies that there is some sort of cycle (like a pendulum) behind the drawing of stochastic disturbances. Figure 10.3 shows two different examples of negative serial correlation. For instance, negative serial correlation might exist in the error term of an equation that is in first differences because *changes* in a variable often follow a cyclical pattern. In most time-series applications, however, negative pure serial correlation is much less likely than positive pure serial correlation. As a result, most econometricians analyzing pure serial correlation concern themselves primarily with positive serial correlation. Serial correlation can take on many forms other than first-order serial correlation. For example, in a quarterly model, the current quarter's error term observation may be functionally related to the observation of the error term from the same quarter in the previous year. This is called *seasonally based serial correlation*:

$$u_t = \rho u_{t-4} + \varepsilon_t$$

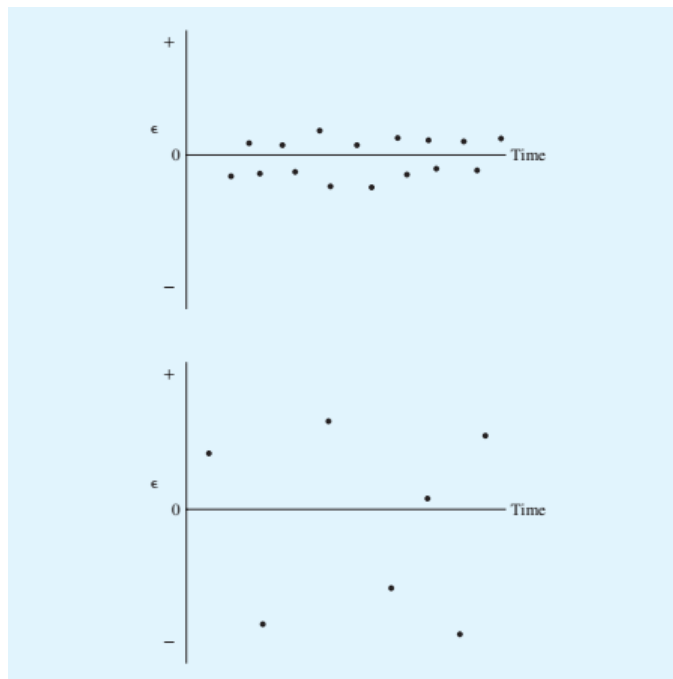


Figure 10.3 Negative serial correlation

With negative first-order serial correlation, the current observation of the error term tends to have the opposite sign from the previous observation of the error term. In most time-series applications, negative serial correlation is much less likely than positive serial correlation.

Impure serial correlation

By **impure serial correlation** we mean serial correlation that is caused by a specification error such as an omitted variable or an incorrect functional form. While pure serial correlation is caused by the underlying distribution of the error term of the true specification of an equation (which cannot be changed by the researcher), impure serial correlation is caused by a specification error that often can be corrected.

How is it possible for a specification error to cause serial correlation? Recall that the error term can be thought of as the effect of omitted variables, nonlinearities, measurement errors, and pure stochastic disturbances on the dependent variable. This means, for example, that if we omit a relevant variable or use the wrong functional form, then the portion of that omitted effect that cannot be represented by the included explanatory variables must be absorbed by the error term. The error term for an incorrectly specified equation thus includes a portion of the effect of any omitted variables and/or a portion of the effect of the difference between the proper functional form and the one chosen by the researcher. This new error term might be serially correlated even if the true one is not. If this is the case, the serial correlation has been caused by the researcher's choice of a specification and not by the pure error term associated with the correct specification.

As you'll see in section 10.5, the proper remedy for serial correlation depends on whether the serial correlation is likely to be pure or impure. Not surprisingly, the best remedy for impure serial correlation is to attempt to find the omitted variable (or at least a good proxy) or the correct functional form for the equation. Both the bias and the impure serial correlation will disappear if the specification error is corrected. As a result, most econometricians try to make sure they have the best specification possible before they spend too much time worrying about pure serial correlation.

To see how an omitted variable can cause the error term to be serially correlated, suppose that the true equation is:

$$y_t = a + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t \quad (10.3)$$

where u_t is a classical error term. As shown in section 7.1, if x_2 is accidentally omitted from the equation (or if data for x_2 are unavailable), then:

$$y_t = a + \beta_1 x_{1t} + u_t^* \quad \text{where} \quad u_t^* = \beta_2 x_{2t} + u_t \quad (10.4)$$

Thus, the error term in the omitted variable case is not the classical error term e . Instead, it's also a function of one of the independent variables, x_2 . As a result, the new error term, e^* , can be serially correlated even if the true error term e is not. In particular, the new error term e^* will tend to exhibit detectable serial correlation when:

1. x_2 itself is serially correlated (this is quite likely in a time series) *and*
2. the size of u is small⁴⁸ compared to the size of $\beta_2 \bar{x}_2$.

These tendencies hold even if there are a number of included and/or omitted variables. Therefore:

$$u_t^* = \rho u_{t-4}^* + \varepsilon_t \quad (10.5)$$

Another common kind of impure serial correlation is caused by an incorrect functional form. Here, the choice of the wrong functional form can cause the error term to be serially correlated. Let's suppose that the true equation is polynomial in nature:

⁴⁸ If typical values of u are significantly larger in absolute size than $\beta_2 \bar{x}_2$, then even a serially correlated omitted variable (x_2) will not change u^* very much. In addition, recall that the omitted variable, x_2 , will cause bias in the estimate of β_1 , depending on the correlation between the two x s. If $\hat{\beta}_1$ is biased because of the omission of x_2 , then a portion of the $\beta_2 \bar{x}_2$ effect must have been absorbed by $\hat{\beta}_1$ and will not end up in the residuals. As a result, tests for serial correlation based on those residuals may give incorrect readings. Such residuals may leave misleading clues as to possible specification errors.

$$y_t = a + \beta_1 x_{1t} + \beta_2 x_{1t}^2 + u_t \quad (10.6)$$

but that instead a linear regression is run:

$$y_t = \alpha_0 + \alpha_1 x_{1t} + u_t^* \quad (10.7)$$

The new error term u_t^* is now a function of the true error term u and of the differences between the linear and the polynomial functional forms. As can be seen in Figure 10.4, these differences often follow fairly autoregressive patterns. That is, positive differences tend to be followed by positive differences, and negative differences tend to be followed by negative differences. As a result, using a linear functional form when a nonlinear one is appropriate will usually result in positive impure serial correlation.

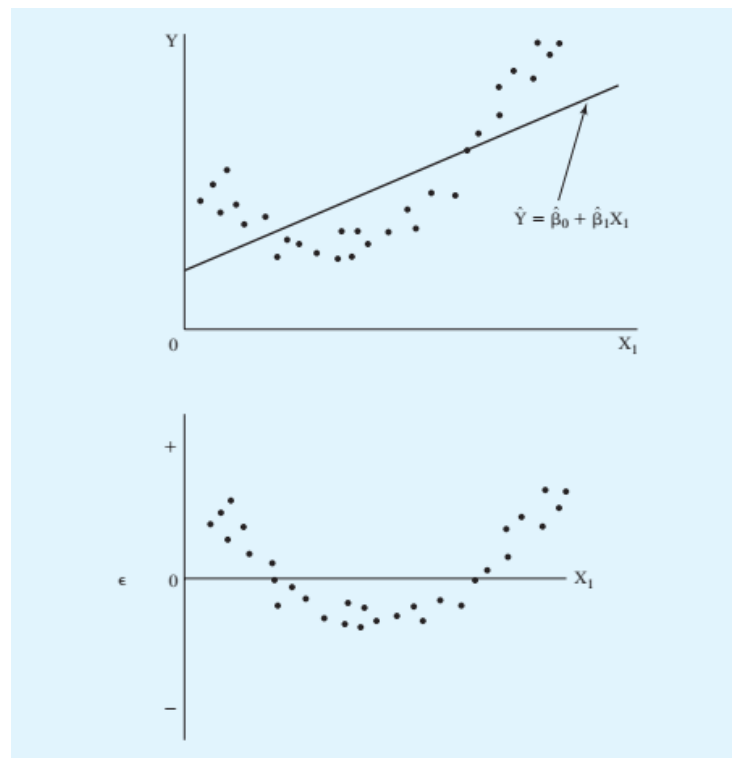


Figure 10.4 Incorrect functional form as a source of impure serial correlation

The use of an incorrect functional form tends to group positive and negative residuals together, causing positive impure serial correlation.

10.3 The consequences of serial correlation

The consequences of serial correlation are quite different in nature from the consequences of the exercises discussed so far in this text. Omitted variables, irrelevant variables, and multicollinearity all have fairly recognizable external symptoms. Each problem changes the estimated coefficients and standard errors in a particular way, and an examination of these changes (and the underlying theory) often provides enough information for the problem to be detected. As we shall see, serial correlation is more likely to have internal symptoms; it affects the estimated equation in a way that is not easily observable from an examination of just the results themselves.

The existence of serial correlation in the error term of an equation violates Classical assumption 4, and the estimation of the equation with OLS has at least three consequences:⁴⁹

1. Pure serial correlation does not cause bias in the coefficient estimates.
2. Serial correlation causes OLS to no longer be the minimum variance estimator (of all the linear unbiased estimators).
3. Serial correlation causes the OLS estimates of the $SE(\hat{\beta})$ s to be biased, leading to unreliable hypothesis testing.

Lets consider each consequences in detail.

1. *Pure serial correlation does not cause bias in the coefficient estimates.* If the error term is serially correlated, one of the assumptions of the Gauss - Markov theorem is violated, but this violation does not cause the coefficient estimates to be biased. If the serial correlation is impure, however, bias may be introduced by the use of an incorrect specification.

This lack of bias does not necessarily mean that the OLS estimates of the coefficients of a serially correlated equation will be close to the true coefficient values. A single estimate observed in practice can come from a wide range of possible

⁴⁹ If the regression includes a lagged dependent variable as an independent variable, then the problems worsen significantly. For more on this topic (called a dynamic model), see chapter 13.

values. In addition, the standard errors of these estimates will typically be increased by the serial correlation. This increase will raise the probability that a β_n will differ significantly from the true β value. What unbiased means in this case is that the distribution of the $\hat{\beta}$ s is still centered around the true β .

2. *Serial correlation causes OLS to no longer be the minimum variance estimator (of all the linear unbiased estimators).* Although the violation of Classical assumption 4 causes no bias, it does affect the other main conclusion of the Gauss - Markov theorem, that of minimum variance. In particular, we cannot prove that the distribution of the OLS $\hat{\beta}$ s is minimum variance (among the linear unbiased estimators) when assumption 4 is violated.

The serially correlated error term causes the dependent variable to fluctuate in a way that the OLS estimation procedure sometimes attributes to the independent variables. Thus, OLS is more likely to misestimate the true β in the face of serial correlation. On balance, the β_n s are still unbiased because overestimates are just as likely as underestimates, but these errors increase the variance of the distribution of the estimates, increasing the amount that any given estimate is likely to differ from the true β .

3. *Serial correlation causes the OLS estimates of the $SE(\hat{\beta})$ s to be biased, leading to unreliable hypothesis testing.* With serial correlation, the OLS formula for the standard error produces biased estimates of the $SE(\hat{\beta})$ s. Because the $SE(\hat{\beta})$ is a prime component in the t -statistic, these biased $SE(\hat{\beta})$ s cause biased t -scores and unreliable hypothesis testing in general. In essence, serial correlation causes OLS to produce *incorrect* $SE(\beta_n)$ s and t -scores! Not surprisingly, most econometricians therefore are very hesitant to put much faith in hypothesis tests that were conducted in the face of pure serial correlation.

What sort of bias does serial correlation tend to cause? Typically, the bias in the estimate of $SE(\hat{\beta})$ is negative, meaning that OLS underestimates the size of the standard errors of the coefficients. This comes about because serial

correlation usually results in a pattern of observations that allows a better fit than the actual (not serially correlated) observations would otherwise justify. This tendency of OLS to underestimate the $SE(\hat{\beta})$ means that OLS typically overestimates the t scores of the estimated coefficients, since:

$$t_k = \frac{(\hat{\beta}_k - \beta_{H_0})}{SE(\hat{\beta}_k)} \quad (10.8)$$

Thus the t -scores printed out by a typical software regression package in the face of serial correlation are likely to be too high. Similarly, confidence intervals for the coefficients will tend to be too narrow.

What will happen to hypothesis testing if OLS underestimates the $SE(\hat{\beta})$ s and therefore overestimates the t -scores? Well, the “too low” $SE(\hat{\beta})$ will cause a “too high” t -score for a particular coefficient, and this will make it more likely that we will reject a null hypothesis (for example $H_0: \beta \leq 0$) when it is in fact true. This increased chance of rejecting H_0 means that we’re more likely to reject a true null hypothesis, so we’re more likely to make the mistake of keeping an irrelevant variable in an equation because its coefficient’s t -score has been overestimated. In other words, hypothesis testing becomes unreliable when we have pure serial correlation.

10.4 The detection of serial correlation

How can we detect serial correlation? While the first indication of serial correlation often occurs when we observe a pattern in the residuals similar to Figure 10.1, most econometricians rely on more formal tests like the Durbin - Watson test and the Lagrange Multiplier test.

The Durbin–Watson test

The **Durbin–Watson test** is used to determine if there is first-order serial correlation in the error term of an equation by examining the *residuals* of a particular estimation of that equation.⁵⁰ It's important to use the Durbin - Watson test only when the assumptions that underlie its derivation are met:

1. The regression model includes an intercept term.
2. The serial correlation is first-order in nature:

$$u_t^* = \rho u_{t-1}^* + \varepsilon_t \quad (10.9)$$

where ρ is the autocorrelation coefficient and u is a classical (normally distributed) error term.

3. The regression model does not include a lagged dependent variable (discussed in Chapter 13) as an independent variable.⁵¹

The equation for the *Durbin – Watson statistic* for T observations is:

$$d = \frac{\sum_{t=2}^T (u_t - u_{t-1})^2}{\sum_{t=1}^T u_t^2} \quad (10.10)$$

where the u_t s are the OLS residuals. Note that the numerator has one fewer observation than the denominator because an observation must be used to calculate u_{t-1} . The Durbin–Watson statistic equals 0 if there is extreme positive serial correlation, 2 if there is no serial correlation, and 4 if there is extreme negative serial correlation. To see this, let's put appropriate residual values into equation (10.10) for these three cases:

1. Extreme positive serial correlation: $d = 0$.
In this case, $u_t = u_{t-1}$, so $(u_t - u_{t-1}) = 0$ and $d = 0$.
2. Extreme negative serial correlation: $d \approx 4$
In this case, $u_t = -u_{t-1}$, and $(u_t - u_{t-1}) = (2u_t)$. Substituting into equation (10.10), we obtain $d = \sum (2u_t)^2 / \sum (u_t)^2$ and $d \approx 4$.

⁵⁰ J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least-Squared Regression," *Biometrika*, 1951, pp. 159–177.

⁵¹ In such a circumstance, the Durbin–Watson test is biased toward 2.

3. No serial correlation: $d \approx 2$

When there is no serial correlation, the mean of the distribution of d is equal to 2.⁵² That is, if there is no serial correlation, $d \approx 2$.

Using the Durbin–Watson test

The Durbin - Watson test is unusual in two respects. First, econometricians almost never test the one-sided null hypothesis that there is negative serial correlation in the residuals because negative serial correlation, as mentioned previously, is quite difficult to explain theoretically in economic or business analysis. Its existence usually means that impure serial correlation has been caused by some specification error.

Second, the Durbin - Watson test is sometimes inconclusive. Whereas previously explained decision rules always have had only “acceptance” regions and rejection regions, the Durbin–Watson test has a third possibility, called the inconclusive region. For reasons outlined in section 10.5, we do not recommend the application of a remedy for serial correlation if the Durbin - Watson test is inconclusive.⁵³

With these exceptions, the use of the Durbin - Watson test is quite similar to the use of the t -test. To test for positive serial correlation, the following steps are required:

1. Obtain the OLS residuals from the equation to be tested and calculate the d statistic by using equation (10.10).
2. Determine the sample size and the number of explanatory variables and then consult Statistical Table B-4 in Appendix

⁵² To see this, multiply out the numerator of Equation 9.10, obtaining

$$d = \left[\sum_1^T u_i^2 - 2 \sum_2^T (u_i u_{i-1}) + \sum_2^T u_{i-1}^2 \right] / \left[\sum_2^T u_i^2 - 2 \sum_2^T (u_i u_{i-1}) + \sum_2^T u_{i-1}^2 \right] / \sum_1^T u_i^2 \approx \left[\sum_2^T u_i^2 + \sum_2^T u_{i-1}^2 \right] / \sum_1^T u_i^2 \approx 2 \quad (10.11)$$

If there is no serial correlation, then u_i and u_{i-1} are not related, and, on average, $\sum (u_i u_{i-1}) = 0$.

⁵³ This inconclusive region is troubling, but the development of exact Durbin - Watson tests may eliminate this problem in the future. Some computer programs allow the user the option of calculating an exact Durbin - Watson probability (of first-order serial correlation). Alternatively, it's worth noting that there is a growing trend toward the use of d_U as a sole critical value. This trend runs counter to our view that if the Durbin–Watson test is inconclusive, then no remedial action should be taken except to search for a possible cause of impure serial correlation.

B of the tutorial Econometric analysis using Eviews: a Practical tutorial to find the upper critical d value, d_U , and the lower critical d value, d_L , respectively. Instructions for the use of this Table are also in that appendix.

3. Given the null hypothesis of no positive serial correlation and a onesided alternative hypothesis:

$$\begin{aligned} H_0: \rho &\leq 0 \text{ (no positive serial correlation)} \\ H_1: \rho &> 0 \text{ (positive serial correlation)} \end{aligned} \quad (10.12)$$

the appropriate decision rule is:

If $d < d_L$	Reject H_0
If $d > d_U$	Do not reject H_0
If $d_L \leq d \leq d_U$	Inconclusive

In rare circumstances, perhaps first differenced equations, a two-sided Durbin Watson test might be appropriate. In such a case, steps 1 and 2 are still used, but step 3 is now:

Given the null hypothesis of no serial correlation and a two-sided alternative hypothesis:

$$\begin{aligned} H_0: \rho &= 0 \text{ (no serial correlation)} \\ H_1: \rho &\neq 0 \text{ (serial correlation)} \end{aligned} \quad (10.13)$$

the appropriate decision rule is:

If $d < d_L$	Reject H_0
If $d > 4-d_L$	Do not reject H_0
If $4-d_U > d > d_U$	Inconclusive

Examples of the use of the Durbin - Watson statistics

Let's work through some applications of the Durbin - Watson test. First, turn to Statistical Table B-4. Note that the upper and lower critical d values (d_U and d_L) depend on the number of

explanatory variables (do not count the constant term), the sample size, and the level of significance of the test.

Now let's set up a one-sided 5% test for a regression with three explanatory variables and 25 observations. As can be seen from 5% Table B-4 in the Appendix of tutorial *Econometric Analysis using Eviews: a Practical tutorial*, the critical values are $d_L = 1.12$ and $d_U = 1.66$. As a result, if the hypotheses are:

$H_0: \rho \leq 0$ (no positive serial correlation)

$H_1: \rho > 0$ (positive serial correlation)

the appropriate decision rule is:

if $d \leq 1.12$	Reject H_0
if $d > 1.66$	Do not reject H_0
if $1.12 \leq d \leq 1.66$	Inconclusive

A computed Durbin - Watson statistic of 1.78, for example, would indicate that there is no evidence of positive serial correlation, a value of 1.28 would be inconclusive, and a value of 0.60 would imply positive serial correlation. Figure (10.5) provides a graph of the “acceptance,” rejection, and inconclusive regions for this example.

For a real-world example, let's look at a simple time-series model of the annual consumption of chicken in the United States. There are a variety of variables that might make sense in such an

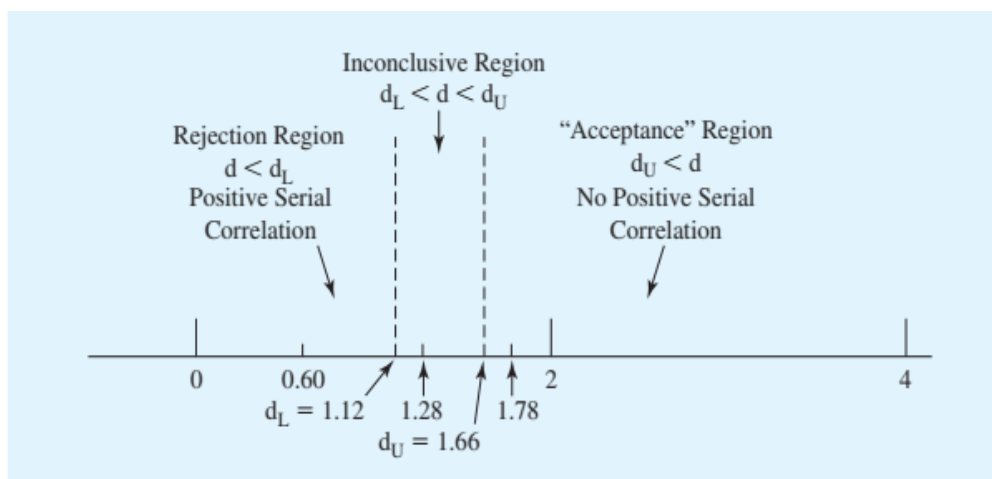


Figure 10.5 An example of a one-sided Durbin - Watson test

In a one-sided Durbin - Watson test for positive serial correlation, only values of d significantly below 2 cause the null hypothesis of no positive serial correlation to be rejected. In this example, a d of 1.78 would indicate no positive serial correlation, a d of 0.60 would indicate positive serial correlation, and a d of 1.28 would be inconclusive.

equation, and at least three variables seem obvious. We'd expect the demand for chicken to be a negative function of the price of chicken and a positive function of income and the price of a substitute (in this case, beef):

$$Y_t = a + \beta_1^- PC_t + \beta_2^+ PB_t + \beta_3^+ YD_t + u_t$$

where: Y_t = per capita chicken consumption (in kg) in year t

PC_t = the price of chicken (in somoni per kg) in year t

PB_t = the price of beef (in somoni per kg) in year t

YD_t = per capita disposable income (in hundreds of somoni) in year t

If we collect data for these variables for the years 1986 through 2014, we can estimate⁵⁴ the following equation:

$$\begin{aligned} \hat{Y}_t &= 27.7 - \underset{(0.03)}{0.11} PC_t + \underset{(0.02)}{0.03} PB_t + \underset{(0.01)}{0.23} YD_t \\ &\quad t = -3.38 \quad +1.86 \quad +15.7 \\ \bar{R}^2 &= 0.9904 \quad N = 29 \end{aligned} \tag{10.14}$$

How does our estimated equation look? The overall fit of equation (10.14) is excellent, and each of the individual regression coefficients is significantly different from zero in the expected direction. The price of chicken does indeed have a significant negative effect (holding the price of beef and disposable income constant), and the price of beef and disposable income do indeed

⁵⁴ As we'll see in Chapter 16, estimating an equation for the demand for chicken without taking into account the simultaneously determined supply of chicken runs the risk of bias, particularly in the coefficient of the price of chicken.

have positive effects (holding the other independent variables constant).

However, this is a time-series equation, so if there's serial correlation, hypothesis testing will be unreliable, and one or more of these t -scores could be artificially high. We'd better run a Durbin - Watson test! When we calculate a Durbin - Watson statistic for equation 10.14,⁵⁵ we get 0.99. Is that cause to be concerned about serial correlation? What would be the result of a one-sided 5-percent test of the null hypothesis of no positive serial correlation? Well, once we've got the Durbin - Watson statistic, the next step is to consult Statistical Table B-4. In that Table, with K (the number of explanatory variables⁵⁶) equal to 3 and N (the number of observations) equal to 29, we find that the critical d values are $d_L = 1.20$ and $d_U = 1.65$. (It's probably a good idea to check these d values yourself to make sure that you know how to look them up.)

The decision rule would thus be:

If $d < 1.20$	Reject H_0
If $d > 1.65$	Do not reject H_0
If $1.20 \leq d \leq 1.65$	Inconclusive

Since 0.99 is less than the critical lower limit of the d statistic, we would reject the null hypothesis of no positive serial correlation, and we would have to decide how to cope with that serial correlation.

The Lagrange multiplier (LM) test

⁵⁵ Luckily, you don't actually need to calculate the Durbin - Watson statistic yourself. Some econometric software programs, including Eviews, calculate the Durbin-Watson statistic automatically, while others, including Stata, allow you to do so quite simply. In Stata, for example, the command is *estat dwatson*.

⁵⁶ Be careful! While we define K as the number of explanatory variables, some other sources, including Stata and the Stanford University Durbin-Watson tables, define K as the number of coefficients (which is equivalent to $K + 1$ in our notation). As long as you're aware of this difference, it won't cause you any problems. Incidentally, the Stanford tables, which are online at <http://web.stanford.edu/~clint/bench/dwcrit.htm>, have many more observations than can be printed in a textbook, so they're quite useful if you have a large sample.

Unfortunately, the Durbin - Watson test has a number of limitations. As mentioned, it can be used only when the serial correlation is first-order, when a constant is included in the equation, and when the equation doesn't include a lagged dependent variable. The Durbin - Watson test's inconclusive region also is a drawback, particularly since the size of the inconclusive region increases as the number of independent variables increases.

A popular alternative to the Durbin-Watson test is the **Lagrange Multiplier (LM) test**, which checks for serial correlation by analyzing how well the lagged residuals explain the residual of the original equation in an equation that also includes all the explanatory variables of the original model. If the lagged residuals are significant in explaining this time's residuals (as shown by the chi-square test), then we can reject the null hypothesis of no serial correlation.⁵⁷ The LM serial correlation test is just one application of a general Lagrange Multiplier testing approach that can be applied to a variety of econometric exercises.⁵⁸

Using the Lagrange Multiplier (LM) test to investigate the possibility of serial correlation involves three steps:

1. Obtain the residuals from the estimated equation. For an equation with two independent variables, this would equal:

$$u_t = y_t - \hat{y}_t = y_t - \hat{a} - \hat{\beta}_1 x_{1t} - \hat{\beta}_2 x_{2t} \quad (10.15)$$

2. Use these residuals as the dependent variable in an auxiliary equation that includes as independent variables all those on the right-hand side of the original equation as well as the lagged residuals:

⁵⁷ The Lagrange Multiplier test for serial correlation is sometimes referred to as the Breusch - Godfrey test. Note that if we're testing for first-order serial correlation, we need to specify that the lag equals 1. If we are concerned with second-order serial correlation, then the lag would equal 2, etc.

⁵⁸ For example, the White test for heteroskedasticity (to be explained in section 11.3) also is an application of the Lagrange Multiplier approach. For a survey of the various uses to which Lagrange Multiplier tests can be put, see Rob Engle, "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics," in Z. Griliches and M. D. Intrilligator (eds.), *Handbook of Econometrics*, Vol. II (Amsterdam, Elsevier Science Publishers, 1984).

$$u_t = \alpha_0 + \alpha_1 x_{1t} + \alpha_2 x_{2t} + \alpha_3 u_{t-1} + \varepsilon_t \quad (10.16)$$

3. Estimate equation (10.16) using OLS and then test the null hypothesis that $\alpha_3 = 0$ with the following test statistic:

$$LM = NR2$$

where N is the sample size and R^2 is the unadjusted coefficient of determination, both of the auxiliary equation, equation (10.16).

For large samples, LM has a chi-square distribution with degrees of freedom equal to one (the number of restrictions in the null hypothesis). If LM is greater than the critical chi-square value from Statistical Table B-6, then we reject the null hypothesis that $\alpha_3 = 0$ and conclude that there is serial correlation in the original equation. Note that even though α_3 tends to be positive in economic examples, this is a two-sided test.

An example of the Lagrange multiplier test

As an example of the Lagrange Multiplier test, let's run a 5% test for serial correlation on our chicken demand model, equation (10.14). The appropriate LM equation to run is:

$$u_t = \alpha_0 + \alpha_1 PC_t + \alpha_2 PB_t + \alpha_3 YD_t + \alpha_4 u_{t-1} + \varepsilon_t \quad (10.17)$$

where u_t is the residual from equation (10.14), the equation that we're testing for serial correlation.

Since there are three independent variables, the null hypothesis becomes $H_0: \alpha_4 = 0$. If we estimate equation (10.17), we get an R^2 of 0.291. Since the sample size is 29, this means that:

$$LM = NR2 = 8.439$$

The decision rule is to reject the null hypothesis if $NR2$ is greater than the critical chi-square value with 1 degree of freedom, so the next step is to consult Table B-6 and look up the critical

value. As you can see if you take a look at Table B-6, the 5% chi-square critical value with 1 degree of freedom is 3.84. Since $8.439 > 3.84$, we can reject the null hypothesis and conclude that we have serial correlation in the chicken demand model. This is a two-sided test, but it confirms the result of our one-sided Durbin - Watson test. It seems clear that the chicken demand equation has serial correlation!

Suppose that the Durbin - Watson or LM test detects serial correlation in the residuals of your equation. Is there a remedy? Some students suggest reordering the observations of y and the x s to avoid serial correlation. They think that if this time's error term appears to be affected by last time's error term, why not reorder the data randomly to get rid of the problem? The answer is that the reordering of the data does not get rid of the serial correlation; it just makes the problem harder to detect. If $u_2 = f(u_1)$ and we reorder the data, then the error term observations are still related to each other, but they now no longer follow each other, and it becomes almost impossible to discover the serial correlation. Interestingly, reordering the data changes the Durbin - Watson statistic but does not change the estimates of the coefficients or their standard errors at all.⁵⁹

The place to start in correcting a serial correlation problem is to look carefully at the specification of the equation for possible errors that might be causing impure serial correlation. Is the functional form correct? Are you sure that there are no omitted variables? Only after the specification of the equation has been reviewed carefully should the possibility of an adjustment for pure serial correlation be considered.

It's worth noting that if an omitted variable increases or decreases over time, as is often the case, or if the data set is logically reordered (say, according to the magnitude of one of the variables), then the Durbin - Watson or LM test can help detect impure serial correlation. A significant Durbin - Watson or LM test

⁵⁹ This can be proven mathematically, but it is usually more instructive to estimate a regression yourself, change the order of the observations, and then reestimate the regression. See exercise 3 at the end of the chapter.

result can easily be caused by an omitted variable or an incorrect functional form. In such circumstances, the Durbin - Watson or LM tests do not distinguish between pure and impure serial correlation, but the detection of negative serial correlation is often a strong hint that the serial correlation is impure.

If you conclude that you have pure serial correlation, then the appropriate response is to consider the application of generalized least squares or Newey - West standard errors, as described in the following sections.

Generalized least squares

Generalized least squares (GLS) is a method of ridding an equation of pure first-order serial correlation and in the process restoring the minimum variance property to its estimation. GLS starts with an equation that does not meet the Classical Assumptions (due in this case to the pure serial correlation in the error term) and transforms it into one equation (10.22) that does meet those assumptions.

At this point, you could skip directly to equation (10.22), but it's easier to understand the GLS estimator if you examine the transformation from which it comes. Start with an equation that has first-order serial correlation:

$$y_t = a + \beta_1 x_{1t} + u_t \quad (10.18)$$

which, if $u_t = \rho u_{t-1} + \varepsilon_t$ (due to pure serial correlation), also equals:

$$y_t = a + \beta_1 x_{1t} + \rho u_{t-1} + \varepsilon_t \quad (10.19)$$

where u is the serially correlated error term, ρ is the autocorrelation coefficient, and ε is a classical (not serially correlated) error term.

If we could get the ρu_{t-1} term out of equation (10.19), the serial correlation would be gone, because the remaining portion of the error term (ε_t) has no serial correlation in it. To rid ρu_{t-1} from

equation (10.19), multiply equation (10.18) by ρ and then lag the new equation by one time period, obtaining

$$\rho y_{t-1} = \rho a + \rho \beta_1 x_{1t-1} + \rho u_t \quad (10.20)$$

Notice that we now have an equation with a pet-1 term in it. If we now subtract equation (10.20) from equation (10.19), the equivalent equation that remains no longer contains the serially correlated component of the error term:

$$y_t - \rho y_{t-1} = a(1 - \rho) + \beta_1(x_{1t} - \rho x_{1t-1}) + u_t \quad (10.21)$$

Equation (10.21) can be rewritten as:

$$y_t^* = a^* + \beta_1 x_{1t}^* + u_t \quad (10.22)$$

$$\begin{aligned} \text{where} \quad y_t^* &= y_t - \rho y_{t-1} \\ x_{1t}^* &= x_{1t} - \rho x_{1t-1} \\ a^* &= a - \rho a \end{aligned} \quad (10.23)$$

Equation (10.22) is called a generalized least squares (or “quasi-differenced”) version of equation (10.19). Notice that:

1. The error term is not serially correlated. As a result, OLS estimation of equation (10.22) will be minimum variance. (This is true if we know ρ or if we accurately estimate ρ .)
2. The slope coefficient β_1 is the same as the slope coefficient of the original serially correlated equation, equation (10.19). Thus coefficients estimated with GLS have the same meaning as those estimated with OLS.
3. The dependent variable has changed compared to that in equation (10.19). This means that the GLS $\bar{R}^2$ is not necessarily comparable to the OLS $\bar{R}^2$.

Unfortunately we can't use OLS to estimate a generalized least squares model because GLS equations are inherently nonlinear in the coefficients. To see why, take a look at equation (10.21). We need to estimate values not only for β_0 and β_1 but also for ρ , and ρ

is multiplied by β_0 and β_1 (which you can see if you multiply out the right-hand side of the equation). Since OLS requires that the equation be linear in the coefficients, we need a different estimation procedure.

Luckily, there are a number of techniques that can be used to estimate GLS equations. While the best-known of these is the *Cochrane–Orcutt method*,⁶⁰ our recommendation is to use a slightly different approach, the Prais - Winsten method. The **Prais - Winsten method**⁶¹ is a two-step iterative technique that rids an equation of serial correlation by first producing an estimate of ρ and then estimating the GLS equation using that ρ . The two steps are:

1. Estimate ρ by running a regression based on the residuals of the equation suspected of having serial correlation:

$$u_t = \rho u_{t-1} + u_t \quad (10.24)$$

where the u_t s are the OLS residuals from the equation suspected of having pure serial correlation and u_t is a classical (non-serially-correlated) error term.

2. Use this ρ to estimate the GLS equation by substituting ρ into equation (10.21) and using OLS to estimate equation (10.21) with the adjusted data.

These two steps are repeated (iterated) until further improvement results in little change in $\hat{\rho}$. Once $\hat{\rho}$ has converged (usually in just a few iterations), the last estimate of step 2 is used as the final estimate of equation (10.21).

Unfortunately, all methods of estimating GLS equations use iterative nonlinear regression techniques that are well beyond the

⁶⁰ D. Cochrane and G. H. Orcutt, "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Association*, 1949, pp. 32–61.

⁶¹ S. J. Prais and C. B. Winsten, "Trend Estimators and Serial Correlation," *Cowles Commission Discussion Paper No. 383* (1954) Chicago. The Prais–Winsten method (sometimes called Yule–Walker) is very similar to Cochrane–Orcutt, but the Prais–Winsten estimate of ρ is more accurate because it uses the first observation in Step 1 while Cochrane–Orcutt does not. For more, see Masahito Kobayashi, "Comparison of Efficiencies of Several Estimators for Linear Regressions with Autocorrelated Errors," *Journal of the American Statistical Association*, 1985, pp. 951–953.

scope of this text. As a result, most researchers rely on their econometric software packages to estimate their GLS equations for them. In Stata, for example, the Prais - Winsten method can be run using the command *prais* followed by a listing of the dependent and independent variables.⁶²

Let's apply generalized least squares, using the Prais - Winsten method, to the chicken demand example that was found to have serial correlation in the previous section. Recall that we estimated the per capita demand for chicken as a function of the price of chicken, the price of beef, and disposable income:

$$\begin{aligned}\hat{Y}_t &= 27.7 - \underset{(0.03)}{0.11}PC_t + \underset{(0.02)}{0.03}PB_t + \underset{(0.01)}{0.23}YD_t \\ t &= -3.38 \quad +1.86 \quad +15.7 \\ \bar{R}^2 &= 0.9904 \quad N = 29 \quad DW = 0.99\end{aligned}\tag{10.14}$$

Note that we have added the Durbin - Watson statistic to the documentation with the notation *DW*. All future time-series results will include the *DW* statistic, but cross-sectional documentation of the *DW* is not required unless the observations are ordered in some meaningful manner (like smallest to largest or youngest to oldest).

If we reestimate equation (10.14) with the Prais - Winsten approach to GLS, we obtain:

$$\begin{aligned}\hat{Y}_t &= 28.5 - \underset{(0.04)}{0.08}PC_t + \underset{(0.021)}{0.016}PB_t + \underset{(0.02)}{0.24}YD_t \\ t &= -2.13 \quad +0.74 \quad +13.2 \\ \bar{R}^2 &= 0.963 \quad N = 29 \quad \hat{\rho} = 0.56\end{aligned}\tag{10.25}$$

Let's compare equations (10.14) and (10.25). Note that the $\hat{\rho}$ used in equation (10.25) is 0.56. This means that *Y* was actually run as $Y^* = Y_t - 0.56Y_{t-1}$, *PC* as $PC^* = PC_t - 0.56PC_{t-1}$, etc. Second, $\hat{\rho}$ replaces *DW* in the documentation of GLS estimates in part because the *DW*

⁶² In Stata, the command to apply GLS (using Prais-Winsten) to Equation 9.14 thus would be *prais Y PC PB YD*. In Eviews, the easiest way to estimate a GLS equation is to add AR(1) to the equation as an independent variable, as in: *LS Y C PC PB YD AR(1)*. The result is a GLS estimate where *pn* will appear as the estimated coefficient of the variable AR(1).

of equation (10.25) isn't strictly comparable to non-GLS DW s (it is biased toward 2).

Generalized least squares estimates, no matter how they are produced, have at least two exercises. First, even though serial correlation causes no bias in the estimates of the $\hat{\beta}$ s, the GLS estimates usually are different from the OLS ones. For example, note that all three slope coefficients change as we move from OLS in equation (10.14) to GLS in equation (10.25). This isn't surprising, since different estimates can have different values even though their expected values are the same. The second problem is more important, however. It turns out that GLS works well if $\hat{\rho}$ is close to the actual ρ , but the GLS $\hat{\rho}$ is biased in small samples. If $\hat{\rho}$ is biased, then the biased $\hat{\rho}$ introduces bias into the GLS estimates of the $\hat{\beta}$ s. Luckily, there is a remedy for serial correlation that helps avoid both of these exercises: Newey - West standard errors.

Newey - West standard errors

Not all corrections for pure serial correlation involve generalized least squares. **Newey–West standard errors** are $SE(\hat{\beta})$ s that take account of serial correlation without changing the $\hat{\beta}$ s themselves in any way.⁶³ The logic behind Newey - West standard errors is powerful. If serial correlation does not cause bias in the $\hat{\beta}$ s but does impact the standard errors, then it makes sense to adjust the estimated equation in a way that changes the $SE(\hat{\beta})$ s but not the $\hat{\beta}$ s.

Thus Newey–West standard errors have been calculated specifically to avoid the consequences of pure first-order serial correlation. The Newey - West procedure yields an estimator of the standard errors that, while they are biased, is generally more accurate than uncorrected standard errors for large samples (greater than 100) in the face of serial correlation. As a result, Newey - West standard errors can be used for t -tests and other hypothesis tests in most samples without the errors of inference potentially caused by serial correlation. Typically, Newey - West $SE(\hat{\beta})$ s are larger than

⁶³ W. K. Newey and K. D. West, "A Simple, Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 1987, pp. 703–708. Newey - West standard errors are similar to HC standard errors (or White standard errors), to be discussed in Section 11.4.

OLS $SE(\hat{\beta})$ s, thus producing lower t -scores and decreasing the probability that a given estimated coefficient will be significantly different from zero. To see how Newey - West standard errors work, let's apply them to the same serially correlated chicken demand equation to which we applied GLS in equation (10.14). If we use Newey - West standard errors in the estimation of equation (10.14), we get:

$$\begin{aligned}\hat{Y}_t &= 27.7 - \underset{(0.03)}{0.11}PC_t + \underset{(0.02)}{0.03}PB_t + \underset{(0.01)}{0.23}YD_t \\ t &= -3.30 \quad + 2.12 \quad + 19.2 \\ \bar{R}^2 &= 0.9904 \quad N = 29\end{aligned}\tag{10.26}$$

Let's compare equations (10.14) and (10.26). First of all, the $\hat{\beta}$ s are identical in equations (10.14) and (10.26). This is because Newey - West standard errors do not change the OLS $\hat{\beta}$ s. Second, while we can't observe the change because of rounding, the Newey - West standard errors must be different from the OLS standard errors because the t -scores have changed even though the estimated coefficients are identical. However, two of the Newey - West $SE(\hat{\beta})$ s are slightly lower than the OLS $SE(\hat{\beta})$ s, which is a surprise even in a small sample like this one. Such a result indicates that there may well be an omitted variable or nonstationarity (to be discussed in chapter 13) in this equation.

SUMMARY

1. Serial correlation, or autocorrelation, is the violation of Classical assumption 4 that the observations of the error term are uncorrelated with each other. Usually, econometricians focus on first-order serial correlation, in which the current observation of the error term is assumed to be a function of the previous observation of the error term and a not-serially-correlated error term (ε):

$$u_t = \rho u_{t-1} + \varepsilon_t \quad -1 < \rho < 1$$

where ρ is “rho,” the autocorrelation coefficient.

2. Pure serial correlation is serial correlation that is a function of the error term of the correctly specified regression equation. Impure serial correlation is caused by specification errors such as an omitted variable or an incorrect functional form. While impure serial correlation can be positive ($0 < \rho < +1$) or negative ($-1 < \rho < 0$) pure serial correlation in economics or business situations is almost always positive (unless first differences are involved).
3. The major consequence of serial correlation is bias in the OLS $SE(\hat{\beta})$ s, causing unreliable hypothesis testing. Pure serial correlation does not cause bias in the estimates of the β s.
4. A commonly used method of detecting first-order serial correlation is the Durbin - Watson test, which uses the residuals of an estimated regression to test the possibility of serial correlation in the error term. An often-preferred alternative is the Lagrange multiplier (LM) test, which is far more general than the Durbin - Watson test.
5. The first step in ridding an equation of serial correlation is to check for possible specification errors. Only once the possibility of impure serial correlation has been reduced to a minimum should remedies for pure serial correlation be considered.
6. Generalized least squares (GLS) is a method of transforming an equation to rid it of pure first-order serial correlation. The use of GLS requires the estimation of ρ .
7. Newey - West standard errors are an alternative remedy for serial correlation that adjusts the OLS estimates of the $SE(\hat{\beta})$ s to take account of the serial correlation without changing the $\hat{\beta}$ s.

CHAPTER 11. HETEROSCEDASTICITY

Heteroskedasticity is the violation of Classical assumption 5, which states that the observations of the error term are drawn from a distribution that has a constant variance.⁶⁴ The assumption of constant variances for different observations of the error term (homoskedasticity) is not always realistic. For example, in a model explaining heights, it's likely that error term observations associated with the height of a basketball player would come from distributions with larger variances than those associated with the height of a mouse. Heteroskedasticity is important because OLS, when applied to heteroskedastic models, is no longer the minimum variance estimator (it still is unbiased, however).

In general, heteroskedasticity is more likely to take place in cross-sectional models than in time-series models. This focus on cross-sectional models is not to say that heteroskedasticity in time-series models is impossible, though. In fact, heteroskedasticity has turned out to be an important factor in timeseries studies of financial markets.

The structure of this chapter will be quite familiar. We'll attempt to answer the same four questions for heteroskedasticity that we answered for multicollinearity and serial correlation in the previous two chapters:

1. What is the nature of the problem?
2. What are the consequences of the problem?
3. How is the problem diagnosed?
4. What remedies for the problem are available?

⁶⁴ Various authors spell this “heteroscedasticity,” but Huston McCulloch appears to settle this controversy in favor of “heteroskedasticity” because of the word’s Greek origin. See J. Huston McCulloch, “On Heteros*edasticity,” *Econometrica*, Vol. 53, No. 2, p. 483. Although heteroskedasticity is a difficult word to spell, at least it’s an impressive comeback when parents ask, “What’d you learn for all that money?”

11.1 Pure versus impure heteroskedasticity

Heteroskedasticity, like serial correlation, can be divided into pure and impure versions. Pure heteroskedasticity is caused by the error term of the correctly specified equation; impure heteroskedasticity is caused by a specification error such as an omitted variable.

Pure heteroskedasticity

Pure heteroskedasticity refers to heteroskedasticity that is a function of the error term of a correctly specified regression equation. As with serial correlation, use of the word “heteroskedasticity” without any modifier (like pure or impure) implies *pure* heteroskedasticity.

Such **pure heteroskedasticity** occurs when Classical assumption 5, which assumes that the variance of the error term is constant, is violated in a correctly specified equation. assumption 5 assumes that:

$$\text{VAR}(u_i) = \sigma^2 = \text{a constant} \quad (i=1, 2, \dots, N) \quad (11.1)$$

If this assumption is met, all the observations of the error term can be thought of as being drawn from the same distribution: a distribution with a mean of zero and a variance of σ^2 . The property of having σ^2 not change for different observations of the error term is called **homoskedasticity**. A homoskedastic error term distribution is pictured in the top half of Figure 11.1; note that the variance of the distribution is constant (even though individual observations drawn from that sample will vary quite a bit).

With heteroskedasticity, this error term variance is not constant; instead, the variance of the distribution of the error term depends on exactly which observation is being discussed:

$$\text{VAR}(u_i) = \sigma_i^2 \quad (i=1, 2, \dots, N) \quad (11.2)$$

Note that the only difference between equations (11.1) and (11.2) is the subscript “i” attached to σ^2 , which implies that instead

of being constant over all the observations, a heteroskedastic error term's variance can change depending on the observation (hence the subscript).

Heteroskedasticity often occurs in data sets in which there is a wide disparity between the largest and smallest observed value of the dependent variable.

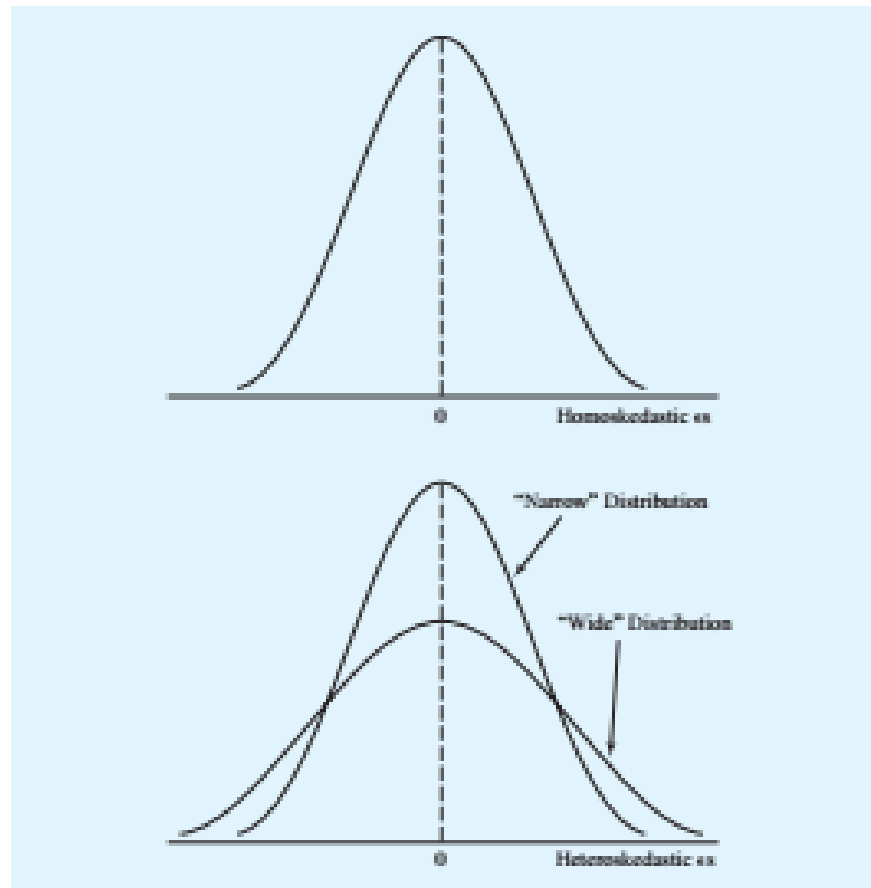


Figure 11.1 Homoskedasticity versus discrete heteroskedasticity

In homoskedasticity, the distribution of the error term has a constant variance, so the observations are continually drawn from the same distribution (shown in the top panel). In the simplest heteroskedastic case, discrete heteroskedasticity, there would be two different error term variances and, therefore, two different distributions (one wider than the other, as in the bottom panel) from which the observations of the error term could be drawn.

The larger the disparity between the size of observations of the dependent variable in a sample is, the larger the likelihood is that the error term observations associated with them will have different

variances and therefore be heteroskedastic. That is, we'd expect that the error term distribution for very large observations might have a large variance, and that the error term distribution for small observations might have a small variance.

In cross-sectional data sets, it's easy to get such a large range between the highest and lowest values of the variables. The difference between California and Rhode Island in terms of the dollar value of the consumption of goods and services, for instance, is quite large (comparable in percentage terms to the difference between the heights of a basketball player and a mouse). Since crosssectional models often include observations of widely different sizes in the same sample (cross-state studies of the United States usually include California and Rhode Island as individual observations, for example), heteroskedasticity is hard to avoid if economic topics are going to be studied cross sectionally.

The simplest way to visualize pure heteroskedasticity is to picture a world in which the observations of the error term could be grouped into just two different distributions, "wide" and "narrow." We'll call this simple version of the problem *discrete heteroskedasticity*. Here, both distributions would be centered around zero, but one would have a larger variance than the other, as indicated in the bottom half of Figure 11.1. Note the difference between the two halves of the Figure. With homoskedasticity, all the error term observations come from the same distribution; with heteroskedasticity, they come from different distributions.

For an example of discrete heteroskedasticity, we need go no further than our discussion of the heights of basketball players and mice. We'd certainly expect the variance of u to be larger for basketball players as a group than for mice, so the distribution of u for the heights of basketball players might look like the "wide" distribution in Figure 11.1, and the distribution of u for mice would be much narrower than the "narrow" distribution in Figure 11.1.

Heteroskedasticity takes on many more complex forms. In fact, the number of different models of heteroskedasticity is virtually limitless, and an analysis of even a small percentage of these alternatives would be a huge task. Instead, we'd like to address the general principles of heteroskedasticity by focusing on the most frequently specified model of pure heteroskedasticity, just as we

focused on pure, positive, first-order serial correlation in the previous chapter. However, don't let this focus mislead you into concluding that econometricians are concerned only with one kind of heteroskedasticity.

In this model of heteroskedasticity, the variance of the error term is related to an exogenous variable z_i . For a typical regression equation:

$$y_i = a + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \quad (11.3)$$

the variance of the otherwise classical error term u might be equal to:

$$\text{VAR}(u_i) = \sigma^2 z_i \quad (11.4)$$

where z may or may not be one of the x s in the equation. The variable z is called a **proportionality factor** because the variance of the error term changes proportionally to z_i . The higher the value of z_i , the higher the variance of the distribution of the i th observation of the error term. There would be N -different distributions, one for each observation, from which the observations of the error term could be drawn depending on the number of different values that z takes. To see what homoskedastic and heteroskedastic distributions of the error term look like with respect to z , compare Figures 11.2 and 11.3. Note that the heteroskedastic distribution gets wider as z increases but that the homoskedastic distribution maintains the same width no matter what value z takes.

What is an example of a proportionality factor z ? How is it possible for an exogenous variable such as z to change the whole distribution of an error term? Think about a function that relates the consumption expenditures in a state to its income. The expenditures of a small state like Rhode Island are not likely to be as variable in absolute value as the expenditures of a large state like California because a 10% change in spending for a large state involves a lot more money than a 10% change for a small one. In such a case, the dependent variable would be consumption expenditures and a likely

proportionality factor, z , would be population. As population rose, so too would the variance of the error term of an equation built to explain expenditures. The error term distributions would look something like those in Figure (11.3), where the z in Figure (11.3) is population.

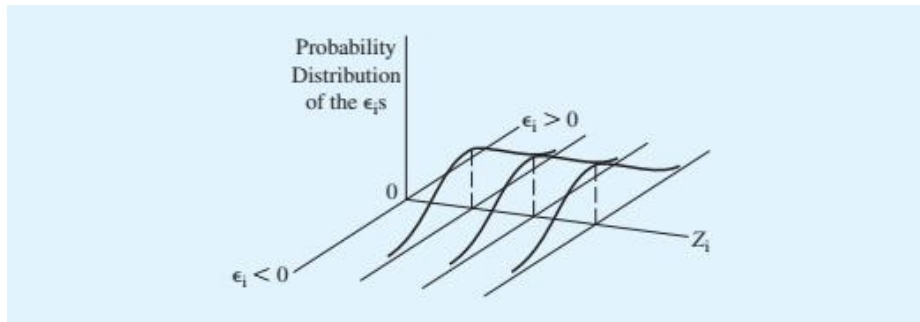


Figure 11.2 A homoskedastic error term with respect to z_i

If an error term is homoskedastic with respect to z_i , the variance of the distribution of the error term is the same (constant) no matter what the value of z_i is: $VAR(u_i) = \sigma^2$

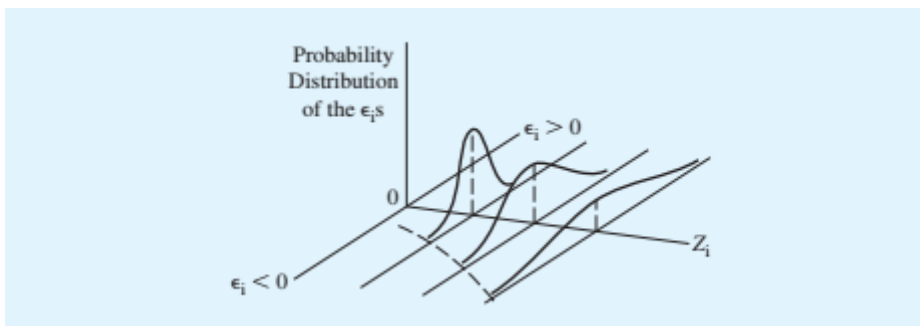


Figure 11.3 A heteroskedastic error term with respect to z_i

If an error term is heteroskedastic with respect to z_i , the variance of the distribution of the error term changes systematically as a function of z_i . In this example, the variance is an increasing function of z_i , as in $VAR(u_i) = \sigma^2 z_i$.

This example helps emphasize that heteroskedasticity is likely to occur in cross-sectional models because of the large variation in the size of the dependent variable involved. An exogenous disturbance that might seem huge to a small state could seem miniscule to a large one, for instance.

Heteroskedasticity can occur in a time-series model with a significant amount of change in the dependent variable. If you were modeling sales of DVD players from 1994 to 2015, it's quite possible that you would have a heteroskedastic error term. As the phenomenal growth of the industry took place, the variance of the error term probably increased as well. Such a possibility is unlikely in time series that have low rates of change, however.

Heteroskedasticity also can occur in any model, time series or cross sectional, where the quality of data collection changes dramatically within the sample. As data collection techniques get better, the variance of the error term should fall because measurement errors are included in the error term. As measurement errors decrease in size, so should the variance of the error term. For more on this topic (called "errors in the variables"), see section 15.6.

Impure heteroskedasticity

Heteroskedasticity that is caused by an error in specification, such as an omitted variable, is referred to as **impure heteroskedasticity**. Impure heteroskedasticity thus is similar to impure serial correlation.

An omitted variable can cause a heteroskedastic error term because the portion of the omitted effect not represented by one of the included explanatory variables must be absorbed by the error term. If this effect has a heteroskedastic component, the error term of the misspecified equation might be heteroskedastic even if the error term of the true equation is not. This distinction is important because with impure heteroskedasticity the correct remedy is to find the omitted variable and include it in the regression. It's therefore important to be sure that your specification is correct before trying to detect or remedy pure heteroskedasticity.

11.2 The consequences of heteroskedasticity

If the error term of your equation is known to be heteroskedastic, what does that mean for the estimation of your

coefficients? If the error term of an equation is heteroskedastic, there are three major consequences:⁶⁵

1. *Pure heteroskedasticity does not cause bias in the coefficient estimates.* Even if the error term of an equation is known to be purely heteroskedastic, that heteroskedasticity will not cause bias in the OLS estimates of the coefficients. This is true because even though large positive errors are more likely, so too are large negative errors. The two tend to average each other out, leaving the OLS estimator still unbiased. As a result, we can say that an otherwise correctly specified equation that has pure heteroskedasticity still has the property that:

$$E(\hat{\beta}) = \beta \quad \text{for all } \beta\text{'s}$$

Lack of bias does not guarantee “accurate” coefficient estimates, especially since heteroskedasticity increases the variance of the estimates, but the distribution of the estimates is still centered around the true β . Equations with impure heteroskedasticity caused by an omitted variable, of course, will have possible specification bias.

2. *Heteroskedasticity typically causes OLS to no longer be the minimum variance estimator (of all the linear unbiased estimators).* Pure heteroskedasticity causes no bias in the estimates of the OLS coefficients, but it does affect the minimum-variance property. If the error term of an equation is heteroskedastic with respect to a proportionality factor z :

$$\text{VAR}(u_i) = \sigma^2 z_i \tag{11.5}$$

then the minimum-variance portion of the Gauss - Markov theorem cannot be proven because there are other linear unbiased estimators that have smaller variances. This is because the heteroskedastic error term causes the dependent

⁶⁵ It turns out that the consequences of heteroskedasticity are almost identical in general framework to those of serial correlation, though the two problems are quite different.

variable to fluctuate, and the OLS estimation procedure attributes this fluctuation to the independent variables. Thus, OLS is more likely to misestimate the true β in the face of heteroskedasticity. The $\hat{\beta}$ s still are unbiased because overestimates are just as likely as underestimates.

3. *Heteroskedasticity causes the OLS estimates of the $SE(\hat{\beta})$ s to be biased, leading to unreliable hypothesis testing and confidence intervals.* With heteroskedasticity, the OLS formula for the standard error produces biased estimates of the $SE(\hat{\beta})$ s. Because the $SE(\hat{\beta})$ is a prime component in the t -statistic, these biased $SE(\hat{\beta})$ s cause biased t -scores and unreliable hypothesis testing in general. In essence, heteroskedasticity causes OLS to produce incorrect $SE(\hat{\beta})$ s and t -scores! Not surprisingly, most econometricians therefore are very hesitant to put much faith in hypothesis tests that were conducted in the face of pure heteroskedasticity.⁶⁶

What sort of bias in the standard errors does heteroskedasticity tend to cause? Typically, heteroskedasticity causes OLS estimates of the standard errors to be biased downward, making them too small. Sometimes, however, they're biased upward; it's hard to predict in any given case. But either way, it's a big problem for hypothesis testing and confidence intervals. What'll happen if OLS underestimates a standard error? Well, the "too low" $SE(\hat{\beta})$ will cause a "too high" t -score for a particular coefficient, and this will make it more likely that we will reject a null hypothesis (for example, $H_0: \beta \leq 0$) when it is in fact true. This increased chance of rejecting H_0 means that we're more likely to make a type I error and we're more likely to make the mistake of keeping an irrelevant variable in an equation. Also, because the confidence interval depends directly on $SE(\hat{\beta})$, the underestimation of $SE(\hat{\beta})$ will fool us into thinking that our estimate is more precise than it really is.⁶⁷

⁶⁶ While our discussion here involves the t -test, the same conclusion of unreliability in the face of heteroskedasticity applies to all other test statistics, including confidence intervals.

⁶⁷ If OLS overestimates the standard error, then we'll have the same problems but in reverse. The "too high" $SE(\hat{\beta})$ will lead to a "too low" t -score. If the t -score is lowered

In other words, pure heteroskedasticity can make quite a mess of our results. Hypothesis testing will become unreliable, and confidence intervals will be misleading.

11.3 Testing for heteroskedasticity

As we've seen, heteroskedasticity is a potentially nasty problem. The good news is that there are many tests for heteroskedasticity. The bad news is heteroskedasticity can take many different forms and no single test can find them all.

In this section, we'll describe two of the most popular and powerful tests for heteroskedasticity, the *Breusch - Pagan test* and the *White test*.⁶⁸ While neither test can "prove" that heteroskedasticity exists, these tests often can give us a pretty good idea of whether or not it's a problem.

Before using any test for heteroskedasticity, it's a good idea to start with the following preliminary questions:

1. Are there any obvious specification errors? Are there any likely omitted variables? Have you specified a linear model when a double-log model is more appropriate? Don't test for heteroskedasticity until the specification is as good as possible. After all, if you find heteroskedasticity in an incorrectly specified model, there's a chance it will be impure.
2. Are there any early warning signs of heteroskedasticity? Just as certain kinds of clouds can warn of potential storms, certain kinds of data can signal possible heteroskedasticity. In particular, if the dependent variable's maximum value is many, many times larger than its minimum, beware of heteroskedasticity.
3. Does a graph of the residuals show any evidence of heteroskedasticity? It sometimes saves time to plot the residuals against a potential Z proportionality factor or

enough, we might be fooled into failing to reject a false null hypothesis, thus increasing the risk that we'll drop a relevant variable from the model. In addition, the confidence intervals will be too wide, leading to similar potential mistakes.

⁶⁸ Both tests belong to a general group of tests based on the Lagrange multiplier (LM), which you first met in chapter 10.

against the dependent variable. If you see a pattern in the residuals, you've got a problem. See Figure 11.4 for a few examples of heteroskedastic patterns in the residuals.

Note that Figure 11.4 shows “textbook” examples of heteroskedasticity. The real world is nearly always a lot messier than textbook graphs. It's not unusual to look at a real-world residual plot and be unsure whether there's a pattern or not. As a result, even if there are no obvious specification errors, no early warning signs, and no visible residual patterns, it's a good idea to do a formal statistical test for heteroskedasticity, so we'd better get started.

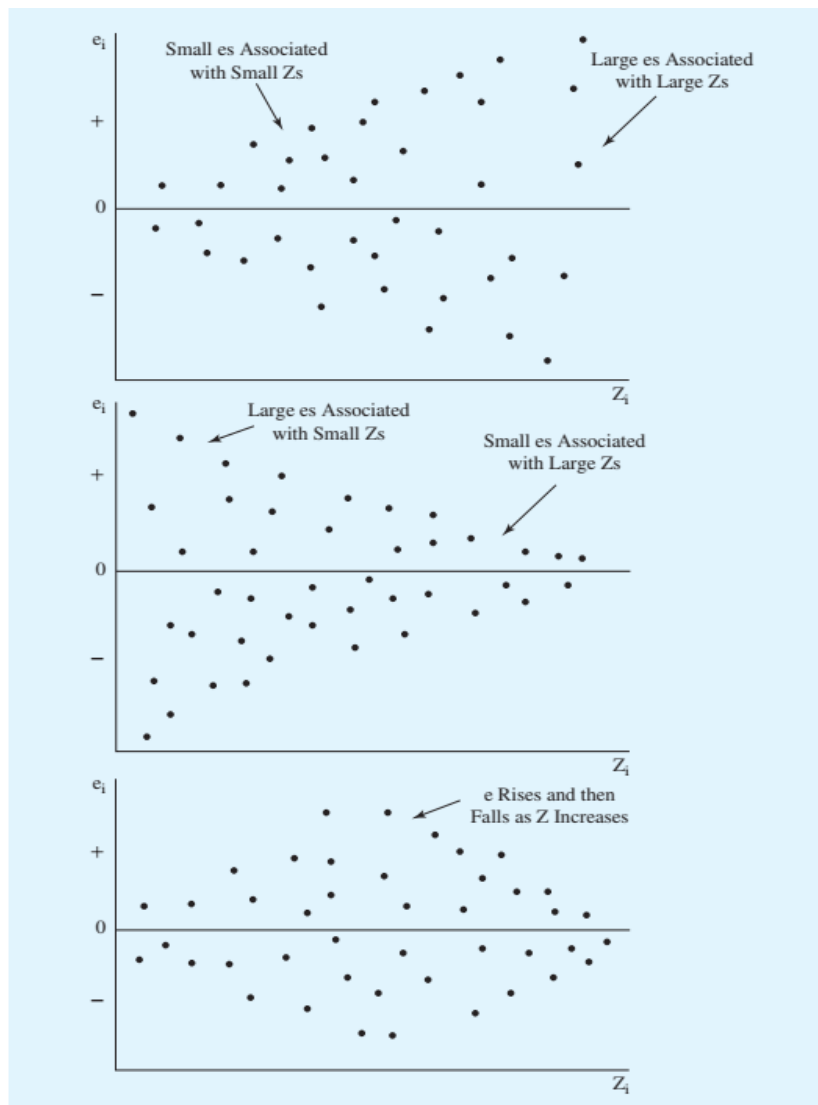


Figure 11.4 Eyeballing residuals for possible heteroskedasticity

If you plot the residuals of an equation with respect to a potential Z proportionality factor, a pattern in the residuals is an indication of possible heteroskedasticity.

The Breusch - Pagan test

The **Breusch–Pagan test** is a method of testing for heteroskedasticity in the error term by investigating whether the squared residuals can be explained by possible proportionality factors.⁶⁹ Here's how it works.

1. *Obtain the residuals from the estimated regression equation.*
For an equation with two independent variables, this would be:

$$u_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} \quad (11.6)$$

2. *Use the squared residuals as the dependent variable in an auxiliary equation.* As the explanatory variables in the auxiliary regression, use right-hand variables from the original regression that you suspect might be proportionality factors. For many researchers, the default option is to include all of them. For instance, if the original equation has two explanatory variables, then the auxiliary regression would be:

$$u_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_i \quad (11.7)$$

3. *Test the overall significance of equation (11.7) with a chi-square test.* The null and alternative hypotheses are:

$$\begin{aligned} H_0: & \alpha_1 = \alpha_2 = 0 \\ H_1: & H_0 \text{ is false} \end{aligned}$$

The null hypothesis is homoskedasticity, because if $\alpha_1 = \alpha_2 = 0$, then the variance equals α_0 , which is a constant. The test

⁶⁹ T. S. Breusch and A. R. Pagan, "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, Vol. 47, pp. 1287–1294.

statistic here is NR^2 , or the sample size (N) times the unadjusted R^2 from equation (11.7). This test statistic has a chi-square distribution⁷⁰ with degrees of freedom equal to the number of slope coefficients in the auxiliary regression (equation 11.7). If NR^2 is greater than or equal to the critical chi-square value, then we reject the null hypothesis of homoskedasticity.

If you strongly suspect that only certain variables are plausible z factors, then you should run the Breusch - Pagan test using only an intercept and the suspect variables. The degrees of freedom for the chi-square statistic of course would change in such a situation, because they're equal to the number of right-hand-side variables in the auxiliary equation. If you're certain you know the one and only proportionality factor z and that there are no other forms of heteroskedasticity present, you don't even need to fool with the chi-square statistic. You can just do a two-sided t -test⁸ on the α_n for z .

The strengths of the Breusch-Pagan test are that it's easy to use and it's powerful if heteroskedasticity is related to one or more linear proportionality factors. Its weakness is that if it fails to find heteroskedasticity, it only means there is no evidence of heteroskedasticity related to the z s you've chosen. If you're pretty certain that the x s in the auxiliary regression are the only plausible proportionality factors, you can rest easy. But if you're not certain, you might want to use the White test, which we'll discuss shortly.

As an example of the use of the Breusch - Pagan test, let's return to the Rohat's restaurants example of section 4.2 and use the residuals of equation (4.4) to test for heteroskedasticity. Recall that the regression explained the number of customers, as measured by the check volume (Y) for a cross section of 33 different Rohat's restaurants as a function of the number of nearby competitors (N), the nearby population (P), and the average household income of the local area (I):

⁷⁰ You might wonder why the test statistic is not the overall F -statistic of the auxiliary regression. It turns out that the F -statistic is valid only if the errors are normally distributed, and in this case, with squared residuals as the dependent variable, it's not safe to assume the errors are normally distributed. With non-normal errors, the proper test is a chi-square test.

$$\begin{aligned} \hat{Y}_i &= 102.192 + 9075N_i + 0.355P_i + 1.288I_i \\ &\quad \begin{matrix} (2053) & (0.073) & (0.543) \end{matrix} \\ t &= -4.42 \quad 4.88 \quad 2.37 \\ N &= 33 \quad \bar{R}^2 = 0.579 \end{aligned} \tag{4.4}$$

The first step in the Breusch–Pagan test is to obtain the residuals from equation (4.4). The second step is to square the residuals and use them as the dependent variable in an auxiliary regression. If we include all the independent variables in equation (4.4) in our auxiliary equation, we get:

$$u_i^2 = \alpha_0 + \alpha_1 N_i + \alpha_2 P_i + \alpha_3 I_i + u_i \tag{11.8}$$

If we estimate Equation (11.8), we find that the unadjusted $R^2 = .0441$. We know that $N = 33$, so we can calculate that the chi-square statistic $= NR^2 = 331.04412 = 1.455$. Since the 5% critical value of chi-square with 3 degrees of freedom is 7.81, we can't reject the null hypothesis that $\alpha_1 = \alpha_2 = \alpha_3 = 0$. As a result, the Breusch - Pagan test doesn't provide any evidence that equation (4.4) suffers from heteroskedasticity. This makes sense. Even though the Rohat's sample is cross-sectional, the largest value of the dependent variable isn't even twice the size of the smallest one, so we have no reason to suspect pure heteroskedasticity.

The White test

Probably the most popular of all the heteroskedasticity tests is the White test⁷¹ because it can find more types of heteroskedasticity than any other test. That's a distinct advantage in a world where just about any variable or combination of variables, linear or nonlinear, could trip us up with a heteroskedastic stumbling block. Let's see how it works.

The **White test** investigates the possibility of heteroskedasticity in an equation by seeing if the squared residuals can be explained

⁷¹ Halbert White, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrica*, Vol. 48, pp. 817–838.

by the equation's independent variables, their squares, and their cross-products. To run the White test:

1. *Obtain the residuals of the estimated regression equation.*
2. *Estimate an auxiliary regression, using the squared residuals as the dependent variable, with each X from the original equation, the square of each x , and the product of each x times every other x as the explanatory variables.* For example, if the original equation's independent variables are x_1 and x_2 , the White test equation is:

$$u_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{1i}^2 + \alpha_4 x_{2i}^2 + \alpha_5 x_{1i} x_{2i} + u_i \quad (11.9)$$

3. *Test the overall significance of equation (11.9) with a chi-square test.* Once again, the test statistic is NR^2 , the sample size (N) times the unadjusted R^2 of equation (11.9). This test statistic has a chi-square distribution with degrees of freedom equal to the number of slope coefficients in the auxiliary regression. The null hypothesis is that all the slope coefficients in equation (11.9) equal zero, and if NR^2 is greater than the chi-square critical value, then we can reject the null hypothesis and conclude that there's evidence of heteroskedasticity.

Check out the explanatory variables in equation (11.9). They include every variable in the original model, their squares, and their cross products. Including all the variables from the original model allows the White test to check to see if any or all of them are z proportionality factors. Including all the squared terms and cross products allows us to test for more exotic and complex types of heteroskedasticity. This is the White test's greatest strength.

However, the White test contains more right-hand-side variables than the original regression, sometimes a *lot* more. This can be its greatest weakness. To see why, note that as the number of explanatory variables in an original regression rises, the number of right-hand variables in the White test auxiliary regression goes up much faster. For example, there are five right-hand variables in equation (11.9) even though the original model had only two, x_1

and x_2 . With three variables in the original model, the White regression could have nine. With 12 explanatory variables in the original model, there could⁷² be 90 in the White regression with all the squares and interactive terms included!

And this is where the weakness becomes a real problem. If the number of right-hand variables in the auxiliary regression exceeds the number of observations, you can't run the White test regression because you would have negative degrees of freedom in the auxiliary equation! Even if the degrees of freedom in the auxiliary equation are positive but small, the White test might do a poor job of detecting heteroskedasticity because the fewer the degrees of freedom there are, the less powerful the statistical test is. In such a situation, you'd be limited to the Breusch–Pagan test or an alternative.⁷³ As an example of the White test, let's again return to the Rohat's restaurants model of section 4.2. As with the Breusch - Pagan test, the first step is to obtain the residuals of the original Rohat's equation. The second step in the White test is to square the residuals and use them as the dependent variable in an auxiliary regression that includes N , P , I , their squares, and their cross products as independent variables:

$$u_i^2 = \alpha_0 + \alpha_1 N_i + \alpha_2 P_i + \alpha_3 I_i + \alpha_4 N_i^2 + \alpha_5 P_i^2 + \alpha_6 I_i^2 \\ + \alpha_7 N_i P_i + \alpha_8 N_i I_i + \alpha_9 P_i I_i + u_i$$

⁷² There could be fewer explanatory variables if one or more of the original independent variables is a dummy, since the square of a dummy is the same as a dummy and the cross product of a variable times a dummy equals the original variable itself or zero. Because of the large number of variables and this possible duplication, it's tedious to create and check all the variables for the White test. Luckily, Eviews and most other econometric software packages have simple commands to do the work for you.

⁷³ For instance, there is an alternative form of the White test that has many fewer degrees of freedom in the auxiliary equation. You perform this alternative White test by replacing all the right-hand variables in equation (11.9) with the fitted Y values and the squares of the fitted Y values from the original model. That is, run: $u_i^2 = \alpha_0 + \alpha_1 \hat{Y}_i + \alpha_2 \hat{Y}_i^2 + u_i$ with the null hypothesis being $\alpha_1 = \alpha_2 = 0$ and the rest of the test being the same. This isn't as good as the full White test, but if the full White test is impossible to run, this may be an excellent alternative. For more, see Christopher F. Baum, *An Introduction to Modern Econometrics Using Stata* (College Station, TX: Stata Press, 2006), p. 146.

If we estimate this equation with the Woody's data, we find that the unadjusted $R^2 = .1218$. Since $N = 33$, the chi-square $NR^2 = 33 \cdot .12182 = 4.02$. That's less than 16.92, the 5% critical chi-square value with nine degrees of freedom (do you see why it's nine?), so we once again can't reject the null hypothesis of homoskedasticity.

11.4 Remedies for heteroskedasticity

The first thing to do if the Breusch - Pagan test or the White test indicates the possibility of heteroskedasticity is to examine the equation carefully for specification errors. Although you should never include an explanatory variable simply because a test indicates the possibility of heteroskedasticity, you ought to rigorously think through the specification of the equation. If this rethinking allows you to discover a variable that should have been in the regression from the beginning, then that variable should be added to the equation. Similarly, if you had the wrong functional form to begin with, the discovery of heteroskedasticity might be the hint you need to rethink the specification and switch to the functional form that best represents the underlying theory. However, if there are no obvious specification errors, the heteroskedasticity is probably pure in nature, and one of the remedies described in this section should be considered.

Heteroskedasticity - corrected standard errors

The most popular remedy for heteroskedasticity is heteroskedasticity - corrected standard errors, which adjust the estimation of the $SE(\hat{\beta})$ s for heteroskedasticity while still using the OLS estimates of the slope coefficients.

The logic behind this approach is powerful. Since heteroskedasticity causes exercises with the $SE(\hat{\beta})$ s but not with the $\hat{\beta}$ s, it makes sense to improve the estimation of the $SE(\hat{\beta})$ s in a way that doesn't alter the estimates of the slope coefficients. This approach is virtually identical to the use of Newey - West standard errors as a remedy for serial correlation.

Thus, **heteroskedasticity-corrected (HC) standard errors** are $SE(\hat{\beta})$ s that have been calculated specifically to avoid the

consequences of heteroskedasticity. The HC procedure yields an estimator of the standard errors that, while they are biased, are generally more accurate than uncorrected standard errors for large samples in the face of heteroskedasticity. As a result, the HC $SE(\hat{\beta})$ s can be used in t -tests and other hypothesis tests in most samples without the errors of inference potentially caused by heteroskedasticity. Typically, the HC $SE(\hat{\beta})$ s are larger than the OLS $SE(\hat{\beta})$ s, thus producing lower t -scores and decreasing the probability that a given estimated coefficient will be significantly different from zero. The technique was suggested by Halbert White in the same article in which he proposed the White test for heteroskedasticity.⁷⁴

There are a few exercises with using heteroskedasticity-corrected standard errors. First, the technique works best in large samples, so it's best to avoid HC $SE(\hat{\beta})$ s in small samples. Second, details of the calculation of the HC $SE(\hat{\beta})$ s are beyond the scope of this text and imply a model that is substantially more general than the basic theoretical construct, $VAR(u_i) = \sigma^2 z_i$, of this chapter. In addition, not all computer regression software packages calculate heteroskedasticity-corrected standard errors.

Redefining the variables

Another approach to ridding an equation of heteroskedasticity is to go back to the basic underlying theory of the equation and redefine the variables in a way that avoids heteroskedasticity. A redefinition of the variables often is useful in allowing the estimated equation to focus more on the behavioral aspects of the relationship. Such a rethinking is a difficult and discouraging process because it appears to dismiss all the work already done. However, once the theoretical work has been reviewed, the alternative approaches that are discovered are often exciting in that they offer possible ways to avoid exercises that had previously seemed insurmountable. Be careful, however. Redefining your variables is a functional form specification change that can dramatically change your equation.

⁷⁴ Note that Newey–West standard errors, introduced in section 10.4, also can be used as HC standard errors. Indeed, some econometric software packages provide a choice between the White and Newey - West procedures. Unless otherwise noted, however, HC standard errors should be assumed to be of the White variety. Most authors refer to this method as HCCM, for heteroskedasticity-consistent covariance matrix.

In some cases, the only redefinition that's needed to rid an equation of heteroskedasticity is to switch from a linear functional form to a double-log functional form. The double-log form has inherently less variation than the linear form, so it's less likely to encounter heteroskedasticity. In addition, there are many research topics for which the double-log form is just as theoretically logical as the linear form. This is especially true if the linear form was chosen by default, as is often the case.

In other situations, it might be necessary to completely rethink the research project in terms of its underlying theory. For example, consider a cross-sectional model of the total expenditures by the governments of different cities. Logical explanatory variables to consider in such an analysis are the aggregate income, the population, and the average wage in each city. The larger the total income of a city's residents and businesses, for example, the larger the city government's expenditures (see Figure 11.5). In this case, it's not very enlightening to know that the larger cities have larger incomes and larger expenditures (in absolute magnitude) than the smaller ones.

Fitting a regression line to such data (see the line in Figure 11.5) also gives undue weight to the larger cities because they would otherwise give rise to large squared residuals. That is, since OLS minimizes the summed squared residuals, and since the residuals from the large cities are likely to be large due simply to the size of the city, the regression estimation will be especially sensitive to the residuals from the larger cities. This is often called "spurious correlation" due to size.

In addition, the residuals may indicate heteroskedasticity. It makes sense to consider reformulating the model in a way that will discount the scale factor (the size of the cities) and emphasize the underlying behavior. In this case, per capita expenditures would be a logical dependent variable. Such a transformation is shown in Figure 11.6. This form of the equation places New York and Los Angeles on the same scale as, say, Pasadena or New Brunswick, and thus gives them the same weight in estimation. If an explanatory variable happened not to be a function of the size of the city, however, it would not need to be adjusted to per capita terms. If the equation included the average wage of city workers, for

example, that wage would not be divided through by population in the transformed equation.

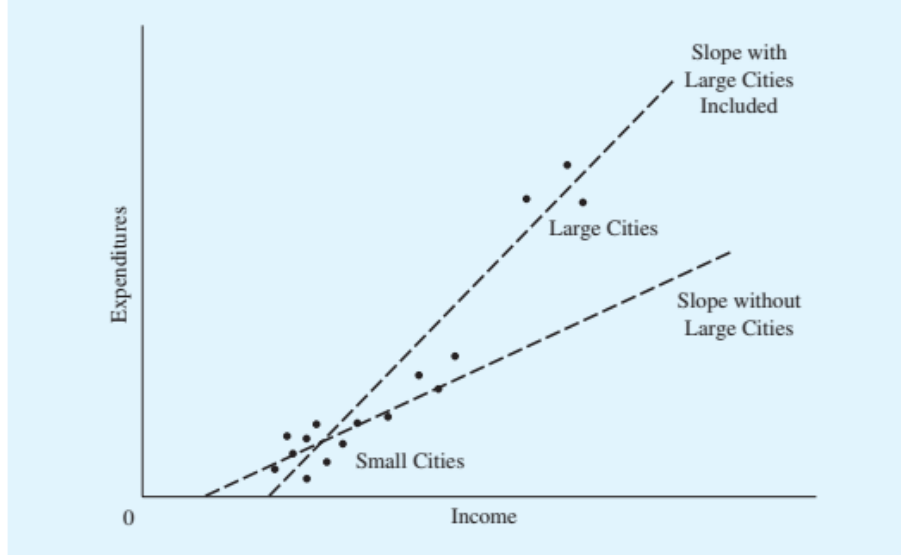


Figure 11.5 An aggregate city expenditures function

If city expenditures are explained in an aggregate model, the larger cities play a major role in the determination of the coefficient values. Note how the slope would be somewhat lower without the heavy influence of the larger cities. In addition, heteroskedasticity is a potential problem in an aggregate model because the wide range of sizes of the dependent variable makes different error term variances more likely.

Suppose your original equation is:

$$EXP_i = a + \beta_1 POP_i + \beta_2 INC_i + \beta_3 WAGE_i + u_i \quad (11.10)$$

where EXP_i refers to expenditures, INC_i refers to income, $WAGE_i$ refers to the average wage, and POP_i refers to the population of the i th city. The transformed equation would be⁷⁵:

⁷⁵ This transformed equation is very similar to the equation for Weighted Least Squares (WLS). *Weighted Least Squares* is a remedy for heteroskedasticity that consists of dividing the entire equation (including the constant and the heteroskedastic error term) by the proportionality factor Z and then re-estimating the equation with OLS. For the example in this section, the WLS equation would be:

$$EXP_i / POP_i = a / POP_i + \beta_1 + \beta_2 INC_i / POP_i + \beta_3 WAGE_i / POP_i + u_i \quad (11.11)$$

where the variables and β s in equation (11.11) are identical to those in equation (11.10). Dividing through by Z means that u is a homoskedastic error term as long as Z is the correct proportionality factor. This is not a trivial problem, however, and other

$$EXP_i / POP_i = \alpha_0 + \alpha_1 INC_i / POP_i + \alpha_2 WAGE_i + u_i \quad (11.12)$$

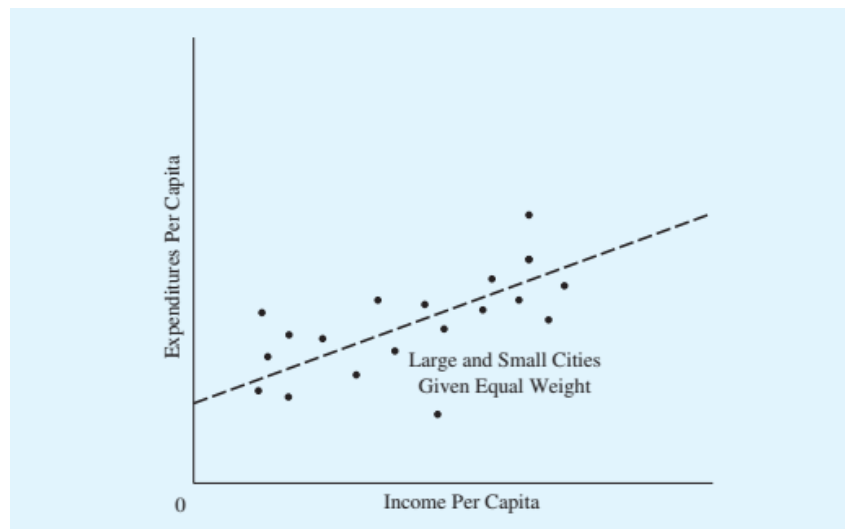


Figure 11.6 A per capita city expenditures function

If city expenditures are explained in a per capita model, then large and small cities have equal weights. In addition, heteroskedasticity is less likely, because the dependent variable does not vary over a wide range of sizes.

where u_i is a classical homoskedastic error term. While the directly transformed equation (11.12) probably avoids heteroskedasticity, such a solution should be considered incidental to the benefits of rethinking the equation in a way that focuses on the basic behavior being examined.

Note that it's possible that the *reformulated* equation (11.12) could have heteroskedasticity; the error variances might be larger for the observations having the larger per capita values for expenditures than they are for smaller per capita values. Thus, it is legitimate to suspect and test for heteroskedasticity even in this transformed model. Such heteroskedasticity in the transformed equation is unlikely, however, because there will be little of the variation in size normally associated with heteroskedasticity.

transformations and HCSEs are much easier to use than WLS is, so we no longer recommend the use of WLS.

11.5 A more complete example

Let's work through a more complete example involving heteroskedasticity, a cross-sectional model of petroleum consumption by state.

Possible explanatory variables include functions of the size of the state (such as the number of miles of roadway, the number of motor vehicle registrations, or the population) and variables that are *not* functions of the size of the state (such as the price of gasoline or the speed limit). Since there is little to be gained by including more than one variable that measures the size of the state (because such an addition would be theoretically redundant and likely to cause needless multicollinearity), and since the speed limit was the same for all states (it would be a useful variable in a time-series model, however) a reasonable model to consider might be:

$$PCON_i = a + \beta_1^+ REG_i + \beta_2^- PRICE_i + u_i \quad (11.13)$$

where: $PCON_i$ = petroleum consumption in the i th state
(trillions of BTUs)

REG_i = motor vehicle registrations in the i th state
(thousands)

$PRICE_i$ = the price of gasoline in the i th state (cents per
gallon)

u_i = a classical error term

The more cars there are registered in a state, we would think, the more petroleum would be consumed, while a high gasoline price would decrease aggregate gasoline purchases in that state.⁷⁶ If we now collect data from 2005 for this example (see Table 11.1) we can estimate:

$$\begin{aligned} PC\hat{O}N_i &= 4101 + 0.16 REG_i - 1885 PRICE_i \\ &\quad \begin{matrix} (0.01) & (750) \\ t = 12.4 & -2.51 \\ N = 50 & \bar{R}^2 = 0.76 \end{matrix} \end{aligned} \quad (11.14)$$

⁷⁶ An alternative to using $PRICE$ as an independent variable in this equation is to use $PRICE*REG$ or $PRICE*POP$ (where POP_i is the population of the i th state). These are more sophisticated examples of the interaction terms we introduced in section 8.4 in our discussion of slope dummy variables.

This equation seems to have no exercises; the coefficients are significant in the hypothesized directions, and the overall equation is statistically significant. No Durbin - Watson statistic is shown because there is no “natural” order of the observations to test for serial correlation (if you’re curious, the DW for the order in Table 11.1 is 2.15).

Given the discussion in the previous sections, let’s investigate the possibility of heteroskedasticity caused by variation in the size of the states. To test this possibility, we obtain the residuals from equation (11.14), and run Breusch - Pagan and White tests on them.

Table 11.1 Data for the petroleum consumption example

PCON	PRICE	REG	POP	STATE
580	2.11	4545	4548	Alabama
284	2.13	673	663	Alaska
537	2.23	3972	5953	Arizona
377	2.10	1940	2776	Arkansas
3837	2.47	32487	36154	California
463	2.19	1808	4663	Colorado
463	2.17	3059	3501	Connecticut
148	2.07	737	842	Delaware
1940	2.21	15691	17768	Florida
1058	2.09	8063	9133	Georgia
270	2.47	948	1273	Hawaii
139	2.14	1374	1429	Idaho
1313	2.22	9458	12765	Illinois
901	2.19	4955	6266	Indiana
393	2.13	3398	2966	Iowa
434	2.17	2368	2748	Kansas
664	2.14	3428	4173	Kentucky
1610	2.10	3819	4507	Louisiana
262	2.16	1075	1318	Maine
561	2.15	4322	5590	Maryland
734	2.08	5420	6433	Massachusetts
1010	2.24	8247	10101	Michigan
694	2.11	4647	5127	Minnesota
484	2.11	1978	2908	Mississippi
737	2.09	4589	5798	Missouri
161	2.17	1009	935	Montana
231	2.21	1703	1758	Nebraska
242	2.38	1349	2412	Nevada
198	2.08	1174	1307	New Hampshire
1233	1.99	6262	8703	New Jersey
250	2.19	1548	1926	New Mexico
1776	2.23	11863	19316	New York
947	2.14	6148	8672	North Carolina
121	2.19	695	635	North Dakota
1340	2.19	10634	11471	Ohio
545	2.08	3725	3543	Oklahoma
370	2.28	2897	3639	Oregon
1466	2.14	9864	12405	Pennsylvania
102	2.12	812	1074	Rhode Island
517	2.06	3339	4247	South Carolina

PCON	PRICE	REG	POP	STATE
113	2.20	854	775	South Dakota
782	2.11	4980	5956	Tennessee
5628	2.07	17470	22929	Texas
276	2.12	2210	2490	Utah
86	2.13	508	622	Vermont
965	2.10	6591	7564	Virginia
793	2.28	5598	6292	Washington
255	2.20	1352	1814	West Virginia
597	2.26	4725	5528	Wisconsin
162	2.08	646	509	Wyoming

Source: 2008 Statistical Abstract (U.S. Department of Commerce).

Datafile = GAS10

Before we can run a Breusch - Pagan test, we must decide which variables to include on the right-hand side of the auxiliary equation. *REG* (motor vehicle registrations) is a measure of market size, so it's an obvious proportionality factor. On the other hand, *PRICE* (gasoline prices) seems unlikely to vary significantly with the size of the state, so it's less likely to be a *z*. However, many researchers automatically include all the independent variables from the original equation in the Breusch - Pagan auxiliary equation, so let's use that approach and estimate the auxiliary equation (equation (11.7)) with both *REG* and *PRICE*:

$$\begin{aligned}
 u_i^2 &= 5164290 + 83.33 \underset{(25.1)}{REG_i} - 2475027 \underset{(1476765)}{PRICE_i} \\
 t &= 3.32 \qquad \qquad -1.68 \\
 N &= 50 \qquad \qquad R^2 = 0.197
 \end{aligned}
 \tag{11.15}$$

The Breusch - Pagan test specifies that we should reject the null hypothesis of homoskedasticity ($\alpha_1 = \alpha_2 = 0$) if $NR^2 >$ the appropriate critical chi-square value. Since $N = 50$ and $R^2 = .197$, $NR^2 = 50(0.197) = 9.85$. If you take a look at Table B-6, you'll see that the 5% critical chi-square value equals 5.99, so since $9.85 > 5.99$, we can reject the null hypothesis of homoskedasticity.⁷⁷ We have heteroskedasticity!

⁷⁷ If we run the Breusch - Pagan test with *REG* as the only proportionality factor, we also can reject the null hypothesis of homoskedasticity at the 5-percent level because the *t*-score of 2.90 is greater than 2.01, the 5-percent two-sided critical *t*-core with 48 degrees of

To see whether the White test also will detect this heteroskedasticity, we'll need to modify equation (11.9) to fit our example. The dependent variable will be the square of the residuals in equation (11.14), and the independent variables will be REG , $PRICE$, their squares, and their cross-product:

$$u_i^2 = \alpha_0 + \alpha_1 REG_i + \alpha_2 PRICE_i + \alpha_3 REG_i^2 + \alpha_4 PRICE_i^2 + \alpha_5 REG_i \times PRICE_i + u_i \quad (11.16)$$

If we estimate equation (11.16) with the residuals from equation (11.14) and the data from Table 11.1, we obtain an R^2 of 0.85.

Since $N = 50$, $NR^2 = 50 \times 0.85 = 42.5$. As you can confirm by looking at Table B-6, the critical chi-square value at the 5% level for 5 degrees of freedom is 11.07. This means that our decision rule is:

Do not reject the null hypothesis of homoskedasticity if $NR^2 \leq 11.07$

Reject the null hypothesis of homoskedasticity if $NR^2 \geq 11.07$

$NR^2 = 42.5 > 11.07$, so we can reject the null hypothesis of homoskedasticity.

Since there appears to be heteroskedasticity in the residuals of equation (11.14), what should we do? First, we should think through the specification of the equation in search of an omitted variable. While there are a number of possible omitted variables, it turns out that the heteroskedasticity in the equation is pure heteroskedasticity for the most part.

Let's apply the most popular of our remedies, heteroskedasticity-corrected standard errors, to this example. If we start with equation (11.14) and use White's suggested method for

freedom. (Table B-1 in Appendix of doesn't include a value for 48 degrees of freedom, but we can interpolate to get a critical value of 2.01.)

estimating $SE(\hat{\beta})$ s that are minimum variance (for large samples) in the face of heteroskedasticity, we obtain:

$$\begin{aligned} PC\hat{O}N_i &= 4101 + \underset{((0.03)}{0.16} REG_i - \underset{(1360)}{1885} PRICE_i \\ t &= 4.85 \qquad \qquad -1.39 \end{aligned} \tag{11.17}$$

Compare equation (11.14) with equation (11.17). Note that the slope coefficients are identical, as you'd expect, since the HC approach uses OLS to estimate the coefficients. Also note that the HC $SE(\hat{\beta})$ s are higher than the OLS $SE(\hat{\beta})$ s, as is usually but not necessarily the case. While the resulting t -scores are lower, they are still reasonably large in the direction we expected, making equation (11.17) very appealing indeed.

A second possible remedy for heteroskedasticity is to change to a double-log functional form. If we use the data in Table 11.1 to estimate a double-log equation, we get:

$$\begin{aligned} \ln PC\hat{O}N_i &= -0.32 + \underset{(0.04)}{0.90} \ln REG_i - \underset{(1.03)}{0.89} \ln PRICE_i \\ t &= 20.3 \qquad \qquad -0.87 \\ N &= 50 \qquad \qquad \bar{R}^2 = 0.89 \end{aligned} \tag{11.18}$$

As can be seen, switching to logs improves $\bar{R}^2$ and the significance of the coefficient of $\ln REG$, but the t -score for the coefficient of $\ln PRICE$ is below 1. We'd normally be concerned about such a low t -score, but the estimated coefficient is in the expected direction, and $PRICE$ unambiguously belongs in the equation for theoretical reasons, so there's no reason to consider the possibility that $PRICE$ might be irrelevant. As we'd hope, the White test indicates that the residuals of equation (11.18) are indeed homoskedastic.

Finally, an alternative is to rethink the purpose of the regression and reformulate the variables of the equation to try to avoid heteroskedasticity resulting from spurious correlation due to size. If we were to rethink equation (11.14), we might decide to attempt to explain per capita petroleum consumption, coming up with:

$$PCON_i/POP_i = a + \beta_1 REG_i/POP_i + \beta_2 PRICE_i + u_i \quad (11.19)$$

where POP_i is the population of the i th state in thousands of people. If we estimate equation (11.19), we obtain:

$$PCON_i/POP_i = 0.23 + \underset{(0.06)}{0.15} REG_i/POP_i - \underset{(0.10)}{0.10} PRICE_i$$

$$t = 2.52 \qquad -1.00 \qquad (11.20)$$

$$N = 50 \qquad \bar{R}^2 = 0.12$$

If we compare equation (11.20) with equations (11.17) and (11.18), we see that this approach is quite different and not necessarily better. The statistical properties of equation (11.20), though not directly comparable to the other equations, do not appear as strong as they might⁷⁸ be, but this is not necessarily the deciding factor. One positive note is that the residuals of equation (11.20) do indeed appear to be homoskedastic.

Which remedy is best: HC standard errors, the double-log functional form, or reformulating the equation? Most econometricians would prefer HC standard errors, though the sample size of 50 makes it unlikely that the large sample properties of HC estimators hold in this case. However, this answer could change depending on the underlying theory of your equation. If theory strongly supports either the double-log or the reformulated functional form, then that model clearly is best. In such a situation, however, it's worth asking why the theoretically superior functional form wasn't chosen in the first place. Finally, in the fairly unusual case that t -scores aren't used to test hypotheses or retain variables, then it's not at all clear that any sort of remedy for heteroskedasticity is required at all.

⁷⁸ Petroleum-producing states like Texas, Louisiana, and Alaska consume petroleum products for many reasons other than their use in motor vehicles, so it might be tempting to add a variable measuring petroleum production to the equation. A better approach, however, would be to limit the dependent variable to the consumption of petroleum for use in motor vehicles. In addition, there is evidence that the *Statistical Abstract of the United States* data for REG may well have been incorrect for Colorado. Adjusting for this possible error improves the fit of the percapita model dramatically.

SUMMARY

1. Heteroskedasticity is the violation of Classical assumption 5 that the observations of the error term are drawn from a distribution with a constant variance. Homoskedastic error term observations are drawn from a distribution that has a constant variance for all observations, and heteroskedastic error term observations are drawn from distributions whose variances differ from observation to observation. Heteroskedasticity occurs most frequently in cross-sectional data sets.
2. The variance of a heteroskedastic error term is not equal to σ^2 , a constant. Instead, it equals σ_i^2 , where the subscript i indicates that the variance can change from observation to observation. Many different kinds of heteroskedasticity are possible, but a common model is one in which the variance changes systematically as a function of some other variable, a proportionality factor z :

$$VAR(u_i) = \sigma^2 z_i$$

The proportionality factor z is usually a variable related in some way to the size or accuracy of the dependent variable.

3. Pure heteroskedasticity is a function of the error term of the correctly specified regression equation. Impure heteroskedasticity is caused by a specification error such as an omitted variable.
4. The major consequence of heteroskedasticity is bias in the OLS $SE(\hat{\beta})$ s, causing unreliable hypothesis testing. Pure heteroskedasticity does not cause bias in the estimates of the β s themselves.
5. Two popular tests for heteroskedasticity are the Breusch - Pagan test and the White test. Both test for heteroskedasticity by analyzing the extent to which the

squared residuals of the original equation can be explained by an auxiliary equation.

6. The first step in correcting heteroskedasticity is to check for a specification error that might be causing impure heteroskedasticity. If the specification is as good as possible, then solutions such as HC standard errors or redefining the variables should be considered.

CHAPTER 12. TIME SERIES MODELS

The purpose of this chapter is to provide an introduction to a number of interesting models that have been designed to cope with and take advantage of the special properties of time-series data. Working with time-series data often causes complications that simply can't happen with cross-sectional data. Most of these complications involve the order of the observations because order matters quite a bit in time-series data but doesn't matter much (if at all) in cross-sectional data.

The most important of the topics concerns a class of dynamic models in which a lagged value of the dependent variable appears on the right-hand side of the equation. As you will see, the presence of a lagged dependent variable on the right-hand side of the equation implies that the impact of the independent variables can be spread out over a number of time periods.

Why would you want to distribute the impact of an independent variable over a number of time periods? To see why, consider the impact of advertising on sales. Most analysts believe that people remember advertising for more than one time period, so advertising affects sales in the future as well as in the current time period. As a result, models of sales should include current and lagged values of advertising, thus distributing the impact of advertising over a number of different lags.

While this chapter focuses on such dynamic models, you'll also learn about models in which different numbers of lags appear and we'll investigate how the presence of these lags affects our estimators. The chapter concludes with a brief introduction to a topic called nonstationarity. If variables have significant changes in basic properties (like their mean or variance) over time, they are said to be nonstationary, and it turns out that nonstationary variables have the potential to inflate t -scores and measures of overall fit in an equation.

12.1 Distributed lag models

As described in Section 7.3, lagged independent variables can be used whenever you expect x to affect y after a period of time. For example, if the underlying theory suggests that x_1 affects y with a one-time-period lag (but x_2 has an instantaneous impact on y), we use equations like:

$$y_t = a + \beta_1 x_{1t-1} + \beta_2 x_{2t} + u_t \quad (7.14)$$

Such lags are called simple lags, and the estimation of β_1 with OLS is no more difficult than the estimation of the coefficients of nonlagged equations, except for possible impure serial correlation if the lag is misspecified. Remember, however, that the coefficients of such equations should be interpreted carefully. For example, β_2 in equation (7.14) measures the effect of a one-unit increase in this time's x_2 on this time's y holding *last time's* x_1 constant.

A case that's more complicated than this one-period lag occurs when the impact of an independent variable is expected to be spread out over a number of time periods. Suppose, for example, that we're interested in studying the impact of a change in the money supply on GDP. Theoretical and empirical studies have provided evidence that, because of rigidities in the marketplace, it takes time for the economy to react completely to a change in the money supply. If it takes two years, some of the effect will take place immediately, some will take place with a lag of one year, and the rest will occur with a lag of two years. In such a case, the appropriate econometric model would be:

$$y_t = a + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} + u_t \quad (12.1)$$

where y would be GDP and x would be the money supply. The right-hand side of equation (12.1) is unusual, because x appears three times, each with a different lag, distributing the impact of x over a number of time periods. This is a **distributed lag model**; it explains the current value of y as a function of current and past values of x .

Can you think of another example of a dependent variable that might be appropriately explained with a distributed lag? For instance, is your grade on an econometrics exam a function only of how much you studied the night before the test, or is it impacted by your work during the previous days and weeks? Most people would agree that with a few notable exceptions, a distributed lag model would indeed be a good way to measure the impact of studying on an exam grade.

The estimation of equation (12.1) with OLS typically is straightforward. There will be some unavoidable multicollinearity between the x s, but otherwise a distributed lag model like equation (12.1) will be quite useful in a variety of applications. However, the impact of x on y often can be expected to continue over a large number of time periods, so in many cases we'll need more lagged values of x than are shown in equation (12.1). If we were building a *quarterly* model of the impact of a change in the money supply on GDP, for example, then we'd need quite a few lagged independent variables, and a more general distributed lag equation would be appropriate:

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \dots + \beta_p x_{t-p} + u_t \quad (12.2)$$

where p is the maximum number of periods by which x is lagged. In our quarterly GDP model, p might be as high as 10 or 11. (Note that in order to have the subscript of β equal the lag in x , we've defined the constant term as α_0 and β_0 now is a slope coefficient.)

Take a careful look at equation (12.2). The slope coefficients β_0 through β_p measure the effects of the various lagged values of x on the current value of y (holding constant the other independent variables in the equation). In most economic applications, including our GDP example, we'd expect the impact of x on y to decrease as the length of the lag (indicated by the subscript of the β) increases. As a result, we'd always expect β_0 and β_1 to be larger in absolute value than β_9 or β_{10} .

Unfortunately, the estimation of equation (12.2) with OLS causes a number of exercises:

1. The various lagged values of x are likely to be severely multicollinear, making coefficient estimates imprecise.
2. In large part because of this multicollinearity, there is no guarantee that the estimated β s will follow the smoothly declining pattern that economic theory would suggest. Instead, it's quite typical for the estimated coefficients of equation (12.2) to follow a fairly irregular pattern, for example:

$$\hat{\beta}_0 = 0.26 \quad \hat{\beta}_1 = 0.07 \quad \hat{\beta}_2 = 0.17 \quad \hat{\beta}_3 = 0.03 \quad \hat{\beta}_4 = 0.08$$

3. The degrees of freedom tend to decrease, sometimes substantially, for two reasons. First, we have to estimate a coefficient for each lagged x , thus increasing K and lowering the degrees of freedom ($N - K - 1$).
4. Second, unless data for lagged x s outside the sample are available, we have to decrease the sample size by 1 for each lagged x we calculate, thus lowering the number of observations, N , and therefore the degrees of freedom.

As a result of these exercises with OLS estimation of distributed lag equations like equation (12.2), it's standard practice to consider a simplifying assumption in such situations. The most commonly used simplification is to replace all the lagged independent variables with a lagged value of the dependent variable, and we'll call that kind of equation a *dynamic model*.

12.2 Dynamic models

The simplest dynamic model is:

$$y_t = a + \beta_0 x_t + \lambda y_{t-1} + u_t \quad (12.3)$$

Note that Y is on both sides of the equation! Luckily, the subscripts are different in that the y on the left-hand side is y_t , and the y on the right-hand side is y_{t-1} . It's this difference in time period that makes the equation dynamic. Thus, the simplest **dynamic**

model is an equation in which the current value of the dependent variable Y is a function of the current value of X and a lagged value of Y itself. Such a model with a lagged dependent variable is often called an *autoregressive* equation.

Let's look at equation (12.3) to try to see why it can be used to represent a distributed lag model or any model in which the impact of X on Y is distributed over a number of lags. Suppose that we lag Equation 12.3 one time period:

$$y_{t-1} = a + \beta_0 x_{t-1} + \lambda y_{t-2} + u_{t-1} \quad (12.4)$$

If we now substitute equation (12.4) for y_{t-1} in equation (12.3), we get:

$$y_t = a + \beta_0 x_t + \lambda(a + \beta_0 x_{t-1} + \lambda y_{t-2} + u_{t-1}) + u_t \quad (12.5)$$

or

$$y_t = (a + \lambda a) + \beta_0 x_t + \lambda \beta_0 x_{t-1} + \lambda^2 y_{t-2} + (\lambda u_{t-1} + u_t) \quad (12.6)$$

If we do this one more time (that is, if we lag equation (12.3) two time periods, substitute it into equation (12.5), and rearrange), we get:

$$y_t = a^* + \beta_0 x_t + \lambda \beta_0 x_{t-1} + \lambda^2 \beta_0 x_{t-2} + \lambda y_{t-3} + u_t^* \quad (12.7)$$

where a^* is the new (combined) intercept and u_t^* is the new (combined) error term. We've shown that a dynamic model can indeed be used to represent a distributed lag model!

In addition, note that the coefficients of the lagged x s follow a clear pattern. To see this, let's go back to equation (12.2):

$$y_t = a + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \dots + \beta_p x_{t-p} + u_t \quad (12.2)$$

and compare the coefficients in equation (12.2) to those in equation (12.7). We get:

$$\begin{aligned}\beta_1 &= \lambda\beta_0 \\ \beta_2 &= \lambda^2\beta_0 \\ \beta_3 &= \lambda^3\beta_0 \\ &\vdots \\ \beta_p &= \lambda^p\beta_0\end{aligned}\tag{12.8}$$

As long as λ is between 0 and 1, these coefficients will indeed smoothly decline,⁷⁹ as shown in Figure 12.1.

Dynamic models like equation (12.3) avoid the three major exercises with distributed lag equations that we outlined in the previous section. The degrees of freedom have increased dramatically, and the multicollinearity problem has disappeared. If u_t is well behaved, OLS estimation of equation (12.3) can be shown to have desirable properties for large samples. How large is “large enough”? Our recommendation, based more on experience than proof, is to aim for a sample of at least 50 observations. The smaller the sample, the more likely you are to encounter bias. In particular, estimates of λ will be biased downward, and the bias will be especially severe for larger values of λ and in the presence of additional independent variables, even irrelevant ones. As a result, samples below 30 should be avoided, in part because of this bias and in part because hypothesis testing can become unreliable.⁸⁰

In addition to this sample size issue, dynamic models face another serious problem. It turns out that serial correlation almost surely will cause bias in the OLS estimates of dynamic models. This problem will be discussed in section 12.3.

⁷⁹ This model is sometimes referred to as a Koyck distributed lag model because it was originally developed by L. M. Koyck in *Distributed Lags and Investment Analysis* (Amsterdam: NorthHolland Publishing, 1954).

⁸⁰ David Grubb and James Symons, “Bias in Regressions with a Lagged Dependent Variable,” *Econometric Theory*, Vol. 3, No. 3, pp. 371–386.

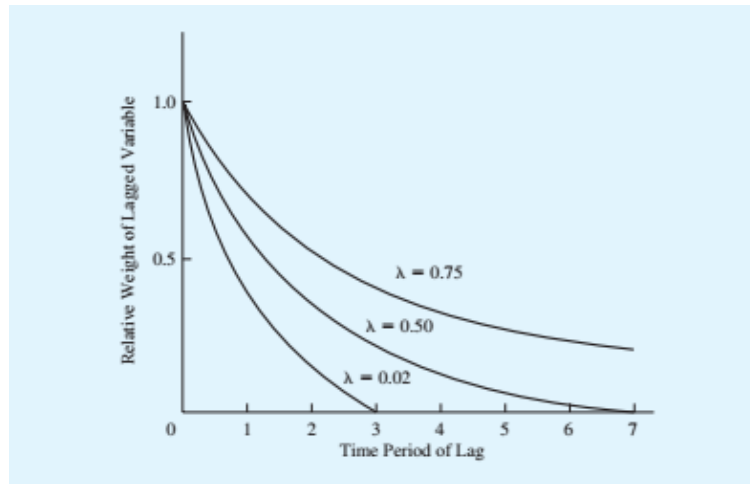


Figure 12.1 Geometric weighting schemes
for various dynamic models

As long as λ is between 0 and 1, a dynamic model has the impact of the independent variable declining as the length of the lag increases.

An Example of a dynamic model

As an example of a dynamic model, let's look at an aggregate consumption function from a macroeconomic equilibrium GDP model. Many economists argue that in such a model, consumption (CO_t) is not just an instantaneous function of disposable income (YD_t). Instead, they believe that current consumption is also influenced by past levels of disposable income (YD_{t-1} , YD_{t-2} , etc.):

$$CO_t = a + \beta_0 YD_t + \beta_1 YD_{t-1} + \beta_2 YD_{t-2} + \dots + \beta_p YD_{t-p} + u_t \quad (12.9)$$

Such an equation fits well with simple models of consumption, but it makes sense only if the coefficients of past levels of income decrease as the length of the lag increases. That is, the impact of lagged income on current consumption should decrease as the lag gets bigger. Thus we'd expect the coefficient of YD_{t-2} to be less than the coefficient of YD_{t-1} , and so on.

As a result, most econometricians would model equation (12.9) with a dynamic model:

$$CO_t = a + \beta_0 YD_t + \lambda CO_{t-1} + u_t \quad (12.10)$$

The OLS estimates of equation (12.10) for this data set are (standard errors in parentheses):

$$\begin{aligned} \hat{CO}_t &= -266.6 + 0.46 YD_t + 0.56 CO_{t-1} \\ &\quad \quad \quad (0.10) \quad \quad (0.10) \\ &\quad \quad \quad t = 4.70 \quad \quad 5.66 \\ \bar{R}^2 &= 0.999 \quad \quad N = 32 \end{aligned} \quad (12.11)$$

If we substitute $\hat{\beta}_0 = 0.46$ and $\hat{\lambda} = 0.56$ into equation (12.3) for $i = 1$, we obtain $\hat{\beta}_1 = \hat{\beta}_0 \hat{\lambda} = (0.46)(0.56) = 0.26$. If we continue this process, it turns out that equation (12.11) is equivalent to:

$$\hat{CO}_t = -605.91 + 0.46 YD_t + 0.26 YD_{t-1} + 0.14 YD_{t-2} + 0.08 YD_{t-3} + \dots \quad (12.12)$$

As can be seen, the coefficients of YD in equation (12.12) do indeed decline as we'd expect in a dynamic model.

To compare this estimate with an OLS estimate of the same equation without the dynamic model format, we'd need to estimate a distributed lag equation with at least three lagged variables.

$$CO_t = a + \beta_0 YD_t + \beta_1 YD_{t-1} + \beta_2 YD_{t-2} + \beta_3 YD_{t-3} + u_t \quad (12.13)$$

If we estimate equation (12.13) using the same data set, we get:

$$\hat{CO}_t = -695.89 + 0.73 YD_t + 0.39 YD_{t-1} + 0.006 YD_{t-2} - 0.08 YD_{t-3} \quad (12.14)$$

How do the coefficients of equation (12.14) look? As the lag increases, the coefficients of YD decrease sharply, actually going negative for $t-3$. Neither economic theory nor common sense leads us to expect this pattern. Such a poor result is due to the severe multicollinearity between the lagged x s. Most econometricians therefore estimate consumption functions with a lagged dependent variable simplification scheme like the dynamic model in equation (12.10).

An interesting interpretation of the results in equation (12.11) concerns the long-run multiplier implied by the model. The long-run multiplier measures the total impact of a change in income on consumption after all the lagged effects have been felt. One way to get this estimate would be to add up all the $\hat{\beta}$ s, but an easier

alternative is to calculate $\hat{\beta}_0[1/(1-\hat{\lambda})]$, which in this case equals $0.46[1/(1-0.56)]$ or 1.05. A sample of this size is likely to encounter small sample bias, however, so we shouldn't overanalyze the results. For more on this data set and the other equations in the model, see section 14.3. For more on testing and adjusting dynamic equations like equation (12.11) for serial correlation, let's move on to the next section.

12.3 Serial correlation and dynamic models

The consequences of serial correlation depend crucially on the type of model we're talking about. For a distributed lag model such as equation (12.2), serial correlation has the effects outlined in Section 9.3: Serial correlation causes: (1) OLS to no longer be the minimum variance unbiased estimator, (2) the $SE(\hat{\beta})$ s to be biased, and (3) no bias in the OLS $\hat{\beta}$ s themselves.

For dynamic models such as equation (12.3), however, all this changes, and serial correlation does indeed cause bias in the β ns produced by OLS. Compounding this is the fact that the detection of and remedies for serial correlation that we discussed in chapter 9 need to be modified in the presence of a lagged dependent variable.

Serial correlation causes bias in dynamic models

If an equation with a lagged dependent variable as an independent variable has a serially correlated error term, then OLS estimates of the coefficients will be biased, even in large samples.

To see where this bias comes from, let's start with a dynamic model:

$$y_t = a + \beta_0 x_t + \lambda y_{t-1} + u_t \quad (12.3)$$

and lag it one time period:

$$\overset{\uparrow}{y}_t = a + \beta_0 x_{t-1} + \lambda y_{t-2} + \overset{\uparrow}{u}_{t-1} \quad (12.15)$$

As you can see from the upward-pointing arrows above equation (12.15), if u_{t-1} is positive, then y_{t-1} will be higher than it would have

been otherwise. In addition, if u_t is serially correlated, then we know that:

$$\overset{\uparrow}{u_t} = \rho \overset{\uparrow}{u_{t-1}} + u_t \quad (12.16)$$

where u_t is a classical error term with an expected value of zero. As you can see from the arrows above equation (12.16), if u_{t-1} is positive, then u_t will be higher than it would have been otherwise as long as ρ is positive, as it typically is in economic applications. If we add the arrows from equations (12.15) and (12.16) to equation (12.3), we get:

$$y_t = a + \beta_0 x_t + \lambda \overset{\uparrow}{y_{t-1}} + \overset{\uparrow}{u_t} \quad (12.3)$$

Take a look at the arrows in equation (12.3). y_{t-1} and u_t are correlated! Such a correlation violates Classical assumption 3, which assumes that the error term is not correlated with any of the explanatory variables. (If u_{t-1} is negative, then both u_t and y_{t-1} will be lower than they would have been otherwise, which again violates Classical assumption 3). The consequences of this correlation include biased estimates, in particular of the coefficient λ , because OLS attributes to y_{t-1} some of the change in y_t actually caused by u_t . In essence, the uncorrected serial correlation acts like an omitted variable (u_{t-1}). Since an omitted variable causes bias whenever it is correlated with one of the included independent variables, and since u_{t-1} is correlated with y_{t-1} , the combination of a lagged dependent variable and serial correlation causes bias in the coefficient estimates.

Testing for serial correlation in dynamic models

If serial correlation causes bias in a dynamic model, then tests for serial correlation are obviously important. Unfortunately, however, the Durbin - Watson test is potentially invalid for an equation that contains a lagged dependent variable as an independent variable because the Durbin - Watson statistic is biased toward 2 in a dynamic equation. This bias toward 2 means that

serial correlation in a dynamic model is more likely to evade detection by the Durbin - Watson test.⁸¹

Luckily, the *Lagrange multiplier (LM)* serial correlation test of section 10.4 still is valid even in the face of a lagged dependent variable. Using the Lagrange multiplier to test for serial correlation in a typical dynamic model involves three steps that should seem familiar:

1. Obtain the residuals from the estimated equation:

$$u_t = y_t - \hat{y}_t = y_t - \hat{a} - \beta_0 x_{1t} - \hat{\lambda} y_{t-1} \quad (12.17)$$

2. Use these residuals as the dependent variable in an auxiliary equation that includes as independent variables all those on the right-hand side of the original equation as well as the lagged residuals:

$$u_t = a_0 + a_1 x_t + a_2 y_{t-1} + a_3 u_{t-1} + u_t \quad (12.18)$$

3. Estimate equation (12.18) using OLS and then test the null hypothesis that $a_3 = 0$ with the following test statistic:

$$LM = NR2 \quad (12.19)$$

where N is the sample size and R2 is the unadjusted coefficient of determination, both of the auxiliary equation, equation (12.18). For large samples, NR2 has a chi-square distribution with degrees of freedom equal to the number of restrictions in the null hypothesis (in this case, one). If NR2 is greater than the critical chi-square value from Statistical Table B-6, then we reject the null hypothesis that $a_3 = 0$ and conclude that there is indeed serial correlation in the original equation.

As an example of testing for serial correlation in a dynamic model, let's run a Lagrange multiplier (LM) serial correlation test on equation (12.11), the consumption function we estimated in the

⁸¹ The opposite is not a problem. A Durbin - Watson test that indicates serial correlation in the presence of a lagged dependent variable, despite the bias toward 2, is an even stronger affirmation of serial correlation.

previous section. If we estimate the auxiliary equation for that model, we get an R^2 of 0.4025 which, when multiplied by the sample size of 31 (do you see why it's 31 and not 32?), produces an NR^2 of 12.48. 12.48 is greater than 3.84, the 5-percent critical chi-square value with one degree of freedom, so we have strong evidence of serial correlation in equation (12.11). What should we do?

Correcting for serial correlation in dynamic models

If the Lagrange multiplier test indicates serial correlation in a dynamic model, the first step is to consider the possibility that the serial correlation is impure, perhaps caused by omission of a relevant variable and/or by failure to capture the actual distributed lag pattern accurately.

If the serial correlation appears to be pure, then the theoretically preferred solution is to transform the equation so as to eliminate the serial correlation and re-estimate the model. The required transformation is quite similar to the generalized least squares approach to serial correlation described in section 10.5. Unfortunately, the iterative nonlinear estimation of the transformed equation is well beyond the scope of this text, so it's not a realistic alternative for most readers. Instead, our suggestion is to use one of two alternatives, depending on the underlying theory of the model and the size of the sample.

If theory indicates that only a few lagged values of x are meaningful in explaining y , then a potential way to avoid the bias due to serial correlation is to estimate a distributed lag model (equation 12.2 with $p = 1$ or 2) instead of a dynamic model. The distributed lag model will have potential multicollinearity but will not likely face the other exercises typically associated with distributed lag models because p is so low. Most econometricians would prefer to deal with the consequences of multicollinearity than to face bias, so this is an improvement. There's a second advantage to distributed lags. The x_{t-p} terms, in essence, are acting as proxies for y_{t-1} . As we'll learn in chapter 14, such proxies are similar to

instrumental variables, and they are especially useful because they will eliminate bias if they're uncorrelated with the error term.

In a small sample, the best approach may well be to continue to use OLS even in the face of serial correlation in a dynamic model. For a nontrivial subset of real-world examples, OLS actually outperforms more sophisticated techniques. One reason is that in a small sample, the estimation bias in λ mentioned in section 12.2 and the bias introduced by serial correlation often are of opposite signs, so they offset each other. As a result, OLS can do better than a technique that eliminates only the bias caused by serial correlation.⁸² The second reason is that there's evidence from Monte Carlo studies that the size of the bias introduced by serial correlation in small samples often is fairly low.⁸³

To sum, unless you're comfortable with iterative nonlinear least squares, our suggestion for dealing with serial correlation in dynamic models depends on theory and the sample size. If theory calls for very few lagged x s, we suggest the distributed lag approach. If the sample is small, we think continuing to use OLS, even in the face of serial correlation, presents the best alternative. If both the sample and the meaningful number of lagged x s are large, we'd recommend using distributed lags because of the benefits of the instrumental variable approach.

12.4 Granger causality

One application of distributed lag models is to provide evidence about the direction of causality in economic relationships. Such a test is useful when we know that two variables are related but we don't know which variable causes the other to move. For example, most economists believe that increases in the money supply stimulate GDP, but others feel that increases in GDP eventually lead the monetary authorities to increase the money supply. Who's right?

⁸² Asatoshi Maeshiro, "Teaching Regressions with a Lagged Dependent Variable and Autocorrelated Disturbances," *Journal of Economic Education*, Vol. 27, No. 1, pp. 72–84.

⁸³ Luke Keele and Nathan Kelly, "Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables," *Political Analysis*, Vol. 14, No. 2, pp. 186–205.

One approach to such a question of indeterminate causality is to theorize that the two variables are determined simultaneously. We'll address the estimation of simultaneous equation models in chapter 14. A second approach to the problem is to test for what is called "Granger causality."

How can we claim to be able to test for causality? After all, didn't we say in chapter 1 that even though most economic relationships are causal in nature, regression analysis can't prove such causality? The answer is that we don't actually test for theoretical causality; instead, we test for Granger causality.

Granger causality, or precedence, is a circumstance in which one timeseries variable consistently and predictably changes before another variable.⁸⁴ Granger causality is important because it allows us to analyze which variable precedes or "leads" the other, and, as we shall see, such leading variables are extremely useful for forecasting purposes.

Despite the value of Granger causality, however, we shouldn't let ourselves be lured into thinking that it allows us to prove economic causality in any rigorous way. If one variable precedes ("Granger causes") another, we can't be sure that the first variable "causes" the other to change. As a result, even if we're able to show that event A always happens before event B, we have not shown that event A "causes" event B.

There are a number of different tests for Granger causality, and all the various methods involve lagged dependent variables in one way or another.⁸⁵ Our preference is to use an expanded version of a test originally developed by Granger. Granger suggested that to see if A Granger-caused Y, we should run:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \dots + \beta_p Y_{t-p} + \alpha_1 A_{t-1} + \dots + \alpha_p A_{t-p} + u_t \quad (12.20)$$

⁸⁴ See C. W. J. Granger, "Investigating Causal Relations by Econometric Models and CrossSpectral Methods," *Econometrica*, Vol. 37, No. 3, pp. 424-438.

⁸⁵ See John Geweke, R. Meese, and W. Dent, "Comparing Alternative Tests of Causality in Temporal Systems," *Journal of Econometrics*, Vol. 21, pp. 161-194, and Rodney Jacobs, Edward Leamer, and Michael Ward, "Difficulties with Testing for Causation," *Economic Inquiry*, Vol. 17, No. 3, pp. 401-413.

and test the null hypothesis that the coefficients of the lagged A s (the α s) jointly equal zero.⁸⁶ If we can reject this null hypothesis using the F -test, then we have evidence that A Granger-causes Y . Note that if $p = 0$, equation (12.20) is similar to the dynamic model, equation (12.3). Applications of this test involve running two Granger tests, one in each direction. That is, run equation (12.20) and also run:

$$A_t = \beta_0 + \beta_1 A_{t-1} + \dots + \beta_p A_{t-p} + \alpha_1 Y_{t-1} + \dots + \alpha_p Y_{t-p} + u_t \quad (12.21)$$

testing for Granger causality in both directions by testing the null hypothesis that the coefficients of the lagged Y s (again, the α s) jointly equal zero. If the F -test is significant for equation (12.20) but not for equation (12.21), then we can conclude that A Granger-causes Y .

12.5 Spurious correlation and nonstationarity

One problem with time-series data is that independent variables can appear to be more significant than they actually are if they have the same underlying trend as the dependent variable. In a country with rampant inflation, for example, almost any nominal variable will appear to be highly correlated with all other nominal variables. Why? Nominal variables are unadjusted for inflation, so every nominal variable will have a powerful inflationary component. This inflationary component will usually outweigh any realcausal relationship, making nominal variables appear to be correlated even if they aren't.

Such a problem is an example of **spurious correlation**, a strong relationship between two or more variables that is not caused by a real underlying causal relationship. Initially we have been considering spurious correlation in chapter. If you run a regression in which the dependent variable and one or more independent variables are spuriously correlated, the result is a *spurious regression*, and the t -scores and overall fit of such spurious regressions are likely to be overstated and untrustworthy.

⁸⁶ Such a joint test requires the use of the F -test of section 6.4.

There are many causes of spurious correlation. In a cross-sectional data set, for example, spurious correlation can be caused by dividing both the dependent variable and at least one independent variable by a third variable that varies considerably more than do the others. The focus of this section, however, will be on time-series data and in particular on spurious correlation caused by *nonstationary time series*.

Stationary and nonstationary time series

A stationary series is one whose basic properties, for example, its mean and its variance, do not change over time. In contrast, a nonstationary series has one or more basic properties that *do* change over time. For instance, the real per capita output of an economy typically increases over time, so it's nonstationary. By contrast, the growth *rate* of real per capita output often does not increase over time, so this variable is stationary even though the variable it's based on, real per capita output, is nonstationary. A time series can be nonstationary even with a constant mean if another property, such as the variance, changes over time.

More formally, a time-series variable, x_t , is **stationary** if:

1. the mean of x_t is constant over time,
2. the variance of x_t is constant over time, and
3. the simple correlation coefficient between x_t and x_{t-k} depends on the length of the lag (k) but on no other variable (for all k).⁸⁷

If one or more of these properties is not met, then x_t is **nonstationary**. And if a series is nonstationary, that problem is referred to as *nonstationarity*.

Although our definition of a stationary series focuses on stationary and nonstationary *variables*, it's important to note that *error terms* (and, therefore, residuals) also can be nonstationary. In fact, we've already had experience with a nonstationary error term. Many cases of heteroskedasticity in timeseries data involve an error

⁸⁷ There are two different definitions of stationarity. The particular definition we use here is a simplification of the most frequently cited definition, referred to by various authors as weak, wide-sense, or covariance stationarity. In addition, there are many models of nonstationarity, for example ARCH and GARCH, that are significantly more sophisticated than the model of nonstationarity introduced in this section.

term with a variance that tends to increase with time. That kind of heteroskedastic error term is also nonstationary!

The major consequence of nonstationarity for regression analysis is spurious correlation that inflates R^2 and the t -scores of the nonstationary independent variables, which in turn leads to incorrect model specification. This occurs because the regression estimation procedure attributes to the nonstationary x_t changes in y_t that were actually caused by some factor (trend, for example) that also affects x_t . Thus, the variables move together because of the nonstationarity, increasing R^2 and the relevant t -scores. This is especially important in macroeconometrics, and the macroeconomic literature includes many articles that examine various series for signs of nonstationarity.⁸⁸

Some variables are nonstationary mainly because they increase rapidly over time. Spurious regression results involving these kinds of variables often can be avoided by the addition of a simple time trend ($t = 1, 2, 3, \dots, T$) to the equation as an independent variable.

Unfortunately, many economic time-series variables are nonstationary even after the removal of a time trend. This nonstationarity often takes the form of the variable behaving as though it were a “random walk.” A **random walk** is a time-series variable in which the next period’s value equals this period’s value plus a stochastic error term. A random-walk variable is nonstationary because it can wander up and down without an inherent equilibrium and without approaching a long-term mean of any sort.

To get a better understanding of how a random walk gives rise to nonstationarity, let’s suppose that y_t is generated by an equation that includes only past values of itself (an *autoregressive* equation):

$$y_t = \mathcal{Y}_{t-1} + v_t \quad (12.22)$$

where v_t is a classical error term.

⁸⁸ See, for example, C. R. Nelson and C. I. Plosser, “Trends and Random Walks in Macroeconomics Time Series: Some Evidence and Implication,” *Journal of Monetary Economics*, Vol. 10, pp. 169-182, and J. Campbell and N. G. Mankiw, “Permanent and Transitory Components in Macroeconomic Fluctuations,” *American Economic Review*, Vol. 77, No. 2, pp. 111-117.

Take a look at equation (12.22). Can you see that if $|\gamma| < 1$, then the expected value of y_t will eventually approach 0 (and therefore be stationary) as the sample size gets bigger and bigger? (Remember, since v_t is a classical error term, its expected value = 0). Similarly, can you see that if $|\gamma| > 1$, then the expected absolute value of y_t will continuously increase, making y_t nonstationary? This is nonstationarity due to a trend, but it still can cause spurious regression results.

Most importantly, what about if $|\gamma| = 1$? In this case,

$$y_t = y_{t-1} + v_t \quad (12.23)$$

It's a random walk! The expected value of y_t does not converge on any value, meaning that it is nonstationary. This circumstance, where $\gamma = 1$ in equation (12.23) (or similar equations), is called a **unit root**. If a variable has a unit root, then equation (12.23) holds, and the variable follows a random walk and is nonstationary. The relationship between unit roots and nonstationarity is so strong that some econometricians use the words interchangeably, even though they recognize that many factors other than unit roots can cause nonstationarity.

EXAMPLE 12.1

Spurious regression

As noted at the beginning of this section, if the dependent variable and at least one independent variable in an equation are trending, as they will if they contain unit roots, it's possible for the results of an OLS regression to be spurious.⁸⁹ Consider the linear regression model

$$y_t = \alpha_0 + \beta_0 x_t + u_t \quad (12.24)$$

If both x and y are nonstationary, then they can be highly correlated for noncausal reasons, and our standard regression inference measures will be very misleading in that they'll overstate $\bar{R}^2$ and the t -score for β_0 . For example, take a look at the following estimated equation:

⁸⁹ See C. W. J. Granger and P. Newbold, "Spurious Regression in Econometrics," *Journal of Econometrics*, Volume 2, pp. 111–120.

$$\begin{aligned} \widehat{PRICE}_t &= -27.8 + 0.070 TUITON_t \\ &\quad \quad \quad (0.006) \\ t &= 11.4 \\ \bar{R}^2 &= 0.94 \quad T = 10 \end{aligned} \tag{12.25}$$

The $\bar{R}^2$ of this equation and the t -score for the coefficient of *TUITON* are clearly significant, but what are the definitions of the variables? Well, *PRICE* is the price of a one liter of gasoline in Khatlon, Tajikistan, and *TUITON* is the tuition for a semester of study at Tajik national university (TNU) in Dushanbe (both measured in nominal somoni). Is it possible that an increase in the tuition at TNU caused gas prices in Khatlon to go up? Not unless every TNU student was the child of a Khatlon gas station owner! What's going on? Well, this regression is from the 2000s, a decade of inflation, so any nominally measured variables are likely to result in an equation that fits as well as equation (12.25). Both variables are nonstationary, and this particular regression result clearly is spurious.

To avoid spurious regression results, it's crucial to be sure that time-series variables are stationary before running regressions.

The Dickey–Fuller test

To ensure that the equations we estimate are not spurious, it's important to test for nonstationarity. If we can be reasonably sure that all the variables are stationary, then we need not worry about spurious regressions. How can you tell if a time series is nonstationary? The first step is to visually examine the data. For many time series, a quick glance at the data (or a diagram of the data) will tell you that the mean of a variable is increasing dramatically over time and that the series is nonstationary.

After this trend has been removed, the standard method of testing for nonstationarity is the **Dickey–Fuller test**,⁹⁰ which examines the hypothesis that the variable in question has a unit root⁹¹ and, as a result, is likely to benefit from being expressed in first-difference form.

⁹⁰ D. A. Dickey and W. A. Fuller, “Distribution of the Estimators for Autoregressive Time-Series with a Unit Root,” *Journal of the American Statistical Association*, Vol. 74, pp. 427–431.

⁹¹ For more on unit roots, see John Y. Campbell and Pierre Peron, “Pitfalls and Opportunities: What Macroeconomists Should Know About Unit Roots,” *NBER Macroeconomics Annual* (Cambridge, MA: MIT Press, 1991), pp. 141–219.

To best understand how the Dickey - Fuller test works, let's return to the discussion of the role that unit roots play in the distinction between stationarity and nonstationarity. Recall that we looked at the value of γ in equation (12.22) to help us determine if y was stationary or nonstationary:

$$y_t = \gamma y_{t-1} + v_t \quad (12.22)$$

We decided that if $|\gamma| < 1$, then y is stationary, and that if $|\gamma| > 1$, then y_t is nonstationary. However, if $|\gamma| = 1$, then y_t is nonstationary due to a unit root. Thus we concluded that the autoregressive model is stationary if $|\gamma| < 1$ and nonstationary otherwise.

From this discussion of stationarity and unit roots, it makes sense to estimate equation (12.22) and determine if $|\gamma| < 1$ to see if y is stationary, and that's almost exactly how the Dickey - Fuller test works. First, we subtract y_{t-1} from both sides of equation (12.22), yielding:

$$(y_t - y_{t-1}) = (\gamma - 1)y_{t-1} + v_t \quad (12.26)$$

If we define $\Delta y_t = y_t - y_{t-1}$ then we have the simplest form of the Dickey - Fuller test:

$$\Delta y_t = \beta_1 y_{t-1} + v_t \quad (12.27)$$

where $\beta_1 = \gamma - 1$. The null hypothesis is that y_t contains a unit root and is therefore nonstationary, and the alternative hypothesis is that y_t is stationary.

If y_t contains a unit root, $\gamma = 1$ and $\beta_1 = 0$. If y_t is stationary, $|\gamma| < 1$ and $\beta_1 < 0$. Hence we construct a one-sided t -test on the hypothesis that $\beta_1 = 0$:

$$\begin{aligned} H_0: \beta_1 &= 0 \\ H_A: \beta_1 &< 0 \end{aligned}$$

The Dickey - Fuller test actually comes in three versions:

1. Equation (12.27),
2. Equation (12.27) with a constant term added (equation (12.28)), and
3. Equation (12.27) with a constant term and a trend term added (equation (12.29)).

The form of the Dickey - Fuller test in equation (12.27) is correct if y_t follows equation (12.22), but most econometricians add a constant term to the equation, for reasons similar to those mentioned in section 8.1, so the basic Dickey - Fuller test equation becomes:

$$\Delta y_t = \beta_0 + \beta_1 y_{t-1} + v_t \quad (12.28)$$

Alternatively, if we believe y_t contains a trend “ t ” ($t = 1, 2, 3, \dots, T$), then we’d add “ t ” to the equation as a *variable* with a coefficient, and the appropriate Dickey - Fuller test equation is:

$$\Delta y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 t + v_t \quad (12.29)$$

How do we decide whether to use equation (12.28) or equation (12.29)? If you compare the two equations, you can see that the only difference between them is a trend term ($\beta_2 t$) that’s in equation (12.29) but not in equation (12.28).

Thus equation (12.29) is appropriate if y is growing and equation (12.28) is appropriate if y is not growing. Perhaps the best way to make this decision is to plot y over time and then judge whether or not it appears to be growing.⁹² GDP and consumption are good examples of variables that usually are growing over time,

⁹² John Elder and Peter Kennedy, “Testing for Unit Roots: What Should Students Be Taught?” *Journal of Economic Education*, Vol. 32, No. 2, pp. 137–146. Elder and Kennedy also investigate what to do in the unlikely case that the growth status of Y is unknown.

while most rates (like interest rates and unemployment rates) are good examples of variables that are not growing.

No matter which form of the Dickey - Fuller test we use, the decision rule is the same. If $\hat{\beta}_1$ is significantly less than 0 as measured by a t -test, then we can reject the null hypothesis of a unit root; this implies that the variable is stationary. If $\hat{\beta}_1$ is not significantly less than 0, then we cannot reject the null hypothesis of a unit root; this implies that the variable is nonstationary. (Recall from chapter 6 that if we're not able to reject the null hypothesis, we still have not "proven" that y is nonstationary).

Be careful, however. The standard t -Table does not apply to Dickey - Fuller tests. Instead the critical values are somewhat higher than those in statistical Table B-1, and they also depend on the version of the Dickey - Fuller test you use. In Table 12.1,⁹³ we list critical values for the two versions of the Dickey - Fuller test we have discussed the most, equations (12.28) and (12.29). For example, a 5% one-sided test of β_1 in equation (12.28) with a sample size of 25 has a critical t -value of 3.00 compared to 1.717 for a standard t -test (with 22 degrees of freedom since $K = 2$).

The Dickey - Fuller specifications in equations (12.28) and (12.29) and the critical values for those specifications are derived under the assumption that the error term is serially uncorrelated. If the error term is serially correlated, then the test must be modified to take this serial correlation into account. This adjustment, called the Augmented Dickey - Fuller test (ADF), adds a series of lagged values of Δy to the Dickey-Fuller test. While the ADF is the most-used version of the Dickey-Fuller test, it is beyond the scope of this textbook, at least in part because choosing how many lagged Δy values to include can be quite complicated.

⁹³ Most sources list negative critical values for the Dickey-Fuller test, because the unit root test is one-sided with a negative expected value. However, the t -test decision rule of this text is based on the absolute value of the t -score, so negative critical values would cause every null hypothesis to be rejected. As a result, the critical values in Table 12.1 are positive. For adjusted critical t -values for the Dickey-Fuller test, including those in Table 12.1, see J. G. MacKinnon, "Critical Values of Cointegration Tests," in Rob Engle and C. W. J. Granger, eds., *Long-Run Economic Relationships: Readings in Cointegration* (New York: Oxford University Press, 1991), chapter 13. Most software packages provide these critical values with the output from a Dickey - Fuller test.

Table 12.1 5% one-sided critical Values for the Dickey–Fuller test

Sample Size (T)	For Equation 12.28 (Y not growing)	For Equation 12.29 (Y growing)
25	3.00	3.60
50	2.93	3.50
100	2.89	3.45
∞	2.86	3.41

Cointegration

If the Dickey - Fuller test reveals nonstationarity, what should we do? The traditional approach is to take the first differences ($\Delta y = y_t - y_{t-1}$ and $\Delta x = x_t - x_{t-1}$) and use them in place of y_t and x_t in the equation. With economic data, taking a first difference usually is enough to convert a nonstationary series into a stationary one. Unfortunately, using first differences to correct for nonstationarity throws away information that economic theory can provide in the form of equilibrium relationships between the variables when they are expressed in their original units (x_t and y_t). As a result, first differences should not be used without carefully weighing the costs and benefits of that shift, and in particular first differences should not be used until the residuals have been tested for *cointegration*.

Cointegration consists of matching the degree of nonstationarity of the variables in an equation in a way that makes the error term (and residuals) of the equation stationary and rids the equation of any spurious regression results. Even though individual variables might be nonstationary, it's possible for linear combinations of nonstationary variables to be stationary, or *cointegrated*. If a long-run equilibrium relationship exists between a set of variables, those variables are said to be cointegrated. If the variables are cointegrated, then you can avoid spurious regressions even though the dependent variable and at least one independent variable are nonstationary. To see how this works, let's return to equation (12.24):

$$y_t = \alpha_0 + \beta_0 x_t + u_t \quad (12.24)$$

As we saw in the previous section, if x_t and y_t are nonstationary, it's likely that we'll get spurious regression results. To understand how it's possible to get sensible results from equation (12.24) if the nonstationary variables are cointegrated, let's focus on the case in which both x_t and y_t contain one unit root. The key to cointegration is the behavior of u_t . If we solve equation (12.24) for u_t , we get:

$$u_t = y_t - \alpha_0 - \beta_0 x_t \quad (12.30)$$

In equation (12.30), u_t is a function of two nonstationary variables, so you'd certainly expect u_t also to be nonstationary, but that's not necessarily the case. In particular, suppose that x_t and y_t are related? More specifically, if economic theory supports equation (12.24) as an equilibrium, then departures from that equilibrium should not be arbitrarily large.

Hence, if y_t and x_t are related, then the error term u_t may well be stationary even though x_t and y_t are nonstationary. If u_t is stationary, then the unit roots in y_t and x_t have "cancelled out" and y_t and x_t are said to be cointegrated.⁹⁴

We thus see that if x_t and y_t are cointegrated, then OLS estimation of the coefficients in equation (12.24) can avoid spurious results. To determine if x_t and y_t are cointegrated, we begin with OLS estimation of equation (12.24) and calculate the OLS residuals:

$$\hat{u}_t = y_t - \hat{\alpha}_0 - \hat{\beta}_0 x_t \quad (12.31)$$

We then perform a Dickey - Fuller test on the residuals. Once again, the standard critical t -values do not apply to this application, so adjusted critical t -values should be used.⁹⁵ If we are able to reject

⁹⁴ For more on cointegration, see Peter Kennedy, *A Guide to Econometrics* (Malden, MA: Blackwell, 2008), pp. 309–313 and 327–330, and B. Bhaskara Rau, ed., *Cointegration for the Applied Economist* (New York: St. Martin's Press, 1994).

⁹⁵ See J. G. MacKinnon, "Critical Values of Cointegration Tests," in Rob Engle and C. W. J. Granger, eds., *Long-Run Economic Relationships: Readings in Cointegration* (New York: Oxford University Press, 1991), Chapter 13, and Rob Engle and C. W. J. Granger, "Co-integration and Error Correction: Representation, Estimation and Testing," *Econometrica*, Vol. 55, No. 2.

the null hypothesis of a unit root in the residuals, we can conclude that Y_t and X_t are cointegrated and our OLS estimates are not spurious.

To sum, if the Dickey–Fuller test reveals that our variables have unit roots, the first step is to test for cointegration in the residuals. If the nonstationary variables are not cointegrated, then the equation should be estimated using first differences (ΔY and ΔX). However, if the nonstationary variables are cointegrated, then the equation can be estimated in its original units.⁹⁶

A Standard sequence of steps for dealing with nonstationary time series

This material is fairly complex, at least compared to previous chapters, so let's pause for a moment to summarize the various steps suggested in section 12.5. To deal with the possibility that nonstationary time series might be causing regression results to be spurious, most empirical work in time series follows a standard sequence of steps:

1. Specify the model. This model might be a time-series equation with no lagged variables, a distributed lag model or a dynamic model.
2. Test all variables for unit roots using the appropriate version of the Dickey - Fuller test.
3. If the variables don't have unit roots, estimate the equation in its original units (y and x).
4. If the variables have unit roots, test the residuals of the equation for cointegration using the Dickey - Fuller test.

⁹⁶ In this case, it's common practice to use a version of the original equation called the Error Correction Model (ECM). While the equation for the ECM is fairly complex, the model itself is a logical extension of the cointegration concept. If two variables are cointegrated, then there is an equilibrium relationship connecting them. A regression on these variables therefore is an estimate of this equilibrium relationship along with a residual, which is a measure of the extent to which these variables are out of equilibrium. When formulating a dynamic relationship between the variables, economic theory suggests that the current change in the dependent variable should be affected not only by the current change in the independent variable but also by the extent to which these variables were out of equilibrium in the preceding period (the residual from the cointegrating process). The resulting equation is the ECM. For more on the ECM, see Peter Kennedy, *A Guide to Econometrics* (Malden, MA: Blackwell, 2008), pp. 299–301 and 322–323.

5. If the variables have unit roots but are not cointegrated, then change the functional form of the model to first differences (Δy and Δx) and estimate the equation.
6. If the variables have unit roots and also are cointegrated, then estimate the equation in its original units.

SUMMARY

1. A distributed lag explains the current value of y as a function of current and past values of x , thus “distributing” the impact of x over a number of lagged time periods. OLS estimation of distributed lag equations without any constraints encounters exercises with multicollinearity, degrees of freedom, and a noncontinuous pattern of estimated coefficients over time.
2. A dynamic model avoids these exercises by assuming that the coefficients of the lagged independent variables decrease in a geometric fashion the longer the lag. Given this, the dynamic model is:

$$y_t = a + \beta_0 x_t + \lambda y_{t-1} + u_t$$

where y_{t-1} is a lagged dependent variable and $0 < \lambda < 1$.

3. In small samples, OLS estimates of a dynamic model are biased and have unreliable hypothesis testing properties. Even in large samples, OLS will produce biased estimates of the coefficients of a dynamic model if the error term is serially correlated.
4. In a dynamic model, the Durbin - Watson test is biased toward 2, so it should not be used. The most-used alternative is the Lagrange multiplier test.
5. Granger causality, or precedence, is a circumstance in which one timeseries variable consistently and predictably changes before another variable does. If one variable precedes

(Granger-causes) another, we still can't be sure that the first variable "causes" the other to change.

6. A nonstationary series is one that exhibits significant changes (for example, in its mean and variance) over time. If the dependent variable and at least one independent variable are nonstationary or trending, a regression may encounter spurious correlation that inflates $\bar{R}^2$ and t -scores.
7. Nonstationarity can be detected using the Dickey - Fuller test. If the variables have unit roots and therefore are nonstationary, then the residuals of the equation should be tested for cointegration using the Dickey - Fuller test. If the variables are nonstationary but are not cointegrated, then the equation should be estimated with first differences. If the variables are nonstationary and also are cointegrated, then the equation can be estimated in its original units.

CHAPTER 13. DUMMY DEPENDENT VARIABLE TECHNIQUE

Until now, our discussion of dummy variables has been restricted to dummy independent variables. However, there are many important research topics for which the *dependent* variable is appropriately treated as a dummy, equal only to 0 or 1.

In particular, researchers analyzing consumer choice often must cope with dummy dependent variables (also called qualitative dependent variables). For example, how do high school students decide whether to go to college? What distinguishes Pepsi drinkers from Coke drinkers? How can we convince people to use public transportation instead of driving? For an econometric study of these topics, or of any topic that involves a *discrete* choice of some sort, the dependent variable is typically a dummy variable. In the first two sections of this chapter, we'll present two frequently used ways to estimate equations that have dummy dependent variables: the linear probability model and the binomial logit model. In the last section, we'll briefly discuss the binomial probit model and multinomial models.

13.1 The linear probability model

What is a linear probability model?

The most obvious way to estimate a model with a dummy dependent variable is to run OLS on a typical linear econometric equation. A **linear probability model** is just that, a linear-in-the-coefficients equation used to explain a dummy dependent variable:

$$d_i = a + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \quad (13.1)$$

where d_i is a dummy variable and the x s, β s, and u are typical independent variables, regression coefficients, and an error term, respectively.

For instance, suppose you're interested in understanding why some states have female governors and others don't. In such a model, the appropriate dependent variable would be a dummy, for example d_i equal to 1 if the i th state has a female governor and

equal to 0 otherwise. If we hypothesize that states with a high percentage of females and a low percentage of social conservatives would be likely to have a female governor, then a linear probability model would be:

$$d_i = a + \beta_1^+ f_i + \beta_2^- r_i + u_i \quad (13.2)$$

where: $d_i = 1$ if the i th state has a female governor, 0 otherwise
 f_i = females as a percent of the i th state's population
 r_i = conservatives as a percent of the i th state's registered voters

The term *linear probability model* comes from the fact that the right side of the equation is linear while the expected value of the left side measures the probability that $d_i = 1$. To understand this second statement, let's assume that we estimate equation (13.2) and get a $\hat{d}_i$ of 0.10 for a particular state. What does that mean? Well, since $d = 1$ if the governor is female and $d = 0$ if the governor is male, a state with a $\hat{d}_i$ of 0.10 can perhaps best be thought of as a state in which there is a 10% chance that the governor will be female, based on the state's values for the independent variables. Thus $\hat{d}_i$ is an estimate of the probability that $d_i = 1$ for the i th observation, and:

$$\hat{d}_i = pr(\hat{d}_i = 1) = \hat{a} + \hat{\beta}_1 f_i + \hat{\beta}_2 r_i \quad (13.3)$$

where $pr(d_i = 1)$ indicates the probability that $d_i = 1$ for the i th observation.

How should we interpret the coefficients of equation (13.3)? Since $\hat{d}_i$ measures the probability that $d_i = 1$, then a coefficient in a linear probability model is an estimate of the change in the probability that $d_i = 1$ caused by a one-unit increase in the independent variable in question, holding constant the other independent variables in the equation.

Let's define p_i to be the true probability that $d_i = 1$. We can never observe p_i , just as we can never observe true β s, because it reflects the underlying situation before a discrete choice is made. After the choice is made, we can observe only the outcome of that choice, and so the dependent variable d_i can take on the values of only 0 or 1. Thus, even though p_i can be any value between 0 and 1, we can observe only the two extremes (0 and 1) in d_i .

Exercises with the linear probability model

Unfortunately, using OLS to estimate the coefficients of an equation with a dummy dependent variable faces at least three exercises:

1. *R² is not an accurate measure of overall fit.* For models with a dummy dependent variable, R^2 tells us very little about how well the model explains the choices of the decision makers. To see why, take a look at Figure 13.1. d_i can equal only 1 or 0, but $\hat{d}_i$ must move in a continuous fashion from one extreme to the other. This means that $\hat{d}_i$ is likely to be quite different from d_i for some range of x_i . Thus, R^2 is likely to be much lower than 1 even if the model actually does an exceptional job of explaining the choices involved. As a result, $\bar{R}^2$ (or R^2) should not be relied on as a measure of the overall fit of a model with a dummy dependent variable.
2. *$\hat{d}_i$ is not bounded by 0 and 1.* Since d_i is a dummy variable, we'd expect $\hat{d}_i$ to be limited to a range of from 0 to 1. After all, the prediction that a probability equals 2.6 (or -2.6, for that matter) is almost meaningless. However, take another look at equation (13.3). Depending on the values of the independent variables and the $\hat{\beta}$ s, the right-hand side might well be outside the meaningful range. For instance, if f , r , and all the $\hat{\beta}$ s in equation (13.3) equal 1.0, then $\hat{d}_i$ equals 3.0, which is substantially greater than 1.0.
3. *The error term is neither homoskedastic nor normally distributed.* In addition, the error term in a linear probability model is heteroskedastic and is not distributed normally, mainly because d_i takes on only two values (0 and 1). In practice, however, the impact of these exercises on OLS estimation is minor, so many researchers ignore potential heteroskedasticity and nonnormality and apply OLS directly to the linear probability model.⁹⁷

The first of these exercises isn't impossible to deal with, because there are a variety of alternatives to R^2 for equations with

⁹⁷ See R.G. McGilvray, "Estimating the Linear Probability Function," *Econometrica*, Vol. 38, pp. 775–776.

dummy dependent variables.⁹⁸ Our preference is to create a measure based on the percentage of the observations in the sample that a particular estimated equation explains correctly.

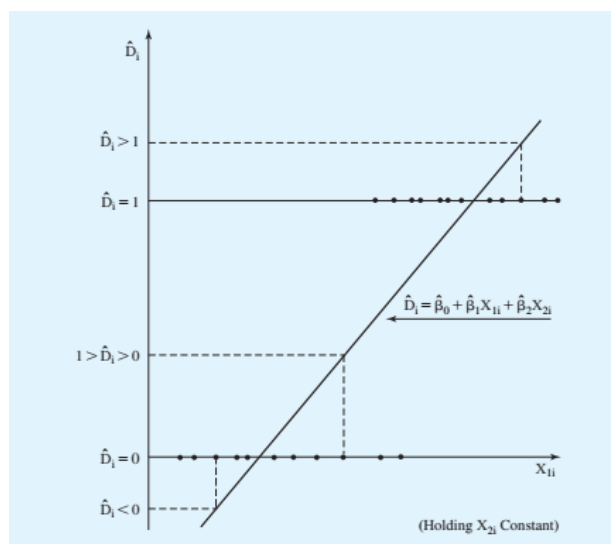


Figure 13.1 A linear probability model

In a linear probability model, all the observed d_i s equal either 0 or 1, but $\hat{d}_i$ moves linearly from one extreme to the other. As a result, $\bar{R}^2$ is often quite low even if the model does an excellent job of explaining the decision maker's choice. In addition, exceptionally large or small values of x_{1i} (holding x_{2i} constant) can produce values of $\hat{d}_i$ outside the meaningful range of 0 to 1.

To use this approach, consider a $\hat{d}_i > 0.5$ to predict that $d_i = 1$ and a $\hat{d}_i < 0.5$ to predict that $d_i = 0$. If we then compare these predictions⁹⁹ with the actual d_i , we can calculate the percentage of observations explained correctly.

Unfortunately, using the percentage explained correctly as a substitute for R^2 for the entire sample has a flaw. Suppose that 85%

⁹⁸ See M. R. Veal and K. F. Zimmerman, "Pseudo- R^2 Measures for Some Common Limited Dependent Variables Models," *Journal of Economic Surveys*, Vol. 10, No. 3, pp. 241–259 and C. S. McIntosh and J. J. Dorfman, "Qualitative Forecast Evaluation: A Comparison of Two Performance Measures," *American Journal of Agricultural Economics*, Vol. 74, pp. 209–214.

⁹⁹ Although it's standard to use $\hat{d}_i > 0.5$ as the value that distinguishes a prediction of $d_i = 1$ from a prediction of $d_i = 0$, there's no rule that requires that 0.5 be used. This is because it's possible to imagine circumstances in which 0.5 is too high or too low. For example, if the payoff when you're right if you classify $d_i = 1$ is much lower than the payoff when you're right if you classify $d_i = 0$, then a value other than 0.5 might make sense.

of your observations are 1s and 15% are 0s. Explaining 85% of the sample correctly sounds good, but your results are no better than naively guessing that every observation is a 1! A better way might be to calculate the percentage of 1s explained correctly, calculate the percentage of zeroes explained correctly, and then report the average of these two percentages. As a shorthand, we'll call this average $\bar{R}_p^2$. That is, we'll define $\bar{R}_p^2$ to be the average of the percentage of 1s explained correctly and the percentage of zeroes explained correctly. Since $\bar{R}_p^2$ is a new statistic, we'll calculate and discuss both $\bar{R}_p^2$ and $\bar{R}^2$ throughout this chapter.

For most researchers, therefore, the major difficulty with the linear probability model is the unboundedness of the predicted d_i s. Take another look at Figure 13.1 for a graphical interpretation of the situation. Because of the linear relationship between the x_i s and $\hat{d}_i$, $\hat{d}_i$ can fall well outside the relevant range of 0 to 1.

One simplistic way to get around the unboundedness problem is to assign $\hat{d}_i = 1.0$ to all values of $\hat{d}_i$ above 1 and $\hat{d}_i = 0.0$ to all negative values. This approach copes with the problem by ignoring it, since an observation for which the linear probability model predicts a probability of 2.0 has been judged to be more likely to be equal to 1.0 than an observation for which the model predicts a 1.0, and yet they are lumped together. Even $\hat{d}_i = 1$ isn't very useful, because it implies that events will happen with certainty, which is surely a foolish prediction to make. What is needed is a systematic method of forcing the $\hat{d}_i$ s to range from 0 to 1 in a smooth and meaningful fashion. We'll present such a method, the binomial logit, in section 13.2.

Using the linear probability model, despite this unboundedness problem, may not cause insurmountable difficulties. In particular, the signs and general significance levels of the estimated coefficients of the linear probability model are often similar to those of the alternatives we will discuss later in this chapter.

EXAMPLE 13.1

An example of a linear probability model

Let's take a look at an example of a linear probability model: a disaggregate study of the labor force participation of women. A person is defined as being in the labor force if she either has a job or is actively looking for a job. Thus, a disaggregate (cross-sectional by person) study of women's labor force participation is appropriately modeled with a dummy dependent variable:

$$d_i = 1 \text{ if the } i\text{th woman has or is looking for a job,} \\ 0 \text{ otherwise (not in the labor force)}$$

A review of the literature reveals that there are many potentially relevant independent variables. Two of the most important are the marital status and the number of years of schooling of the woman. The expected signs for the coefficients of these variables are fairly straightforward, since a woman who is unmarried and well educated is much more likely to be in the labor force than her opposite. If we choose a linear functional form, we've got a linear probability model:

$$d_i = a + \beta_1^- m_i + \beta_2^+ s_i + u_i \quad (13.4)$$

where: $m_i = 1$ if the i th woman is married and 0 otherwise

s_i = the number of years of schooling of the i th woman

The data are presented in Table 13.1. The sample size is limited to 30 in order to make it easier for readers to enter the dataset on their own. Unfortunately, such a small sample will make hypothesis testing fairly unreliable. Table 13.1 also includes the age of the i th woman, a_i . Another typically used variable, o_i = other income available to the i th woman, is not available for this sample, introducing possible omitted variable bias.

If we now estimate equation (13.4) with the data on the labor force participation of women from Table 13.1, we obtain (standard errors in parentheses):

$$\hat{d}_i = -0.28 - \underset{(0.15)}{0.38} m_i + \underset{(0.03)}{0.09} s_i \quad (13.5)$$

$$N = 30 \quad \bar{R}^2 = 0.32 \quad \bar{R}_p^2 = 0.81$$

How do these results look? Despite the small sample and the possible bias due to omitting o_i , both independent variables have estimated coefficients that are significant in the expected direction. In addition, the R^2 of 0.32 is fairly high for a linear probability model. (Since d_i equals only 0 or 1, it's almost impossible to get an R^2 much higher than 0.70.) Further evidence of good fit is the fairly high $\bar{R}_p^2$ of 0.81, meaning that an average of 81% of the choices were explained "correctly" by equation (13.5).

We need to be careful when we interpret the estimated coefficients in equation (13.5), however. Remember that the slope coefficient in a linear probability model represents the change in the probability that d_i equals 1 caused by a one-unit increase in the independent variable (holding the other independent variables constant). Viewed in this context, do the estimated coefficients make economic sense? The answer is yes: The probability of a woman participating in the labor force falls by 38 percentage points if she is married (holding constant schooling). Each year of schooling increases the probability of labor force participation by 9 percentage points (holding constant marital status).

Observation #	D_i	M_i	A_i	S_i	$\hat{D}_i$
1	1.0	0.0	31.0	16.0	1.20
2	1.0	1.0	34.0	14.0	0.63
3	1.0	1.0	41.0	16.0	0.82
4	0.0	0.0	67.0	9.0	0.55
5	1.0	0.0	25.0	12.0	0.83
6	0.0	1.0	58.0	12.0	0.45
7	1.0	0.0	45.0	14.0	1.01
8	1.0	0.0	55.0	10.0	0.64
9	0.0	0.0	43.0	12.0	0.83
10	1.0	0.0	55.0	8.0	0.45
11	1.0	0.0	25.0	11.0	0.73
12	1.0	0.0	41.0	14.0	1.01
13	0.0	1.0	62.0	12.0	0.45
14	1.0	1.0	51.0	13.0	0.54
15	0.0	1.0	39.0	9.0	0.17
16	1.0	0.0	35.0	10.0	0.64
17	1.0	1.0	40.0	14.0	0.63
18	0.0	1.0	43.0	10.0	0.26
19	0.0	1.0	37.0	12.0	0.45
20	1.0	0.0	27.0	13.0	0.92
21	1.0	0.0	28.0	14.0	1.01
22	1.0	1.0	48.0	12.0	0.45
23	0.0	1.0	66.0	7.0	-0.01
24	0.0	1.0	44.0	11.0	0.35
25	0.0	1.0	21.0	12.0	0.45
26	1.0	1.0	40.0	10.0	0.26
27	1.0	0.0	41.0	15.0	1.11
28	0.0	1.0	23.0	10.0	0.26
29	0.0	1.0	31.0	11.0	0.35
30	1.0	1.0	44.0	12.0	0.45

Table 13.1 Data on the labor force participation of women

The values for $\hat{d}_i$ have been included in Table 13.1. Note that $\hat{d}_i$ is indeed often outside the meaningful range of 0 and 1, causing most of the exercises cited earlier. To attack this problem of the unboundedness of $\hat{d}_i$, however, we need a new estimation technique, so let's take a look at one.

13.2 The binomial logit model

What Is the Binomial Logit?

The **binomial logit** is an estimation technique for equations with dummy dependent variables that avoids the unboundedness problem of the linear probability model by using a variant of the cumulative logistic function:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}} \quad (13.6)$$

where p_i is the true probability that $d_i = 1$. We can't observe p_i , so we need to use observed d_i s to estimate a logit equation like equation (13.6). That estimation will produce $\hat{d}_i$ s that we can compare to the $\hat{d}_i$ s produced by an estimated linear probability model like equation 13.3. Are the $\hat{d}_i$ s produced by a logit now limited by 0 and 1? The answer is yes, but to see why we need to take a close look at equation (13.6). What is the largest that $\hat{d}_i$ can be? Well, if $\hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$ equals infinity, then:

$$\hat{d}_i = \frac{1}{1 + e^{-\infty}} = \frac{1}{1} = 1 \quad (13.7)$$

because e to the minus infinity equals 0. What's the smallest that $\hat{d}_i$ can be? If $\hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$ equals minus infinity, then:

$$\hat{d}_i = \frac{1}{1 + e^{\infty}} = \frac{1}{\infty} = 0 \quad (13.8)$$

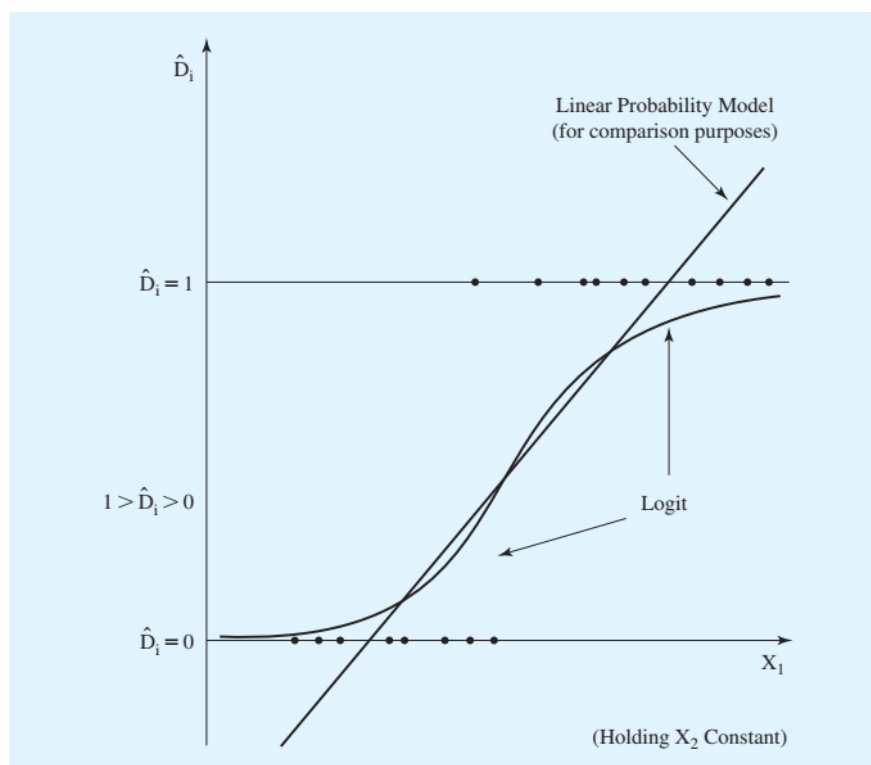


Figure 13.2 $\hat{d}_i$ is bounded by 0 and 1 in a binomial Logit Model

In a binomial logit model, DN i is nonlinearly related to X_1 , so even exceptionally large or small values of x_{1i} , holding x_{2i} constant, will not produce values of $\hat{d}_i$ outside the meaningful range of 0 to 1.

Thus, $\hat{d}_i$ is bounded by 1 and 0. As can be seen in Figure 13.2, $\hat{d}_i$ approaches 1 and 0 very slowly (asymptotically). The binomial logit model therefore avoids the major problem that the linear probability model encounters in dealing with dummy dependent variables. In addition, the logit is quite satisfying to most researchers because it turns out that real-world data often are described well by S-shape patterns like that in Figure 13.2.

Logits cannot be estimated using OLS. Instead, we use **maximum likelihood (ML)**, an iterative estimation technique that is especially useful for equations that are nonlinear in the coefficients. ML estimation is inherently different from least squares in that it chooses coefficient estimates that *maximize the likelihood* of the sample data set being observed.¹⁰⁰ Interestingly, OLS and ML

¹⁰⁰ Actually, the ML program chooses coefficient estimates that maximize the probability (or likelihood) of observing the particular set of values of the dependent variable in the

estimates are not necessarily different; for a linear equation that meets the classical assumptions (including the normality assumption), ML estimates are identical to the OLS ones.

One of the reasons that maximum likelihood is used is that ML has a number of desirable large sample properties; ML is consistent (homes in on true parameter values) and asymptotically efficient (minimum variance for large samples). With very large samples, ML often has the added advantage of converging to a normal distribution, allowing the use of typical hypothesis testing techniques. As a result, sample sizes for logits should be substantially larger than they are for linear regressions. Some researchers aim for samples of 500 or more.

It's also important to make sure that a logit sample contains a reasonable representation of both alternative choices. For instance, if 98% of a sample chooses alternative A and 2% chooses B, a random sample of 500 might have only 10 observations that choose B. In such a situation, our estimated coefficients would be overly reliant on the characteristics of those 10 observations. A better technique would be to disproportionately sample from those who choose B. It turns out that using different sampling rates for subgroups within the sample does not cause bias in the slope coefficients of a logit model,¹⁰¹ even though it might do so in a linear regression. When we estimate a logit, we apply the ML technique to equation (13.6), but that equation's functional form is complex, so let's try to simplify it a bit. First, a few mathematical steps can allow us to rewrite equation (13.6) so that the right side of the equation looks like the linear probability model:

$$\ln\left(\frac{p_i}{[1-p_i]}\right) = a + \beta_1 x_{1i} + \beta_2 x_{2i} \quad (13.9)$$

sample $(y_1, y_2, \dots, y_N)$ for a given set of x s. For more on maximum likelihood, see Robert S. Pindyck and Daniel L. Rubinfeld, *Economic Models and Economic Forecasts* (New York: McGraw-Hill, 1998), pp. 51–53 and 329–330.

¹⁰¹ The constant term, however, needs to be adjusted. Multiply $\hat{a}$ by $[\ln(p_1) - \ln(p_2)]$, where p_1 is the proportion of the observations chosen if $d_i = 1$ and p_2 is the proportion of the observations chosen if $d_i = 0$. See G. S. Maddala, *Limited-Dependent and Qualitative Variables in Econometrics* (Cambridge: Cambridge University Press, 1983), pp. 90–91.

where p_i is the true probability that $d_i = 1$.

Even equation (13.9) is a bit cumbersome, however, since the left side of the equation contains the log of the ratio of p_i to $(1 - p_i)$, sometimes called the “log of the odds.” To make things simpler still, let’s adopt a shorthand for the logit functional form on the left side of equation (13.9). Let’s define:

$$l: pr(d_i = 1) = \ln\left(\frac{p_i}{[1 - p_i]}\right) \quad (13.10)$$

The l indicates that the equation is a logit of the functional form in equation (13.9) (derived from equation (13.6)), and the “ $pr(d_i = 1)$ ” is a reminder that the dependent variable is a dummy and that a $\hat{d}_i$ produced by an estimated logit equation is an estimate of the probability that $d_i = 1$. If we now substitute equation (13.10) into equation (13.9), we get:

$$l: pr(d_i = 1) = a + \beta_1 x_{1i} + \beta_2 x_{2i} \quad (13.11)$$

Equation (13.11) will be our standard documentation format for estimated logit equations.

Interpreting estimated logit coefficients

Once you’ve estimated a binomial logit, then hypothesis testing and the analysis of potential econometric exercises can be undertaken using techniques similar to those discussed in previous chapters. The signs of the coefficients have the same meaning as they do in a linear probability model, and tests of hypotheses about logit coefficients can be run.¹⁰²

¹⁰² Different econometric software programs provide a variety of information in support of this hypothesis testing, at least in part because the t -test isn’t appropriate for hypothesis testing for logits with small samples. Stata produces z -statistics, which require the Normal Distribution table in Appendix B-5 of the tutorial *Econometric analysis using EvIEWS: a Practical tutorial*. SAS, on the other hand, produces chi-square statistics, which use Table B-6 of the *Econometric analysis using EvIEWS: a Practical tutorial*. . Our suggestion is to use p -values because they’re produced by virtually all the econometric packages and

When it comes to the economic interpretation of the estimated logit coefficients, however, all this changes. In particular, the absolute sizes of estimated logit coefficients tend to be quite different from the absolute sizes of estimated linear probability model coefficients for the same specification and the same data. What's going on?

There are two powerful reasons for these differences. First, as you can see by comparing Equations 13.1 and 13.9, the dependent variable in a logit equation isn't the same as the dependent variable in a linear probability model. Since the dependent variable is different, it makes complete sense that the coefficients are different. The second reason that logit coefficients are different is even more dynamic. Take a look at Figure 13.2. The slope of the graph of the logit changes as $\hat{d}_i$ moves from 0 to 1! Thus the change in the probability that $\hat{d}_i = 1$ caused by a one-unit increase in an independent variable (holding the other independent variables constant) will vary as we move from $\hat{d}_i = 0$ to $\hat{d}_i = 1$.

Given all this, how can we interpret estimated logit coefficients? How can we use them to measure the impact of an independent variable on the probability that $\hat{d}_i = 1$? It turns out that there are three reasonable ways of answering this question:

1. *Change an average observation.* Create an "average" observation by plugging the means of all the independent variables into the estimated logit equation and then calculating an "average" $\hat{d}_i$. Then increase the independent variable of interest by one unit and recalculate the $\hat{d}_i$. The difference between the two $\hat{d}$ is tells you the impact of a one-unit increase in that independent variable on the probability that $\hat{d}_i = 1$ (holding constant the other independent variables) for an average observation. This approach has the weakness of not being very meaningful when one or more of the independent variables is a dummy variable (after all, what is an average gender?), but it's

because their use doesn't require the researcher to decide whether the t -distribution, normal distribution, or chi-square distribution is appropriate in any given case.

possible to work around this weakness if you estimate the impact for an “average female” and an “average male” by setting the dummy independent variable equal first to 0 and then to 1.

2. *Use a partial derivative.* It turns out that if you take a derivative of the logit, you’ll find that the change in the expected value of $\hat{d}_i$ caused by a one-unit increase in x_{1i} , holding constant the other independent variables in the equation, equals $\beta_1 p_i(1-p_i)$. To use this formula, plug in your estimates of β_1 and $p_i(\hat{\beta}_1$ and $\hat{d}_i)$. As you can see, the marginal impact of x does indeed depend on the value of $\hat{d}_i$.
3. *Use a rough estimate of 0.25.* The previous two methods are reasonably accurate, but they’re hardly very handy. However, if you plug $\hat{d}_i = 0.5$ into $\beta_1 p_i(1-p_i)$, you get the much more useful result that if you multiply a logit coefficient by 0.25, you’ll get an equivalent linear probability model coefficient.

On balance, what do we recommend? For all situations except those requiring precise accuracy, we find ourselves gravitating toward the third approach. To get a rough approximation of the economic meaning of a logit coefficient, multiply by 0.25 (or, equivalently, divide by 4). Remember, however, that the dependent variable in question still is the probability that $\hat{d}_i = 1$.

Measuring the overall fit also is not straightforward. Recall from section 8.5 that since the functional form of the dependent variable has been changed, R^2 should not be used to compare the fit of a logit with an otherwise comparable linear probability model. In addition, don’t forget the general faults inherent in using R^2 with equations with dummy dependent variables. Our suggestion is to use the mean percentage of correct predictions, $\bar{R}_p^2$, from section 13.1.

To get some practice interpreting logit estimates, let’s estimate a logit on the same women’s labor force participation data that we used in the previous section. The OLS linear probability model estimate of that model, equation (13.5), was:

$$\hat{d}_i = -0.28 - 0.38m_i + 0.09s_i$$

$(0.15) \qquad (0.03)$
 $N = 30 \qquad \bar{R}^2 = 0.32 \qquad \bar{R}_p^2 = 0.81$

(13.5)

where: $d_i = 1$ if the i th woman is in the labor force, 0 otherwise

$m_i = 1$ if the i th woman is married and 0 otherwise

s_i = the number of years of schooling of the i th woman

If we estimate a logit on the same data (from Table 13.1) and the same independent variables, we obtain:

$$l: pr(\hat{d}_1 = 1) = -5.90 - 2.59m_i + 0.69s_i$$

$(1.18) \qquad (0.32)$
 $N = 30 \qquad \bar{R}_p^2 = 0.81 \qquad iterations = 5$

(13.6)

Let's compare equations (13.5) and (13.12). As expected, the signs and general significance of the slope coefficients are the same. Even if we divide the logit coefficients by 4, as suggested earlier, they still are larger than the linear probability model coefficients. Despite these differences, the overall fits are comparable, especially after taking account of the different dependent variables and estimation techniques. In this example, then, the two estimation procedures differ mainly in that the logit does not produce $\hat{d}_i$ s outside the range of 0 and 1.

However, if the size of the sample in this example is too small for a linear probability model, it certainly is too small for a logit, making any in-depth analysis of equation (13.12) problematic. Instead, we're better off finding an example with a much larger sample.

A more complete example of the use of the binomial logit

For a more complete example of the binomial logit, let's look at a model of the probability of passing the California State Department of Motor Vehicles drivers' license test. To obtain a license, each driver must pass a written and a behind-the-wheel test.

Even though the tests are scored from 0 to 100, all that matters is that you pass and get your license.

Since the written test requires some boning up on traffic and safety laws, driving students have to decide how much time to spend studying. If they don't study enough, they waste time because they have to retake the test. If they study too much, however, they also waste time, because there's no bonus for scoring above the minimum, especially since there is no evidence that doing well on the written test has much to do with driving well after the test (this, of course, might be worth its own econometric study).

Recently, two students decided to collect data on test takers in order to build an equation explaining whether someone passed the Department of Motor Vehicles written test. They hoped that the model, and in particular the estimated coefficient of study time, would help them decide how much time to spend studying for the test. (Of course, it took more time to collect the data and run the model than it would have taken to memorize the entire traffic code, but that's another story.)

After reviewing the literature, choosing variables, and hypothesizing signs, the students realized that the appropriate functional form was a binomial logit because their dependent variable was a dummy variable:

$$d_i = \begin{cases} 1 & \text{if the } i\text{th test taker passed the test on the first try} \\ 0 & \text{if the } i\text{th test taker failed the test on the first try} \end{cases}$$

They chose four independent variables (all with positive expected coefficients):

a_i = the age of the i th test taker

h_i = the number of hours the i th test taker studied (usually less than one hour!)

e_i = a dummy variable equal to 1 if the i th test taker's primary language was English, 0 otherwise

c_i = a dummy variable equal to 1 if the i th test taker had any college education, 0 otherwise

After collecting data from 480 test takers, the students estimated the following equation:

$$l: pr(\hat{d}_i = 1) = -1.18 + \underset{(0.009)}{0.011a_i} + \underset{(0.54)}{2.70h_i} + \underset{(0.34)}{1.62e_i} + \underset{(0.99)}{3.97c_i} \quad (13.13)$$

$N = 480 \quad \bar{R}_p^2 = 0.74 \quad iterations = 5$

Note how similar these results look to a typical linear regression result. All the estimated coefficients have the expected signs, and all but one appear to be significantly different from 0. Remember that the logit coefficients need to be divided by 4 to get meaningful estimates of the impact of the independent variables on the probability of passing the test. Note that $\bar{R}_p^2$ is .74, indicating that the equation correctly explained almost three quarters of the sample based on nothing but the four variables in equation (13.13).

And what about the two students? Did the equation help them? How much did they end up deciding to study? They found that given their ages, their college education, and their English-speaking backgrounds, the expected value of $\hat{d}_i$ for each of them was quite high, even if h_i was set equal to 0. So what did they actually do? They studied for a half hour “just to be on the safe side” and passed with flying colors, having devoted more time to passing the test than anyone else in the history of the state.

13.3 Other dummy dependent variable techniques

Although the binomial logit is the most frequently used estimation technique for equations with dummy dependent variables, it's by no means the only one. In this section, we'll mention two alternatives, but our main goal is to briefly describe these estimation techniques, not to cover them in any detail.¹⁰³

¹⁰³ For more, see G. S. Maddala, *Limited Dependent Variables and Qualitative Variables in Econometrics* (Cambridge: Cambridge University Press, 1983) and T. Amemiya, “Qualitative Response Models: A Survey,” *Journal of Economic Literature*, Vol. 19, pp. 1483–1536. These surveys also cover additional techniques, like the Tobit model, that are useful with bounded dependent variables or other special situations.

The binomial probit model

The **binomial probit model** is an estimation technique for equations with dummy dependent variables that avoids the unboundedness problem of the linear probability model by using a variant of the cumulative normal distribution.

$$p_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_i} e^{-s^2/2} ds \quad (13.14)$$

where p_i = the probability that the dummy variable $d_i = 1$

$$Z_i = a + \beta_1 x_{1i} + \beta_2 x_{2i}$$

s = a standardized normal variable

As different as this probit looks from the logit that we examined in the previous section, it can be rewritten to look quite familiar:

$$z_i = \Phi^{-1}(p_i) = a + \beta_1 x_{1i} + \beta_2 x_{2i} \quad (13.15)$$

where Φ^{-1} is the inverse of the normal cumulative distribution function. Probit models typically are estimated by applying maximum likelihood techniques to the model in the form of equation (13.14), but the results often are presented in the format of equation (13.15).

The fact that both the logit and the probit are cumulative distribution functions means that the two have similar properties. For example, a graph of the probit looks almost exactly like the logit in Figure 13.2. In addition, the probit has the same requirement of a fairly large sample before hypothesis testing becomes meaningful. Finally, $\bar{R}^2$ continues to be of questionable value as a measure of overall fit.

For an example of a probit, let's estimate one using the same women's labor force participation data employed in the previous logit and linear probability examples (standard errors in parentheses):

$$\hat{z}_i = \Phi^{-1}(\hat{p}_i) = -3.44 - \underset{(0.62)}{1.44}m_i + \underset{(0.17)}{0.40}s_i \quad (13.16)$$

$$N = 30 \quad \bar{R}_p^2 = 0.81 \quad \text{iterations} = 4$$

Compare this result with equation (13.12) from the previous section. Note that except for a slight difference in the scale of the coefficients, the logit and probit models provide virtually identical results in this example.

Multinomial models

In many cases, there are more than two qualitative choices available. In some cities, for instance, a commuter has a choice of car, bus, or subway for the trip to work. How could we build and estimate a model of choosing from more than two alternatives?

One answer is to hypothesize that choices are made sequentially and to model a multichoice decision as a series of binary decisions. For example, we might hypothesize that the commuter would first decide whether to drive to work, and we could build a binary model of car versus public transportation. For those commuters who choose public transportation, the next step would be to choose whether to take the bus or the subway, and we could build a second binary model of that choice. This method, called a **sequential binary logit**, is cumbersome and at times unrealistic, but it does allow a researcher to use a binary technique to model an inherently multichoice decision.

If a decision among more than two alternatives truly is made simultaneously, then the sequential binary logit can't be used. There are a number of alternative estimation procedures that are appropriate in this situation, but unfortunately they are beyond the scope of this text.¹⁰⁴

¹⁰⁴ These alternative procedures include the multinomial logit and the ordered logit. See William H. Greene, *Econometric Analysis* (Boston: Pearson Education, 2012), pp. 803–806 and pp. 824–827.

SUMMARY

1. A linear probability model is a linear-in-the-coefficients equation used to explain a dummy dependent variable (d_i). $\hat{d}_i$ is an estimate of the probability that d_i equals 1.
2. The estimation of a linear probability model with OLS faces at least three major exercises:
 - a. $\bar{R}^2$ is not an accurate measure of overall fit.
 - b. The expected value of $\hat{d}_i$ is not limited by 0 and 1.
 - c. The error term is neither homoskedastic nor normally distributed.
3. When measuring the overall fit of equations with dummy dependent variables, an alternative to $\bar{R}^2$ is $\bar{R}_p^2$, the average percentage of the observations in the sample that a particular estimated equation would have explained correctly.
4. The binomial logit is an estimation technique for equations with dummy dependent variables that avoids the unboundedness problem of the linear probability model by using a variant of the cumulative logistic function:

$$l: pr(d_i = 1) = \ln\left(\frac{p_i}{[1 - p_i]}\right) = a + \beta_1 x_{1i} + \beta_2 x_{2i}$$

5. The binomial logit is best estimated using the maximum likelihood (ML) technique and a large sample. A slope coefficient from a logit measures the impact of a one-unit increase of the independent variable in question (holding the other explanatory variables constant) on the log of the odds of a given choice.
6. The binomial probit model is an estimation technique for equations with dummy dependent variables that uses the cumulative normal distribution function. The binomial probit has properties quite similar to those of the binomial logit.

CHAPTER 14. SIMULTANEOUS EQUATIONS

The most important models in economics and business are simultaneous in nature. Supply and demand, for example, is obviously simultaneous. To study the demand for chicken without also looking at the supply of chicken is to take a chance on missing important linkages and thus making significant mistakes. Virtually all the major approaches to macroeconomics, from Keynesian aggregate demand models to rational expectations schemes, are inherently simultaneous. Even models that appear to be inherently single equation in nature often turn out to be much more simultaneous than you might think. The price of housing, for instance, is dramatically affected by the level of economic activity, the prevailing rate of interest in alternative assets, and a number of other simultaneously determined variables.

All this wouldn't mean much to econometricians if it weren't for the fact that the estimation of simultaneous equations systems with OLS causes a number of difficulties that aren't encountered with single equations. Most important, Classical assumption 3, which states that all explanatory variables should be uncorrelated with the error term, is violated in simultaneous models. Mainly because of this, OLS coefficient estimates are biased in simultaneous models. As a result, an alternative estimation procedure called **Two - Stage Least Squares** (TSLS) usually is employed in such models instead of OLS.

You're probably wondering why we've waited until now to discuss simultaneous equations if they're so important in economics and if OLS encounters bias when estimating them. The answer is that the simultaneous estimation of an equation changes every time the specification of any equation in the entire system is changed, so a researcher must be well equipped to deal with specification exercises like those of previous chapters. As a result, it does not make sense to learn how to estimate a simultaneous system until you are fairly adept at estimating a single equation.

14.1 Structural and reduced-form equations

Before we can study the exercises encountered in the estimation of simultaneous equations, we need to introduce a few concepts.

The nature of simultaneous equations systems

Which came first, the chicken or the egg? This question is impossible to answer satisfactorily because chickens and eggs are *jointly determined*; there is a two-way causal relationship between the variables. The more eggs you have, the more chickens you'll get, but the more chickens you have, the more eggs you'll get.¹⁰⁵ More realistically, the economic world is full of the kind of *feedback effects* and *dual causality* that require the application of simultaneous equations. Besides the supply and demand and simple macroeconomic model examples mentioned previously, we could talk about the dual causality of population size and food supply, the joint determination of wages and prices, or the interaction between foreign exchange rates and international trade and capital flows. In a typical econometric equation:

$$y_t = a + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t \quad (14.1)$$

a simultaneous system is one in which y clearly has an effect on at least one of the x s in addition to the effect that the x s have on y .

Such topics are usually modeled by distinguishing between variables that are simultaneously determined (the y s, called **endogenous variables**) and those that are not (the x s, called **exogenous variables**):

$$y_{1t} = \alpha_0 + \alpha_1 y_{2t} + \alpha_2 x_{1t} + \alpha_3 x_{2t} + u_{1t} \quad (14.2)$$

$$y_{2t} = \beta_0 + \beta_1 y_{1t} + \beta_2 x_{3t} + \beta_3 x_{2t} + u_{2t} \quad (14.3)$$

¹⁰⁵ This also depends on how hungry you are, which is a function of how hard you're working, which depends on how many chickens you have to take care of. (Although this chicken/egg example is simultaneous in an annual model, it would not be truly simultaneous in a quarterly or monthly model because of the time lags involved.)

For example, y_1 and y_2 might be the quantity and price of chicken (respectively), x_1 the income of the consumers, x_2 the price of beef (beef is a substitute for chicken in both consumption and production), and x_3 the price of chicken feed. With these definitions, equation (14.2) would characterize the behavior of consumers of chickens and equation (14.3) the behavior of suppliers of chickens. These behavioral equations are also called *structural equations*.

Structural equations characterize the underlying economic theory behind each endogenous variable by expressing it in terms of both endogenous and exogenous variables. Researchers must view them as an entire system in order to see all the feedback loops involved. For example, the y s are jointly determined, so a change in y_1 will cause a change in y_2 , which will in turn cause y_1 to change *again*. Contrast this feedback with a change in x_1 , which will not eventually loop back and cause x_1 to change again. The α s and the β s in the equation are *structural coefficients*, and hypotheses should be made about their signs just as we did with the regression coefficients of single equations.

Note that a variable is endogenous because it is jointly determined, not just because it appears in both equations. That is, x_2 , which is the price of beef but could be another factor beyond our control, is in both equations but is still exogenous in nature because it is not simultaneously determined within the chicken market. In a large general equilibrium model of the entire economy, however, such a price variable would also likely be endogenous.

How do you decide whether a particular variable should be endogenous or exogenous? Some variables are almost always exogenous (the weather, for example), but most others can be considered either endogenous or exogenous, depending on the number and characteristics of the other equations in the system. Thus, the distinction between endogenous and exogenous variables usually depends on how the researcher defines the scope of the research project.

Sometimes, lagged endogenous variables appear in simultaneous systems, usually when the equations involved are

dynamic models (described in chapter 12). Be careful! Such lagged endogenous variables are not simultaneously determined in the current time period. They thus have more in common with exogenous variables than with nonlagged endogenous variables.

To avoid exercises, we'll define the term **predetermined variable** to include all exogenous variables and lagged endogenous variables. "Predetermined" implies that exogenous and lagged endogenous variables are determined outside the system of specified equations or prior to the current period. Endogenous variables that are not lagged are not predetermined, because they are jointly determined by the system in the current time period. Therefore, econometricians tend to speak in terms of endogenous and predetermined variables when discussing simultaneous equations systems.

Let's look at the specification of a simple supply and demand model, say for the "cola" soft-drink industry:

$$Q_{Dt} = \alpha_0 + \alpha_1 P_t + \alpha_2 x_{1t} + \alpha_3 x_{2t} + u_{Dt} \quad (14.4)$$

$$Q_{St} = \beta_0 + \beta_1 P_t + \beta_2 x_{3t} + u_{St} \quad (14.5)$$

$$Q_{St} = Q_{Dt} \text{ (equilibrium condition)} \quad (14.6)$$

where Q_{Dt} = the quantity of cola demanded in time period t

Q_{St} = the quantity of cola supplied in time period t

P_t = the price of cola in time period t

x_{1t} = dollars of advertising for cola in time period t

x_{2t} = another "demand-side" exogenous variable (e.g., income or the prices or advertising of other drinks)

x_{3t} = a "supply-side" exogenous variable (e.g., the price of artificial flavors or other factors of production)

u_t = classical error terms (each equation has its own error term, subscripted "D" and "S" for demand and supply)

In this case, price and quantity are simultaneously determined, but price, one of the endogenous variables, is not on the left side of any of the equations. It's incorrect to assume automatically that the endogenous variables are those that appear on the left side of at

least one equation; in this case, we could have just as easily written equation (14.5) with price on the left side and quantity supplied on the right side, as we did in the chicken example in equations (14.2) and (14.3). Although the estimated coefficients would be different, the underlying relations would not. Note also that there must be as many equations as there are endogenous variables. In this case, the three endogenous variables are Q_D , Q_S , and P .

What would be the expected signs for the coefficients of the price variables in equations (14.4) and (14.5)? We'd expect price to enter negatively in the demand equation but to enter positively in the supply equation. The higher the price, after all, the less quantity will be demanded, but the more quantity will be supplied. These signs would result in the typical supply and demand diagram that we're all used to. Look at equations (14.4) and (14.5) again, and note that they would be identical except for the different predetermined variables. What would happen if we accidentally put a supply-side predetermined variable in the demand equation or vice versa? We'd have a very difficult time identifying which equation was which, and the expected signs for the coefficients of the endogenous variable P would become ambiguous. As a result, we must take care when specifying the structural equations in a system.

Simultaneous systems violate Classical assumption 3

Recall from chapter 5 that Classical assumption 3 states that the error term and each explanatory variable must be uncorrelated with each other. If there is such a correlation, then the OLS regression estimation program is likely to attribute to the explanatory variable variations in the dependent variable that are actually being caused by variations in the error term. The result will be biased estimates.

To see why simultaneous equations violate the assumption of independence between the error term and the explanatory variables, look again at a simultaneous system, equations (14.2) and (14.3) (repeated with directional errors):

$$\begin{array}{c} \uparrow \\ y_{1t} = \alpha_0 + \alpha_1 y_{2t} + \alpha_2 x_{1t} + \alpha_3 x_{2t} + \epsilon_{1t} \end{array} \quad (14.2)$$

$$\begin{array}{c} \uparrow \\ y_{2t} = \alpha_0 + \beta_1 y_{1t} + \beta_2 x_{3t} + \beta_3 x_{2t} + \epsilon_{2t} \end{array} \quad (14.3)$$

Let's work through the system and see what happens when one of the error terms increases, holding everything else in the equations constant:

1. If ϵ_1 increases in a particular time period, y_1 will also increase due to equation (14.2).
2. If y_1 increases, y_2 will also rise¹⁰⁶ due to equation (14.3).
3. But if y_2 increases in equation (14.3), it also increases in equation (14.2) where it is an explanatory variable.

Thus, an increase in the error term of an equation causes an increase in an explanatory variable in the same equation: If ϵ_1 increases, y_1 increases, and then y_2 increases, violating the assumption of independence between the error term and the explanatory variables.

For a visual understanding of this, take a look at Figure 14.1. If the error term in the demand equation increases, then the demand curve will shift from D to D' , and both price and quantity will increase. Thus an increase in the error term will be correlated with an increase in an independent variable. We've violated Classical assumption III!

¹⁰⁶ This assumes that β_1 is positive. If β_1 is negative, Y_2 will decrease and there will be a negative correlation between ϵ_1 and Y_2 , but this negative correlation will still violate Classical Assumption III. Also note that both equations (14.2) and (14.3) could have Y_{1t} on the left side; if two variables are jointly determined, it doesn't matter which variable is considered dependent and which explanatory, because they are actually mutually dependent. We used this kind of simultaneous system in the cola model portrayed in equations (14.4) and (14.5).

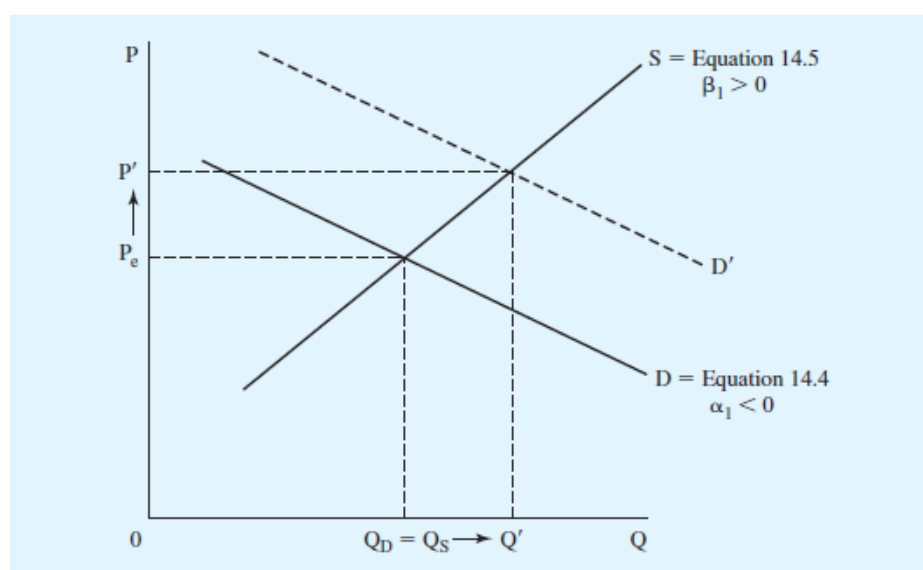


Figure 14.1 Supply and demand simultaneous equations

An example of simultaneous equations that jointly determine two endogenous variables is the supply and demand for a product. In this case, equation (14.4), the downward sloping demand function, and Equation 14.5, the upward-sloping supply function, intersect at the equilibrium price and quantity for this market.

This is not an isolated result that depends on the particular equations involved. Instead, this result works for other error terms, equations, and simultaneous systems. All that is required for the violation of Classical assumption III is that there be endogenous variables that are jointly determined in a system of simultaneous equations.

Reduced-form equations

An alternative way of expressing a simultaneous equations system is through the use of **reduced-form equations**, equations that express a particular endogenous variable solely in terms of an error term and all the predetermined (exogenous plus lagged endogenous) variables in the simultaneous system.

The reduced-form equations for the structural equations (14.2) and (14.3) would thus be:

$$y_{1t} = \pi_0 + \pi_1 x_{1t} + \pi_2 x_{2t} + \pi_3 x_{3t} + v_{1t} \quad (14.2)$$

$$y_{2t} = \beta_0 + \beta_1 y_{1t} + \beta_2 x_{3t} + \beta_3 x_{2t} + u_{2t} \quad (14.3)$$

where the v s are stochastic error terms and the s are called *reduced-form coefficients* because they are the coefficients of the predetermined variables in the reduced-form equations. Where do these reduced-form equations come from? If you substitute equation (14.3) into equation (14.2), solve for y_1 , and then regroup terms, you'll get equation (14.6).

Note that each reduced-form equation includes only one endogenous variable (the dependent variable) and that each equation has exactly the same set of predetermined variables. The reduced-form coefficients, such as π_1 and π_5 , are known as **impact multipliers** because they measure the impact on the endogenous variable of a one-unit increase in the value of the predetermined variable, after allowing for the feedback effects from the entire simultaneous system.

There are at least three reasons for using reduced-form equations:

1. Since the reduced-form equations have no inherent simultaneity, they do not violate Classical assumption III. Therefore, they can be estimated with OLS without encountering the problems discussed in this chapter.
2. The interpretation of the reduced-form coefficients as impact multipliers means that they have economic meaning and useful applications of their own. For example, if you wanted to compare a government spending increase with a tax cut in terms of the per-dollar impact in the first year, estimates of the impact multipliers (reduced-form coefficients or π s) would allow such a comparison.
3. Perhaps most importantly, reduced-form equations play a crucial role in the estimation technique most frequently used for simultaneous equations. This technique, Two-stage least squares, will be explained in section 14.3.

To conclude, let's return to the cola supply and demand model and specify the reduced-form equations for that model. (To test yourself, flip back to equations (14.4) and (14.5) and see if you can get the right answer before going on). Since the equilibrium

condition forces Q_D to be equal to Q_S , we need only two reduced-form equations:

$$Q_t = \pi_0 + \pi_1 x_{1t} + \pi_2 x_{2t} + \pi_3 x_{3t} + v_{1t} \quad (14.8)$$

$$P_t = \pi_4 + \pi_5 x_{1t} + \pi_6 x_{2t} + \pi_7 x_{3t} + v_{2t} \quad (14.9)$$

Even though P never appears on the left side of a structural equation, it's an endogenous variable and should be treated as such.

14.2 The Bias of ordinary least squares

The first six Classical assumptions must be met for OLS estimates to be BLUE; when an assumption is violated, we must determine which of the properties no longer holds. It turns out that applying OLS directly to the structural equations of a simultaneous system produces biased and inconsistent estimates of the coefficients. Such bias is called simultaneous equations bias or simultaneity bias.

Understanding simultaneity bias

Simultaneity bias refers to the fact that in a simultaneous system, the expected values of the OLS-estimated structural coefficients ($\hat{\beta}$ s) are not equal to the true (β s). We are therefore faced with the problem that in a simultaneous system:

$$E(\hat{\beta}) \neq \beta \quad (14.10)$$

Why does this simultaneity bias exist? Recall from section 14.1 that in simultaneous equations systems, the error terms (the ϵ s) tend to be correlated with the endogenous variables (the y s) whenever the y s appear as explanatory variables. Let's follow through what this correlation means (assuming positive coefficients for simplicity) in typical structural equations like (14.11) and (14.12):

$$y_{1t} = \beta_0 + \beta_1 y_{2t} + \beta_2 x_t + \epsilon_{1t} \quad (14.11)$$

$$y_{2t} = \alpha_0 + \alpha_1 y_{1t} + \alpha_2 z_t + \epsilon_{2t} \quad (14.12)$$

Since we cannot observe the error term (ϵ_1) and don't know when ϵ_{1t} is above average, it will appear as if every time y_1 is above average, y_2 is also above average (as long as α_1 is positive). As a result, the OLS estimation program will tend to attribute increases in y_1 caused by the error term ϵ_1 to y_2 , thus typically overestimating β_1 . This overestimation is simultaneity bias. If the error term is abnormally negative, y_{1t} will be less than it would have been otherwise, causing y_{2t} to be less than it would have been otherwise, and the computer program will attribute the decrease in y_1 to y_2 , once again causing us to overestimate β_1 (that is, induce upward bias).

Recall that the causation between y_1 and y_2 runs in both directions because the two variables are interdependent. As a result, β_1 , when estimated by OLS, can no longer be interpreted as the impact of y_2 on y_1 , holding x constant. Instead, $\hat{\beta}_1$ now measures some mix of the effects of the two endogenous variables on each other! In addition, consider β_2 . It's supposed to be the effect of x on y_1 holding y_2 constant, but how can we expect y_2 to be held constant when a change in y_1 takes place? As a result, there is potential bias in all the estimated coefficients in a simultaneous system.

What does this bias look like? It's possible to derive an equation for the expected value of the regression coefficients in a simultaneous system that is estimated by OLS. This equation shows that as long as the error term and any of the explanatory variables in the equation are correlated, then the coefficient estimates will be biased and inconsistent. In addition, it also shows that the bias will have the same sign as the correlation between the error term and the endogenous variable that appears as an explanatory variable in that error term's equation. Since that correlation is often positive in economic and business examples, the bias often will be positive, although the direction of the bias in any given situation will depend on the specific details of the structural equations and the model's underlying theory.

This does not mean that every coefficient from a simultaneous system estimated with OLS will be a bad approximation of the true

population coefficient. However, it's vital to consider an alternative to OLS whenever simultaneous equations systems are being estimated. Before we investigate the alternative estimation technique most frequently used (Two-Stage Least Squares), let's look at an example of simultaneity bias.

An Example of simultaneity bias

To show how the application of OLS to simultaneous equations estimation causes bias, we used a Monte Carlo experiment¹⁰⁷ to generate an example of such biased estimates. Since it's impossible to know whether any bias exists unless you also know the true β s, we arbitrarily picked a set of coefficients to be considered "true." We then stochastically generated data sets based on these "true" coefficients, and obtained repeated OLS estimates of these coefficients from the generated data sets. The expected value of these estimates turned out to be quite different from the true coefficient values, thus exemplifying the bias in OLS estimates of coefficients in simultaneous systems.

We used a supply and demand model as the basis for our example:

$$Q_t = \beta_0 + \beta_1 P_t + \beta_2 X_t + \epsilon_{Dt} \quad (14.4)$$

$$Q_t = \alpha_0 + \alpha_1 P_t + \alpha_2 Z_t + \epsilon_{St} \quad (14.5)$$

where: Q_t = the quantity demanded and supplied in time period t

P_t = the price in time period t

X_t = a "demand-side" exogenous variable, such as income

Z_t = a "supply-side" exogenous variable, such as weather

¹⁰⁷ 3. *Monte Carlo* experiments are computer-generated simulations that typically follow seven steps: 1. Assume a "true" model with specific coefficient values and an error term distribution. 2. Select values for the independent variables. 3. Select an estimating technique (usually OLS). 4. Create various samples of the dependent variable, using the assumed model, by randomly generating error terms from the assumed distribution; often, the number of samples created runs into the thousands. 5. Compute the estimates of the β s from the various samples using the estimating technique. 6. Summarize and evaluate the results. 7. Consider sensitivity analyses using different values, distributions, or estimating techniques.

ϵ_t = classical error terms (different for each equation)

The first step was to choose a set of true coefficient values that corresponded to our expectations for this model:

$$\beta_1 = -1 \quad \beta_2 = +1 \quad \alpha_1 = +1 \quad \alpha_2 = +1$$

In other words, we have a negative relationship between price and quantity demanded, a positive relationship between price and quantity supplied, and positive relationships between the exogenous variables and their respective dependent variables.

The next step was to randomly generate a number of data sets based on the true values. This also meant specifying some other characteristics of the data¹⁰⁸ before generating the different data sets (5,000 in this case).

The final step was to apply OLS to the generated data sets and to calculate the estimated coefficients of the demand equation (14.13). (Similar results were obtained for the supply equation.) The arithmetic means of the results for the 5,000 regressions were:

$$\hat{Q}_{Dt} = \hat{\beta}_0 - 0.37P_t + 1.84x_t \quad (14.5)$$

In other words, the expected value of $\hat{\beta}_1$ should have been -1.00, but instead it was roughly -0.37; the expected value of $\hat{\beta}_2$ should have been +1.00, but instead it was around 1.84:

$$E(\hat{\beta}_1) \approx -0.37 \neq -1.00$$

$$E(\hat{\beta}_2) \approx 1.84 \neq 1.00$$

This is simultaneity bias! As the diagram of the sampling distributions of the $\hat{\beta}$ s in Figure 14.2 shows, the OLS estimates of β_1 were almost never very close to -1.00, and the OLS estimates of β_2 were distributed over a wide range of values.

108

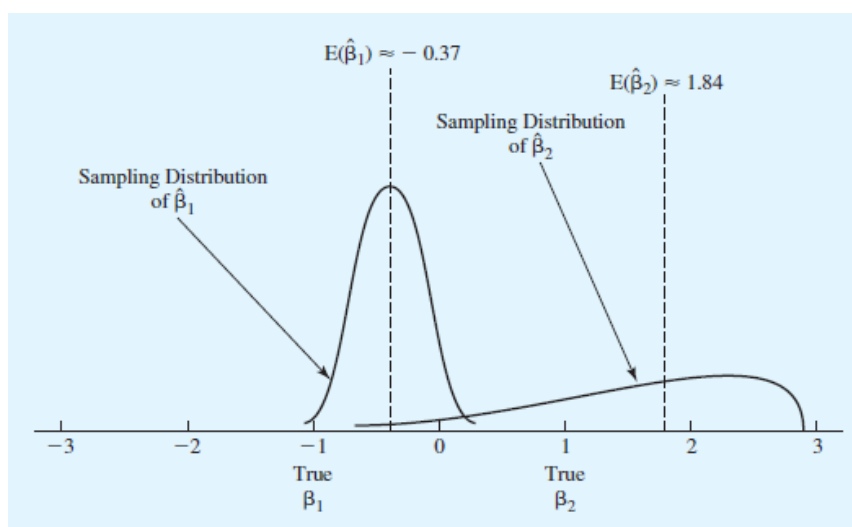


Figure 14.2 Sampling distributions showing simultaneity bias of OLS estimates

In the experiment in Section 14.2, simultaneity bias is evident in the distribution of the estimates of β_1 , which had a mean value of -0.37 compared with a true value of -1.00, and in the estimates of β_2 , which had a mean value of 1.84 compared with a true value of 1.00.

14.3 Two-stage least squares (2SLS)

How can we get rid of (or at least reduce) simultaneity bias? There are a number of estimation techniques that help mitigate simultaneity bias, but the most frequently used alternative to OLS is called Two-stage least squares (2SLS).

Since OLS encounters bias in the estimation of simultaneous equations mainly because such equations violate Classical assumption III, one solution to the problem is to try to avoid violating that assumption. The first step in doing this is to find a variable that is:

1. highly correlated with the endogenous variable, and
2. uncorrelated with the error term.

Such a variable is called an **instrumental variable**; it is highly correlated with the endogenous variable, but is uncorrelated with the error term. More generally, instrumental variables (IV) regression is

a method of avoiding the violation of Classical assumption III by producing predicted values of endogenous variables that can be substituted for the endogenous variables where they appear on the right-hand side of structural equations. These predicted values typically are produced by running an OLS equation to explain the endogenous variable as a function of one or more instrumental variables.

To see how this works, take a look at equation (14.16) in the following system:

$$y_{1t} = \beta_0 + \beta_1 y_{2t} + \beta_2 x_{1t} + \epsilon_{1t} \quad (14.16)$$

$$y_{2t} = \alpha_0 + \alpha_1 y_{1t} + \alpha_2 x_{2t} + \epsilon_{2t} \quad (14.17)$$

If we could find a variable (or variables) highly correlated with y_{2t} but uncorrelated with ϵ_{1t} , then we could produce a predicted value of y_{2t} by running an OLS regression with y_{2t} as a function of the instrumental variable(s). The fitted value $\hat{y}_{2t}$ will be uncorrelated with ϵ_{1t} (because it was produced using variables that are uncorrelated with ϵ_{1t}), so if we substitute $\hat{y}_{2t}$ for y_{2t} on the right side of equation (14.16), then we'll no longer violate Classical assumption III.

This approach avoids the violation of Classical assumption III, but it doesn't give us any insight into where to find appropriate instrumental variables (sometimes called instruments). How do we systematically find variables that are highly correlated with the endogenous variable but uncorrelated with the error term? For simultaneous equations systems, there's a straightforward answer. We use Two-Stage Least Squares.

What is Two-stage least squares?

Two-Stage Least Squares (2SLS) is a method of avoiding simultaneity bias by systematically creating variables to replace the endogenous variables where they appear as explanatory variables in simultaneous equations systems. The simplest form of 2SLS does this by running an OLS regression on the reduced form of every right-side endogenous variable and then using the $\hat{y}$ (or fitted value)

from the reduced-form estimated equation in place of the endogenous variable where it appears on the right side of a structural equation.

Why does 2SLS do this? Every predetermined variable in the simultaneous system is a candidate to be an instrumental variable for every endogenous variable, but if we choose only one instrumental variable, then we'll be throwing away information. To avoid this, 2SLS uses a linear combination of *all* the predetermined variables. We form this linear combination by running a regression for a given endogenous variable as a function of all the predetermined variables in the reduced-form equation to generate a predicted value of the endogenous variable. Thus the 2SLS two-step estimation procedure is:

STAGE ONE: *Run OLS on the reduced-form equations for each of the endogenous variables that appear as explanatory variables in the structural equations in the system.*

Since the predetermined (exogenous plus lagged endogenous) variables are uncorrelated with the reduced-form error term, the OLS estimates of the reduced-form coefficients (the $\hat{\pi}$ s) are unbiased. These $\hat{\pi}$ s can then be used to calculate estimates of the endogenous variables:

$$\hat{y}_{1t} = \hat{\pi}_0 + \hat{\pi}_1 x_{1t} + \hat{\pi}_2 x_{2t} \quad (14.18)$$

$$\hat{y}_{2t} = \hat{\pi}_3 + \hat{\pi}_4 x_{1t} + \hat{\pi}_5 x_{2t} \quad (14.19)$$

These $\hat{y}$ s then are used in place of the y s on the right-hand side of the structural equations.

STAGE TWO: *Substitute the reduced form $\hat{y}$ s for the y s that appear on the right side (only) of the structural equations, and then estimate these revised structural equations.*

That is, stage two consists of estimating the following equations with OLS:

$$\hat{y}_{1t} = \beta_0 + \beta_1 \hat{y}_{2t} + \beta_2 x_{1t} + u_{1t} \quad (14.20)$$

$$y_{2t} = \alpha_0 + \alpha_1 \hat{y}_{1t} + \alpha_2 x_{2t} + u_{2t} \quad (14.21)$$

Note that the dependent variables are still the original endogenous variables and that the substitutions are only for the endogenous variables where they appear on the right-hand side of the structural equations. This procedure produces consistent estimates of the coefficients of the structural equations.

Be careful! If second-stage equations such as equations (14.20) and (14.21) are estimated with OLS, the SE ($\hat{\beta}$)s will be incorrect, so be sure to use your computer's 2SLS estimation procedure.

This description of 2SLS can be generalized to any number of simultaneous structural equations. Each reduced-form equation has as explanatory variables every predetermined variable in the entire system of equations.

The OLS estimates of the reduced-form equations are used to compute the estimated values of all the endogenous variables that appear as explanatory variables in the structural equations. After substituting these fitted values for the original values of the endogenous independent variables, OLS is applied to each stochastic equation in the set of structural equations.¹⁰⁹

14.4 The Identification problem

Two-Stage Least Squares cannot be applied to an equation unless that equation is *identified*. Before estimating any equation in a simultaneous system, you therefore must address the identification problem. Once an equation is found to be identified, then it can be estimated with 2SLS, but if an equation is not identified (*underidentified*), then 2SLS cannot be used no matter how large the sample. Such

¹⁰⁹ Detailed example of simultaneous equation macroeconomic model for Tajikistan can be seen at Хакимова М.Ф. Динамическое моделирование макроэкономических процессов. – Душанбе: «Эр-граф», 2013, 200 стр.

underidentified equations can be estimated with OLS, but OLS estimates of underidentified equations are difficult to interpret. It's important to point out that an equation being identified (and therefore capable of being estimated with 2SLS) does not ensure that the resulting 2SLS estimates will be good ones. The question being asked is not how good the 2SLS estimates will be but whether the 2SLS estimates can be obtained at all.

What Is the Identification Problem?

Identification is a precondition for the application of 2SLS to equations in simultaneous systems; a structural equation is identified only when enough of the system's predetermined variables are omitted from the equation in question to allow that equation to be distinguished from all the others in the system. Note that one equation in a simultaneous system might be identified and another might not. How can we have equations that we cannot identify? To see how, let's consider a supply and demand simultaneous system in which only price and quantity are specified:

$$Q_{Dt} = \alpha_0 + \alpha_1 P_t + \epsilon_{Dt} \quad (\text{demand}) \quad (14.22)$$

$$Q_{St} = \beta_0 + \beta_1 P_t + \epsilon_{St} \quad (\text{supply}) \quad (14.23)$$

where $Q_{Dt} = Q_{St}$

Although we've labeled one equation as the demand equation and the other as the supply equation, the computer will not be able to identify them from the data because the right-side and the left-side variables are exactly the same in both equations; without some predetermined variables included to differentiate these two equations, it would be impossible to distinguish supply from demand.

What if we added a predetermined variable like weather (W) to the supply equation for an agricultural product? Then, equation (14.23) would become:

$$Q_{St} = \beta_0 + \beta_1 P_t + \beta_2 W_t + \epsilon_{St} \quad (14.24)$$

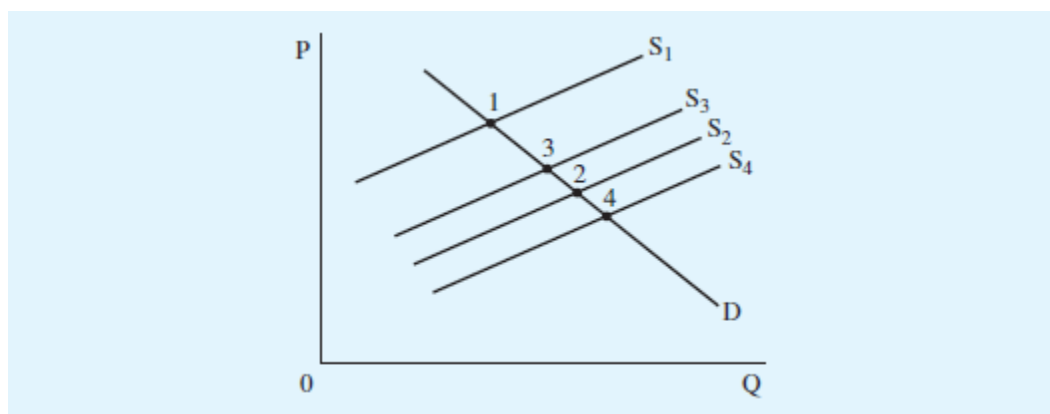


Figure 14.3 A Shifting supply curve allows the identification of the demand curve

If the supply curve shifts but the demand curve does not, then we move along the demand curve, which allows us to identify and estimate the demand curve (but not the supply curve).

In such a circumstance, every time W changed, the supply curve would shift, but the demand curve would not, so that eventually we would be able to collect a good picture of what the demand curve looked like.

Figure 14.3 demonstrates this. Given four different values of W , we get four different supply curves, each of which intersects with the constant demand curve at a different equilibrium price and quantity (intersections 1 - 4). These equilibria are the data that we would be able to observe in the real world and are all that we could feed into the computer. As a result, we would be able to identify the demand curve because we left out at least one predetermined variable. When this predetermined variable changed, but the demand curve didn't, the supply curve shifted so that quantity demanded moved along the demand curve and we gathered enough information to estimate the coefficients of the demand curve. The supply curve, on the other hand, remains as much a mystery as ever because its shifts give us no clue whatsoever about its shape. In essence, the demand curve was identified by the predetermined variable that was included in the system but excluded from the demand equation. The supply curve is not identified because there is no such excluded predetermined variable for it.

Even if we added W to the demand curve as well (which wouldn't make sense from a theoretical point of view), that would not identify the supply curve. In fact, if we had W in both equations, the two would be identical again, and although both would shift when W changed, those shifts would give us no information about either curve! As illustrated in Figure 14.4, the observed equilibrium prices and quantities would be almost random intersections describing neither the demand nor the supply curve. That is, the shifts in the supply curve are the same as before, but now the demand curve also shifts with W . In this case, it's not possible to identify either the demand curve or the supply curve.

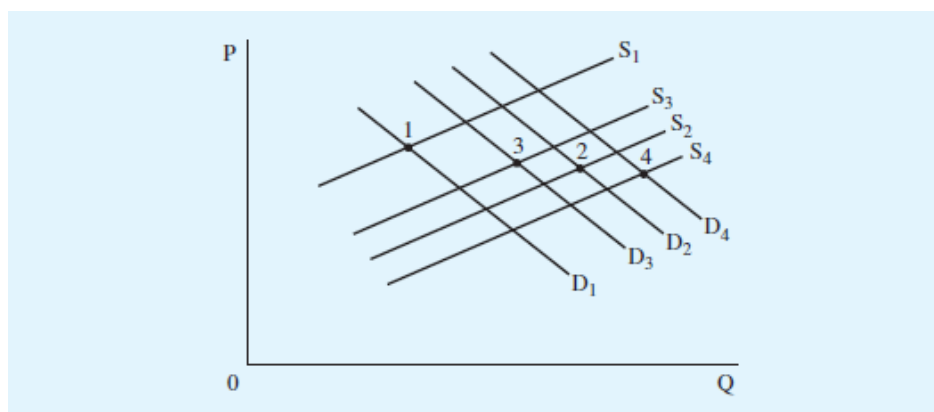


Figure 14.4 If Both the Supply Curve and the Demand Curve Shift, Neither Curve Is Identified

If both the supply curve and the demand curve shift in response to the same variable, then we move from one equilibrium to another, and the resulting data points identify neither curve. To allow such an identification, at least one predetermined variable must cause one curve to shift while allowing the other to remain constant.

The way to identify both curves is to have at least one predetermined variable in each equation that is not in the other, as in:

$$Q_{Dt} = \alpha_0 + \alpha_1 P_t + \alpha_2 X_t + \epsilon_{Dt} \quad (14.25)$$

$$Q_{St} = \beta_0 + \beta_1 P_t + \beta_2 W_t + \epsilon_{St} \quad (14.26)$$

Now when W changes, the supply curve shifts, and we can identify the demand curves from the data on equilibrium prices and

quantities. When x changes, the demand curve shifts, and we can identify the supply curve from the data.

To sum, identification is a precondition for the application of 2SLS to equations in simultaneous systems. A structural equation is identified only when the predetermined variables are arranged within the system so as to allow us to use the observed equilibrium points to distinguish the shape of the equation in question. Most systems are quite a bit more complicated than the previous ones, however, so econometricians need a general method by which to determine whether equations are identified. The method typically used is the *order condition* of identification.

The Order condition of identification

The order condition is a systematic method of determining whether a particular equation in a simultaneous system has the potential to be identified. If an equation can meet the order condition, then it is identified in all but a very small number of cases. We thus say that the order condition is a necessary but not sufficient condition of identification.¹¹⁰

What is the order condition? Recall that we have used the phrases *endogenous* and *predetermined* to refer to the two kinds of variables in a simultaneous system. Endogenous variables are those that are jointly determined in the system in the current time period. Predetermined variables are exogenous variables plus any lagged endogenous variables that might be in the model.

For each equation in the system, we need to determine:

1. The number of predetermined (exogenous plus lagged endogenous) variables in the entire simultaneous system.
2. The number of slope coefficients estimated in the equation in question.

THE ORDER CONDITION: *A necessary condition for an equation to be identified is that the number of predetermined*

¹¹⁰ A sufficient condition for an equation to be identified is called the *rank condition*, but most researchers examine just the order condition before estimating an equation with 2SLS. These researchers let the computer estimation procedure tell them whether the rank condition has been met (by its ability to apply 2SLS to the equation). Those interested in the rank condition are encouraged to consult an advanced econometrics text.

(exogenous plus lagged endogenous) variables in the system be greater than or equal to the number of slope coefficients in the equation of interest.

In equation form, a structural equation meets the order condition if:

The number of predetermined variables $\geq$ (in the simultaneous system)	The number of slope coefficients (in the equation)

Example of the application of the order condition

Let's apply the order condition to some of the simultaneous equations systems encountered in this chapter. For example, consider once again the cola supply and demand model of Section 14.1:

$$Q_{Dt} = \alpha_0 + \alpha_1 P_t + \alpha_2 x_{1t} + \alpha_3 x_{2t} + u_{Dt} \quad (14.27)$$

$$Q_{St} = \beta_0 + \beta_1 P_t + \beta_2 x_{3t} + u_{St} \quad (14.28)$$

$$Q_{Dt} = Q_{St} \quad (14.29)$$

Equation (14.27) is identified by the order condition because the number of predetermined variables in the system (three, x_1 , x_2 , and x_3) is equal to the number of slope coefficients in the equation (three: α_1 , α_2 , and α_3). This particular result (equality) implies that equation (14.27) is *exactly identified* by the order condition. Equation (14.28) is also identified by the order condition because there are still three predetermined variables in the system, but there are only two slope coefficients in the equation; this condition implies that equation (14.28) is *overidentified*. 2SLS can be applied to equations that are identified (which includes exactly identified and overidentified), but not to equations that are underidentified.

SUMMARY

1. Most economic and business models are inherently simultaneous because of the dual causality, feedback loops, or joint determination of particular variables. These simultaneously determined variables are called endogenous, and nonsimultaneously determined variables are called exogenous.
2. A structural equation characterizes the theory underlying a particular variable and is the kind of equation we have used to date in this text. A reduced-form equation expresses a particular endogenous variable solely in terms of an error term and all the predetermined (exogenous and lagged endogenous) variables in the simultaneous system.
3. Simultaneous equations models violate Classical assumption III that the error term is uncorrelated with the explanatory variables. This occurs because of the feedback effects of the endogenous variables. For example, an unusually high observation of an equation's error term works through the simultaneous system and eventually causes a high (with positive coefficients) value for the endogenous variables that appear as explanatory variables in the equation in question, thus violating Classical assumption III.
4. If OLS is applied to the coefficients of a simultaneous system, the resulting estimates are biased and inconsistent. This occurs mainly because of the violation of Classical assumption III; the OLS regression package attributes to explanatory variables changes in the dependent variable actually caused by the error term (with which the explanatory variables are correlated).
5. Two-stage least squares is a method of estimating simultaneous equations systems. It works by systematically using the reduced-form equations of the system to create substitutes for the endogenous variables that are independent of the error terms. It then estimates the structural equations of the system with these substitutes replacing the endogenous variables where they appear as explanatory variables.
6. Two-stage least squares estimates are consistent but biased. Luckily, the expected bias due to 2SLS usually is smaller than the

expected bias due to OLS. If the fit of the reduced-form equations is poor, then 2SLS will not work very well.

7. 2SLS cannot be applied to an equation that's not identified. A necessary (but not sufficient) requirement for identification is the order condition, which requires that the number of predetermined variables in the system be greater than or equal to the number of slope coefficients in the equation of interest. Sufficiency usually is determined by the ability of 2SLS to estimate the coefficients.

CHAPTER 15. FORECASTING

Of the uses of econometrics outlined in Chapter 1, we have discussed forecasting the least. Accurate forecasting is vital to successful planning, so it's the primary application of many business and governmental uses of econometrics. For example, manufacturing firms need sales forecasts, banks need interest rate forecasts, and governments need unemployment and inflation rate forecasts.

To many business and government leaders, the words *econometrics* and *forecasting* mean the same thing. Such a simplification gives econometrics a bad name because many econometricians overestimate their ability to produce accurate forecasts, resulting in unrealistic claims and unhappy clients.

Some of their clients would probably applaud the nineteenth century New York law (luckily unenforced but apparently also unrepealed) that provides that persons "pretending to forecast the future" shall be liable to a \$250 fine and/or six months in prison.¹ Although many econometricians might wish that such consultants would call themselves "futurists" or "soothsayers," it's impossible to ignore the importance of econometrics in forecasting in today's world.

The ways in which the prediction of future events is accomplished are quite varied. At one extreme, some forecasters use models with thousands of equations. At the other extreme, quite accurate forecasts can be created with nothing more than a good imagination and a healthy dose of self-confidence.

Unfortunately, it's unrealistic to think we can cover even a small portion of the topic of forecasting in one short chapter. Indeed, there are a number of excellent books and journals on this subject alone. Instead, this chapter is meant to be a brief introduction to the use of econometrics in forecasting.

We will begin by using simple linear equations and then move on to investigate a few more complex forecasting situations. The chapter concludes with an introduction to a technique, called ARIMA, that calculates forecasts entirely from past movements of the dependent variable without the use of any independent variables at all. ARIMA is almost universally used as a benchmark forecast,

so it's important to understand even though it's not based on economic theory.

15.1 What is forecasting?

In general, forecasting is the act of predicting the future; in econometrics, forecasting is the estimation of the expected value of a dependent variable for observations that are outside the sample data set. In most forecasts, the values being predicted are for time periods in the future, but cross-sectional predictions of values for countries or people not in the sample are also common. To simplify terminology, the words prediction and forecast will be used interchangeably in this chapter. (Some authors limit the use of the word forecast to out-of-sample prediction for a time series.)

We've already encountered an example of a forecasting equation. Think back to the weight/height example of Section 2.2 and recall that the purpose of that model was to guess the weight of a male customer based on his height. In that example, the first step in building a forecast was to estimate

Equation (2.13):

$$\text{estimated weight} = 60.40 + 0.638 \times \text{height}$$

That is, we estimated that a customer's weight on average equaled a base of 60.40 kg plus 0.638 kg for each centimeter of height over 1.5 meters. To actually make the prediction, all we had to do was to substitute the height of the individual whose weight we were trying to predict into the estimated equation. For a male who is 1 meter 80 centimeters tall, for example, we'd calculate:

$$\text{predicted weight} = 60.40 + 0.638 \times 1.80 \quad (15.1)$$

or

$$60.40 + 0.638 \times 1.80 = 61.548$$

The weight-guessing equation is a specific example of using a single linear equation to predict or forecast. Our use of such an equation to make a forecast can be summarized into two steps:

1. *Specify and estimate an equation that has as its dependent variable the item that we wish to forecast.* We obtain a forecasting equation by specifying and estimating an equation for the variable we want to predict:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{1t} + \hat{\beta}_2 x_{2t} \quad (t = 1, 2, \dots, T) \quad (15.2)$$

Such specification and estimation have been the topics of the first 14 chapters of this book. The use of $(t = 1, 2, \dots, T)$ to denote the sample size is fairly standard for time-series forecasts (t stands for “time”).

2. *Obtain values for each of the independent variables for the observations for which we want a forecast and substitute them into our forecasting equation.* To calculate a forecast with equation (15.2), this would mean finding values for period $T + 1$ for x_1 and x_2 and substituting them into the equation:

$$\hat{y}_{T+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{1T+1} + \hat{\beta}_2 x_{2T+1} \quad (15.3)$$

What is the meaning of this $\hat{y}_{T+1}$? It is a prediction of the value that y will take in observation $T + 1$ (outside the sample) based upon our values of x_{1T+1} and x_{2T+1} and based upon the particular specification and estimation that produced equation (15.2).

To understand these steps more clearly, let’s look at two applications of this forecasting approach.

EXAMPLE 15.1

Let’s return to the chicken demand model, equation (10.14) on page 286, to see how well that equation forecasts aggregate per capita chicken consumption:

$$\begin{aligned} \hat{Y}_t = 27.7 - \underset{(0.03)}{0.11} PC_t + \underset{(0.02)}{0.03} PB_t + \underset{(0.01)}{0.23} YD_t \\ t = -3.38 \quad +1.86 \quad +15.7 \end{aligned} \quad (10.14)$$

$$\bar{R}^2 = 0.9904 \quad N = 29 \quad DW = 0.99$$

where: Y_t = per capita chicken consumption (in kg) in year t

PC_t = the price of chicken (in somoni per kg) in year t

PB_t = the price of beef (in somoni per kg) in year t

YD_t = per capita disposable income (in hundreds of somoni) in year t

To make these forecasts as realistic as possible, we held out the last three available years from the data set used to estimate equation (10.14). We'll thus be able to compare the equation's forecasts with what actually happened. To forecast with the model, we first obtain values for the three independent variables and then substitute them into equation (10.14). For 2003, $PC = 34.1$, $PB = 374.6$, and $YD = 280.2$, giving us:

$$\hat{y}_{2003} = 27.7 - 0.11(34.1) + 0.03(374.6) + 0.23(280.2) = 99.63 \quad (15.4)$$

Continuing on through 2005, we end up with:¹¹¹

Year	Forecast	Actual	Percent Error
2003	99.63	99.63	4.2
2004	105.06	98.58	6.6
2005	107.44	100.60	6.8

How does the model do? Well, forecasting accuracy, like beauty, is in the eye of the beholder, and there are many ways to answer the question.¹¹² The simplest method is to take the mean of the percentage errors (in absolute value), an approach called, not surprisingly, **the mean absolute percentage error (MAPE)** method. The MAPE for our forecast is 5.9 percent.

The most popular alternative method of evaluating forecast accuracy is the **root mean square error criterion (RMSE)**, which is calculated by squaring the forecasting error for each time period, averaging these squared amounts, and then taking the square root of this average. One advantage of the RMSE is that it penalizes large

¹¹¹ The rest of the actual values are PC : 2004 = 24.8, 2005 = 26.8; PB : 2004 = 406.5, 2005 = 409.1; YD : 2004 = 295.17, 2005 = 306.16. Many software packages, including Stata and EViews, have forecasting modules that will allow you to calculate forecasts using equations like Equation 15.4 automatically. If you use that module, you'll note that the forecasts differ slightly because we rounded the coefficient estimates.

¹¹² For a summary of seven different methods of measuring forecasting accuracy, see Peter Kennedy, *A Guide to Econometrics* (Malden, MA: Blackwell, 2008), pp. 334–335.

errors because the errors are squared before they're added together. For the chicken demand forecasts, the RMSE of our forecast is 5.97 kg (or 6 percent).

As you can see in Figure 15.1, it really doesn't matter which method you use, because the unconditional forecasts generated by equation (10.14) track quite well with reality. We missed by around 6 percent.

Forecasting Stock Prices: Some students react to the previous example by wanting to build a model to forecast stock prices and make a killing on the stock market. "If we could predict the price of a stock three years from now to within six percent," they reason, "we'd know which stocks to buy".

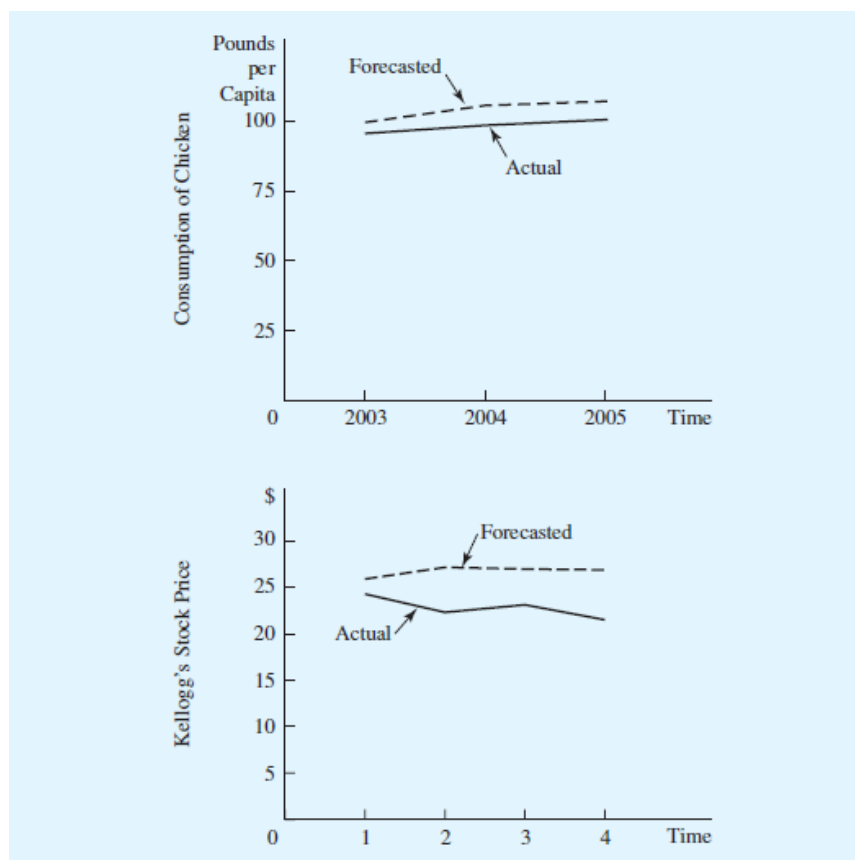


Figure 15.1 Forecasting Examples

In the chicken consumption example, the equation's forecast errors averaged around 6 percent. For the stock price model, even

actual values for the independent variables and an excellent fit within the sample could not produce an accurate forecast.

EXAMPLE 15.2

To see how such a forecast might work, let's look at a simplified model of the quarterly price of a particular individual stock, that of the Gazpromneft Company:

$$\begin{aligned}
 \hat{PG}_t = & -7.80 + 0.0096 DJA_t + 2.68 GPN_t + 16.18 DIV_t + 4.84 BVPS_t \\
 & \quad \quad \quad (0.0024) \quad \quad (2.83) \quad \quad (22.70) \quad \quad (1.47) \\
 & \quad \quad \quad t = 3.91 \quad \quad 0.95 \quad \quad 0.71 \quad \quad 3.29 \quad (15.5) \\
 & \quad \quad \quad \bar{R}^2 = 0.95 \quad \quad N = 35 \quad \quad DW = 1.88
 \end{aligned}$$

where: PG_t = the dollar price of Gazpromneft's stock in quarter t
 DJA_t = the Dow Jones industrial average in quarter t
 GPN_t = Gazpromneft's earnings growth (percent change in annual earnings over the previous five years) that quarter
 DIV_t = Gazpromneft's declared dividends (in dollars) that quarter
 $BVPS_t$ = per-share book value of the Gazpromneft company that quarter

The signs of the estimated coefficients all agree with those hypothesized before the regression was run, $\bar{R}^2$ indicates a good overall fit, and the Durbin - Watson test indicates that the hypothesis of no positive serial correlation cannot be rejected. The low t -scores for GPN and DIV are caused by multicollinearity ($r = 0.9852$), but both variables are left in the equation because of their theoretical importance. Note also that most of the variables in the equation are nonstationary, surely causing some of the good fit.

To forecast with equation (15.5), we collected actual values for all of the independent variables for the next four quarters and substituted them into the right side of the equation, obtaining:

Quarter	Forecast	Actual	Percent error
1	\$26.32	\$24.38	8.0
2	27.37	22.38	22.3
3	27.19	23.00	18.2
4	27.13	21.88	24.0

How did our forecasting model do? Even though the $\bar{R}^2$ within the sample was 0.95, even though we used actual values for the independent variables, and even though we forecasted only four quarters beyond our sample, the model was something like 20 percent off. If we had decided to buy Gazpromneft's stock based on our forecast, we'd have *lost* money!

Since other attempts to forecast stock prices have also encountered difficulties, this doesn't seem like a reasonable use for econometric forecasting. The poor performance of forecasting in the stock market can be explained by an economic theory called the efficient markets hypothesis, which suggests that accurate predictions of stock prices are practically impossible. The efficient markets hypothesis theorizes that "security prices fully reflect all available information."⁶ Thus, forecasting stock prices becomes a game of chance to the extent that markets are efficient and current prices reflect the available information.

Although an investor who has inside information (or who perhaps invents a superior forecasting approach) has an opportunity to create better than average stock price forecasts, the use of insider information is illegal in most equity markets.

15.2 More complex forecasting problems

The forecasts generated in the previous section are quite simple, however, and most actual forecasting involves one or more additional questions. For example:

1. *Unknown xs*: It's unrealistic to expect to know the values for the independent variables outside the sample. For instance, we'll almost never know what the Dow Jones industrial average will be in the future when we are making forecasts of the price of a given stock, and yet we assumed that knowledge when making our Gazpromneft price forecasts. What happens when we don't know the values of the independent variables for the forecast period?

2. *Serial Correlation*: If there is serial correlation involved, the forecasting equation may be estimated with GLS. How should predictions be adjusted when forecasting equations are estimated with GLS?
3. *Confidence Intervals*: All the previous forecasts were single values, but such single values are almost never exactly right. Wouldn't it be more helpful if we forecasted an interval within which we were confident that the actual value would fall a certain percentage of the time? How can we develop these confidence intervals?
4. *Simultaneous Equations Models*: As you saw in Chapter 14, many economic and business equations are part of simultaneous models. How can we use an independent variable to forecast a dependent variable when we know that a change in the value of the dependent variable will change, in turn, the value of the independent variable that we used to make the forecast?

Even a few questions like these should be enough to convince you that forecasting is more complex than is implied by Section 15.1.

Conditional forecasting (unknown X values for the forecast period)

A forecast in which all values of the independent variables are known with certainty can be called an unconditional forecast, but, as mentioned previously, the situations in which one can make such unconditional forecasts are rare. More likely, we will have to make a conditional forecast, for which actual values of one or more of the independent variables are *not* known. We are forced to obtain forecasts for the independent variables before we can use our equation to forecast the dependent variable, which makes our forecast of y conditional on our forecast of the x s.

One key to an accurate conditional forecast is accurate forecasting of the independent variables. If the forecasts of the

independent variables are unbiased, then using a conditional forecast will not introduce bias into the forecast of the dependent variable in a linear model. Anything but a perfect forecast of the independent variables will contain some amount of forecast error, however, and so the expected error variance associated with conditional forecasting will be larger than that associated with unconditional forecasting.

Thus, one should try to find unbiased, minimum variance forecasts of the independent variables when using conditional forecasting. To get good forecasts of the independent variables, take the forecastability of potential independent variables into consideration when making specification choices. For instance, when deciding which of two redundant variables should be included in an equation to be used for forecasting, you should choose the one that is easier to forecast accurately. When you can, you should choose an independent variable that is regularly forecasted by someone else (an econometric forecasting firm, for example) so that you don't have to forecast x yourself.

The careful selection of independent variables can sometimes help you avoid the need for conditional forecasting in the first place. This opportunity can arise when the dependent variable can be expressed as a function of leading indicators. A leading indicator is an independent variable whose movements anticipate movements in the dependent variable. The best known leading indicator, the Index of Leading Economic Indicators, is produced each month.

For instance, the impact of interest rates on investment typically is not felt until two or three quarters after interest rates have changed. To see this, let's look at the investment function of the small macroeconomic model of Section 14.3:

$$I_t = \beta_0 + \beta_1 Y_t + \beta_2 r_{t-1} + \epsilon_t \quad (15.6)$$

where I equals gross investment, Y equals GDP, and r equals the interest rate. In this equation, actual values of r can be used to help forecast I_{T+1} . Note, however, that to predict I_{T+2} , we need to forecast r_{T+1} . Thus, leading indicators like r help us avoid conditional forecasting for only a time period or two.

For long-range predictions, a conditional forecast is usually necessary. Forecasting with serially correlated error terms recall from Chapter 10 that pure first-order serial correlation implies that the current observation of the error term ϵ_t is affected by the previous error term and an autocorrelation coefficient, ρ :

$$\epsilon_t = \rho \epsilon_{t-1} + u_t$$

where u_t is a non-serially correlated error term. Also recall that when serial correlation is severe, one remedy is to run Generalized Least Squares (GLS) as noted in equation (10.21):

$$y_t - \rho y_{t-1} = \beta_0(1 - \rho) + \beta_1(x_t - \rho x_{t-1}) + u_t \quad (10.21)$$

Unfortunately, whenever the use of GLS is required to rid an equation of pure first-order serial correlation, the procedures used to forecast with that equation become a bit more complex. To see why this is necessary, note that if equation (10.21) is estimated, the dependent variable will be:

$$y_t^* = y_t - \hat{\rho} y_{t-1} \quad (15.7)$$

Thus, if a GLS equation is used for forecasting, it will produce predictions of y_{T+1}^* rather than of y_{T+1} . Such predictions thus will be of the wrong variable. If forecasts are to be made with a GLS equation, equation (10.21) should first be solved for y_t before forecasting is attempted:

$$y_t = \rho y_{t-1} + \beta_0(1 - \rho) + \beta_1(x_t - \rho x_{t-1}) + u_t \quad (15.8)$$

We now can forecast with Equation 15.8 as we would with any other equation. If we substitute $T + 1$ for t (to forecast time period $T + 1$) and insert estimates for the coefficients, ρ s and x s into the right side of the equation, we obtain:

$$\hat{y}_{T+1} = \hat{\rho} y_T + \beta_0(1 - \hat{\rho}) + \hat{\beta}_1(\hat{x}_{T+1} - \hat{\rho} x_T) \quad (15.9)$$

Equation (15.9) thus should be used for forecasting when an equation has been estimated with GLS to correct for serial correlation.¹¹³

We now turn to an example of such forecasting with serially correlated error terms. In particular, recall from Chapter 10 that the Durbin - Watson statistic of the chicken demand equation used as an example in Section 15.1 was 0.99, indicating significant positive first-order serial correlation. As a result, we estimated the chicken demand equation with GLS, obtaining equation (10.25):

$$\begin{aligned} \hat{Y}_t &= 28.5 - 0.08 PC_t + 0.016 PB_t + 0.24 YD_t \\ &\quad \begin{matrix} (0.04) & (0.021) & (0.02) \\ t = -2.13 & +0.74 & +13.2 \end{matrix} \\ \bar{R}^2 &= 0.963 \quad N = 29 \quad \hat{\rho} = 0.56 \end{aligned} \quad (10.25)$$

Since equation (10.25) was estimated with GLS, Y is actually Y_t^* , which equals $Y_t - \hat{\rho}Y_{t-1}$, PC_t is actually PC_t^* , which equals $PC_t - \hat{\rho}PC_{t-1}$, and so on. Thus, to forecast with equation (10.25), we have to convert it to the form of equation 15.9, or:

$$\begin{aligned} \hat{Y}_{T+1} &= 0.56Y_T + 28.5(1 - 0.56) - 0.08(PC_{T+1} - 0.56PC_T) + \\ &+ 0.016(PB_{T+1} - 0.56PB_T) + 0.24(YD_{T+1} - 0.56YD_T) \end{aligned} \quad (15.10)$$

Substituting the actual values for the independent variables into equation (15.10), we obtain:

Year	Forecast	Actual	Percent error
2003	97.57	95.63	2.0
2004	101.02	98.58	2.5
2005	102.38	100.60	1.8

The MAPE of the GLS forecasts is 2.1 percent, far better than that of the OLS forecasts. In general, GLS usually will provide

¹¹³ If $\hat{\rho}$ is less than 0.3, many researchers prefer to use the OLS forecast plus $\hat{\rho}$ times the lagged residual as their forecast instead of the GLS forecast from equation (15.9).

superior forecasting performance to OLS in the presence of serial correlation.

Forecasting confidence intervals

Until now, the emphasis in this text has been on obtaining point (or single-value) estimates. This has been true whether we have been estimating coefficient values or estimating forecasts. Recall, though, that a point estimate is only one of a whole range of such estimates that could have been obtained from different samples (for coefficient estimates) or different independent variable values or coefficients (for forecasts). The usefulness of such point estimates is improved if we can also generate some idea of the variability of our forecasts. The measure of variability typically used is the *confidence interval*, which was defined in Section 6.4 as the range of values that contains the true value of the item being estimated a specified percentage of the time. This is the easiest way to warn forecast users that a sampling distribution exists.

Suppose you are trying to decide how many hot dogs to order for your city's holiday show and the best point forecast is that you'll sell 24,000 hot dogs. How many hot dogs should you order? If you order 24,000, you're likely to run out about half the time! A point forecast is usually an estimate of the mean of the distribution of possible sales figures; you will sell more than 24,000 about as frequently as you will sell less than 24,000. It would be easier to decide how many dogs to order if you also had a confidence interval that told you the range within which hot dog sales would fall 95 percent of the time. The usefulness of the 24,000 hot dog forecast changes dramatically depending on the confidence interval; an interval of 22,000 to 26,000 would pin down the likely sales, but an interval of 4,000 to 44,000 would leave you virtually in the dark about what to do.

The decision as to how many hot dogs to order would also depend on the costs of ordering the wrong number. These costs may not be the same per hot dog for overestimates as they are for underestimates. For example, if you don't order enough, then you lose the entire retail price of the hot dog minus the wholesale price

of the dog (and bun) because your other costs, like hiring employees and building hot dog stands, are essentially fixed. On the other hand, if you order too many, you lose the wholesale cost of the dog and bun minus whatever salvage price you might be able to get for day-old buns, etc. As a result, the right number to order would depend on your profit margin and the importance of nonreturnable inputs in your total cost picture.

The same techniques we use to test hypotheses can also be adapted to create confidence intervals. Given a point forecast, $\hat{Y}_{T+1}$, all we need to generate a confidence interval around that forecast are t_c , the critical t -value (for the desired level of confidence), and S_F , the estimated standard error of the forecast:

$$\text{Confidence interval} = \hat{Y}_{T+1} \pm S_F t_c \quad (15.11)$$

or, equivalently,

$$\hat{Y}_{T+1} - S_F t_c \leq Y_{T+1} \leq \hat{Y}_{T+1} + S_F t_c \quad (15.12)$$

The critical t -value, t_c , can be found in Statistical Table B-1 in Appendix of Econometric analysis using Eviews: a practical tutorial (for a two-tailed test with $T - K - 1$ degrees of freedom). The standard error of the forecast, S_F , for an equation with just one independent variable, equals the square root of the forecast error variance:

$$S_F = \sqrt{s^2 \left[1 + 1/T + (\hat{X}_{T+1} - \bar{X})^2 / \sum_{t=1}^T (X_t - \bar{X})^2 \right]} \quad (15.13)$$

where s^2 = the estimated variance of the error term

T = the number of observations in the sample

$\hat{X}_{T+1}$ = the forecasted value of the single independent variable

$\bar{X}$ = the arithmetic mean of the observed X s in the sample

Note that equation (15.13) implies that the forecast error variance decreases the larger the sample is, the more X varies within the sample, and the closer $\hat{x}$ is to its within-sample mean. An

important implication is that the farther the X used to forecast Y is from the within-sample mean of the X s, the wider the confidence interval around the $\hat{Y}$ is going to be. This can be seen in Figure 15.2, in which the confidence interval actually gets wider as $\hat{X}_{T+1}$ is farther from $\bar{X}$. Since forecasting outside the sample range is common, researchers should be aware of this phenomenon. Also note that equation (15.13) is for unconditional forecasting. If there is any forecast error in $\hat{X}_{T+1}$, then the confidence interval is larger and more complicated to calculate. Finally, equation (15.13) should not be used in conjunction with HC standard errors.

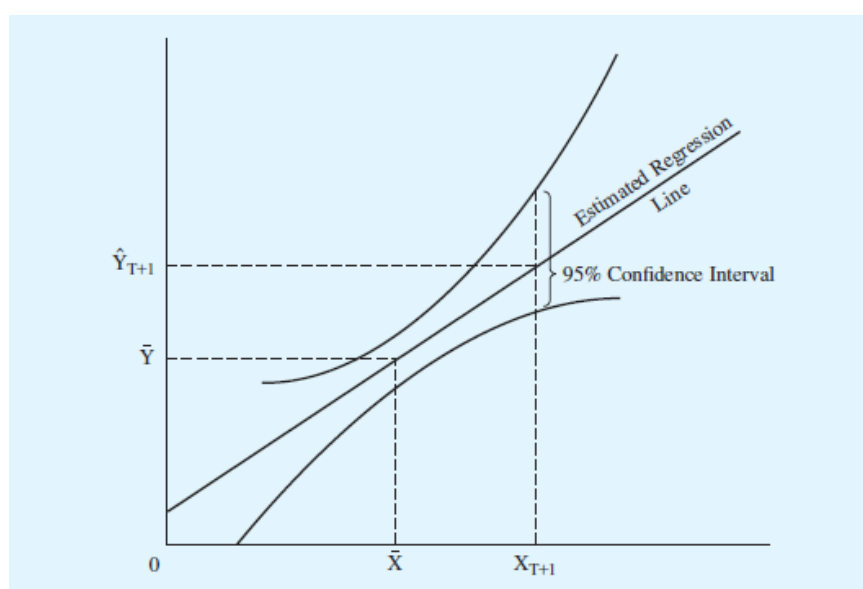


Figure 15.2 A Confidence Interval for $\hat{Y}_{T+1}$

A 95 percent confidence interval for $\hat{Y}_{T+1}$ includes the range of values within which the actual Y_{T+1} will fall 95 percent of the time. Note that the confidence interval widens as X_{T+1} differs more from its within-sample mean, $\bar{X}$.

As mentioned, equation (15.13) assumes that there is only one independent variable. The equation to be used with more than one variable is similar but more complicated. Equation (15.13) is valid whether Y_t is in the sample period or outside the sample period, but it applies only to point forecasts of individual Y_t s. If a confidence interval for the expected value of Y , $E(Y_t)$, is desired, then the correct equation to use is:

$$S_F^* = \sqrt{s^2 \left[1/T + (\hat{X}_{T+1} - \bar{X})^2 / \sum (X_t - \bar{X})^2 \right]} \quad (15.14)$$

Forecasting with simultaneous equations systems

As you learned in Chapter 14, most economic and business models are actually simultaneous in nature; for example, simultaneous macromodel estimated with 2SLS for Tajikistan.¹¹⁴ Since GDP is one of the independent variables in the consumption equation, when consumption rises, so will GDP, causing a feedback effect that is not captured if we just forecast with a single equation. How should forecasting be done in the context of a simultaneous model? There are two approaches to answering this question, depending on whether there are lagged endogenous variables on the right side of any of the equations in the system.

If there are no lagged endogenous variables in the system, then the reduced-form equation for the particular endogenous variable can be used for forecasting because it represents the simultaneous solution of the system for the endogenous variable being forecasted. Since the reduced-form equation is the endogenous variable expressed entirely in terms of the predetermined variables in the system, it allows the forecasting of the endogenous variable without any feedback or simultaneity impacts. This result explains why some researchers forecast potentially simultaneous dependent variables with single equations that appear to combine supply-side and demand-side predetermined variables; they are actually using modified reduced-form equations to make their forecasts.

If there are lagged endogenous variables in the system, then the approach must be altered to take into account the dynamic interaction caused by the lagged endogenous variables. For simple models, this sometimes can be done by substituting for the lagged endogenous variables where they appear in the reduced-form equations. If such a manipulation is difficult, however, then a technique called simulation analysis can be used. *Simulation*

¹¹⁴ Хакимова М.Ф. Динамическое моделирование макроэкономических процессов. Душанбе – Эр-граф, - 2013 – 200 стр.

involves forecasting for the first postsample period by using the reduced-form equations to forecast all endogenous variables where they appear in the reduced-form equations. The forecast for the second postsample period, however, uses the endogenous variable *forecasts* from the last period as lagged values for any endogenous variables that have one-period lags while continuing to use sample values for endogenous variables that have lags of two or more periods. This process continues until all forecasting is done with reduced-form equations that use as data for lagged endogenous variables the forecasts from previous time periods. Although such dynamic analyses are beyond the scope of this chapter, they're important to remember when considering forecasting with a simultaneous system.

15.3 ARIMA models

The forecasting techniques of the previous two sections are applications of familiar regression models. We use linear regression equations to forecast the dependent variable by plugging likely values of the independent variables into the estimated equations and calculating a predicted value of Y ; this bases the prediction of the dependent variable on the independent variables (and on their estimated coefficients).

ARIMA (the name will be explained shortly) is an increasingly popular forecasting technique that completely ignores independent variables in making forecasts. **ARIMA** is a highly refined curve-fitting device that uses current and past values of the dependent variable to produce often accurate short-term forecasts of that variable. While a traditional econometric model attempts to describe and estimate a variable's underlying structure (like a consumption function or a money demand function), an ARIMA model takes these structures as "black boxes" and simply analyzes the correlation pattern of a variable's movements over time in order to forecast it. Examples of such ARIMA forecasts are stock market price predictions created by brokerage analysts based entirely on past patterns of the movement of the stock price.

Any forecasting technique that ignores independent variables also essentially ignores all potential underlying theories except those that hypothesize repeating patterns in the variable under study. Since we have emphasized the advantages of developing the theoretical underpinnings of particular equations before estimating them, why would we advocate using ARIMA? The answer is that the use of ARIMA is appropriate when little or nothing is known about the dependent variable being forecasted, when the independent variables known to be important really cannot be forecasted effectively, or when all that is needed is a one- or two-period forecast. In these cases, ARIMA has the potential to provide short-term forecasts that are superior to more theoretically satisfying regression models. ARIMA models are particularly well suited to forecast a system that has not undergone a profound structural change within the sample or forecasting period. In such a situation, a naïve ARIMA model often can beat a moderately sophisticated econometric model in terms of forecasting outside the sample and has come close to the performance of state-of-the-art macro models in terms of forecasting key macro variables. In addition, ARIMA can sometimes produce better explanations of the residuals from an existing regression equation (in particular, one with known omitted variables or other problems). This introduction to ARIMA is intentionally brief; a more complete coverage of the topic can be obtained from a number of other sources.

The ARIMA approach combines two different specifications (called *processes*) into one equation. The first specification is an *autoregressive* process (hence the AR in ARIMA), and the second specification is a *moving average* (hence the MA).

An **autoregressive process** expresses a dependent variable Y_t as a function of past values of the dependent variable. This is similar to the serial correlation error term function of Chapter 10 and to the dynamic model of Chapter 12. If we have p different lagged values of Y , the equation is often referred to as a “ p th-order” autoregressive process.

A **moving-average process** expresses a dependent variable Y_t as a function of past values of the error term. Such a function is a

moving average of past error term observations that can be added to the mean of Y to obtain a moving average of past values of Y . If we used q past values of e , we'd call it a q th-order moving-average process.

To create an ARIMA model, we begin with an econometric equation with no independent variables $Y_t = \beta_0 + \epsilon_t$ and add to it both the autoregressive and moving-average processes:

$$Y_t = \beta_0 + \underbrace{\theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \dots + \theta_p Y_{t-p}}_{\text{autoregressive process}} + \underbrace{\phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \dots + \phi_q \epsilon_{t-q}}_{\text{moving-average process}} + \epsilon_t \quad (15.15)$$

where the θ s and the ϕ s are the coefficients of the autoregressive and moving-average processes, respectively, and p and q are the number of past values used of Y and e , respectively.

Before this equation can be applied to a time series, however, it must be ensured that the time series is *stationary*, as defined in Section 12.5. If a series is nonstationary, then steps must be taken to convert the series into a stationary one before the ARIMA technique can be applied. For example, a nonstationary series can often be converted into a stationary one by taking the first difference of the variable in question:

$$Y_t^* = \Delta Y_t = Y_t - Y_{t-1} \quad (15.16)$$

If the first differences do not produce a stationary series, then first differences of this first-differenced series can be taken.¹¹⁵ The resulting series is a second-difference transformation:

$$Y_t^{**} = \Delta Y_t^* = Y_t^* - Y_{t-1}^* = \Delta Y_t - \Delta Y_{t-1} \quad (15.17)$$

¹¹⁵ For variables that are growing in percentage terms rather than absolute amounts, it often makes sense to take logs before taking first differences.

In general, successive differences are taken until the series is stationary. The number of differences required to be taken before a series becomes stationary is denoted with the letter d . For example, suppose that GDP is increasing by a fairly consistent amount each year. A plot of GDP with respect to time would depict a nonstationary series, but a plot of the first differences of GDP might depict a fairly stationary series. In such a case, d would be equal to 1 because one first difference was necessary to convert the nonstationary series into a stationary one.

The dependent variable in equation (15.15) must be stationary, so the Y in that equation may be Y , Y^* , or even Y^{**} , depending on the variable in question.¹¹⁶

If a forecast of Y^* or Y^{**} is made, then it must be converted back into Y terms before its use; for example, if $d = 1$, then:

$$\hat{Y}_{T+1} = Y_T + \hat{Y}_{T+1}^* \quad (15.18)$$

This conversion process is similar to integration in mathematics, so the “I” in ARIMA stands for “integrated.” ARIMA thus stands for *AutoRegressive Integrated Moving Average*. (If the original series is stationary and d therefore equals 0, this is sometimes shortened to ARMA.)

As a shorthand, an ARIMA model with p , d and q specified is usually denoted as ARIMA (p,d,q) with the specific integers chosen inserted for p , d and q , as in ARIMA (2,1,1). ARIMA (2,1,1) would indicate a model with two autoregressive terms, one first difference, and one moving-average term:

$$ARIMA (2,1,1): Y_t^* = \beta_0 + \theta_1 Y_{t-1}^* + \theta_2 Y_{t-2}^* + \epsilon_t + \phi_1 \epsilon_{t-1} \quad (15.19)$$

where $Y_t^* = Y_t - Y_{t-1}$.

¹¹⁶ If Y in equation (15.15) is Y^* , then β_0 represents the coefficient of the linear trend in the original series, and if Y is Y^{**} , then β_0 represents the coefficient of the second-difference trend in the original series. In such cases - for example, equation (15.19) - it's not always necessary that β_0 be in the model.

It's remarkable how very small values of p and q can model extremely rich dynamics.

SUMMARY

1. Forecasting is the estimation of the expected value of a dependent variable for observations that are not part of the sample data set. Forecasts are generated (via regressions) by estimating an equation for the dependent variable to be forecasted, and substituting values for each of the independent variables (for the observations to be forecasted) into the equation.
2. An excellent fit within the sample period for a forecasting equation does not guarantee that the equation will forecast well outside the sample period.
3. A forecast in which all the values of the independent variables are known with certainty is called an unconditional forecast, but if one or more of the independent variables have to be forecasted, it is a conditional forecast. Conditional forecasting introduces no bias into the prediction of Y (as long as the X forecasts are unbiased), but increased forecast error variance is unavoidable with conditional forecasting.
4. If the coefficients of an equation have been estimated with GLS (to correct for pure first-order serial correlation), then the forecasting equation is:

$$\hat{Y}_{T+1} = \hat{\rho}Y_T + \hat{\beta}_0(1 - \hat{\rho}) + \hat{\beta}_1(\hat{X}_{T+1} - \hat{\rho}X_T)$$

where ρ is the autocorrelation coefficient rho.

5. Forecasts are often more useful if they are accompanied by a confidence interval, which is a range within which the actual value of the dependent variable should fall a given percentage of the time (the level of confidence). This is:

$$\hat{Y}_{T+1} \pm S_F t_c$$

where S_F is the estimated standard error of the forecast and t_c is the critical two-tailed t -value for the desired level of confidence.

6. ARIMA is a highly refined curve-fitting technique that uses current and past values of the dependent variable (and only the dependent variable) to produce often accurate short-term forecasts of that variable. The first step in using ARIMA is to make the dependent variable series stationary by taking d first differences until the resulting transformed variable has a constant mean and variance. The $ARIMA(p,d,q)$ approach then combines an autoregressive process (with $\theta_1 Y_{t-1}$ terms) of order p with a moving-average process (with $\phi_1 \epsilon_{t-1}$ terms) of order q to explain the d th differenced dependent variable.

REFERENCE

1. Айвазян С.А. Эконометрика -2. Продвинутый курс. – 2014. – 944 стр.
2. Доугерти К. Введение в эконометрику. – М: Инфра-М. – 2009. – 465 стр.
3. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. – М.: Дело. – 2004. – 576 стр.
4. Шанченко Н.И., Эконометрика: лабораторный практикум. – Ульяновск. – 2008. – 139 стр.
5. Frank, C. R., Jr.: *Statistics and Econometrics*, Holt, Rinehart and Winston, New York, 1971.
6. Goldberger, Arthur S.: *Introductory Econometrics*, Harvard University Press, 1998.
7. Gujarati, Damodar N.: *Essentials of Econometrics*, 2d ed., McGraw-Hill, New York, 1999.
8. Hill, Carter, William Griffiths, and George Judge: *Undergraduate Econometrics*, John Wiley & Sons, New York, 2001.
9. Hu, Teh-Wei: *Econometrics: An Introductory Analysis*, University Park Press, Baltimore, 1973.
10. Katz, David A.: *Econometric Theory and Applications*, Prentice Hall, Englewood Cliffs, N.J., 1982.
11. Klein, Lawrence R.: *An Introduction to Econometrics*, Prentice Hall, Englewood Cliffs, N.J., 1962.
12. Koop, Gary: *Analysis of Economic Data*, John Wiley & Sons, New York, 2000.
13. Walters, A. A.: *An Introduction to Econometrics*, Macmillan, London, 1968.

Khakimova M.F.

TEXTBOOK
INTRODUCTORY ECONOMETRICS

Reviewers:

Jurabaev G.J., Doctor of Sciences, Full professor of the public and local management of the Institute of tourism, entrepreneurship and services of the Ministry of industry and new technologies of the Republic of Tajikistan;

Tabarov O.S., Doctor of Science, Associate professor, head of economic cybernetics department of the Tajik national university;

Najmiddinov A.A., Doctor of Science, Associate professor, head of mathematical and informational modelling in economics department of the Tajik state university of finance and economics

Commissioned print 20.04.2020.

Signed for printing 01.05.2020. Offset paper. Offset printing.

Format 60x84 1/16. Size 13. 100 units.

Printed in typography LLC “Bebok”