ECONOMETRIC METHODS

FOURTH EDITION

JACK JOHNSTON
JOHN DiNARDO

3.5"
data disk enclosed.
Formatted in ASCII
and adaptable to
most statistical analysis
programs.

http://www.mhcollege.com

P/N 032720-3
PART OF
ISBN 0-07-913121-2

90000

9 780079 131218

McGraw-Hill
A Division of The McGraw-Hill Companies

Appendix A

Appendix B

CHAPTER 1

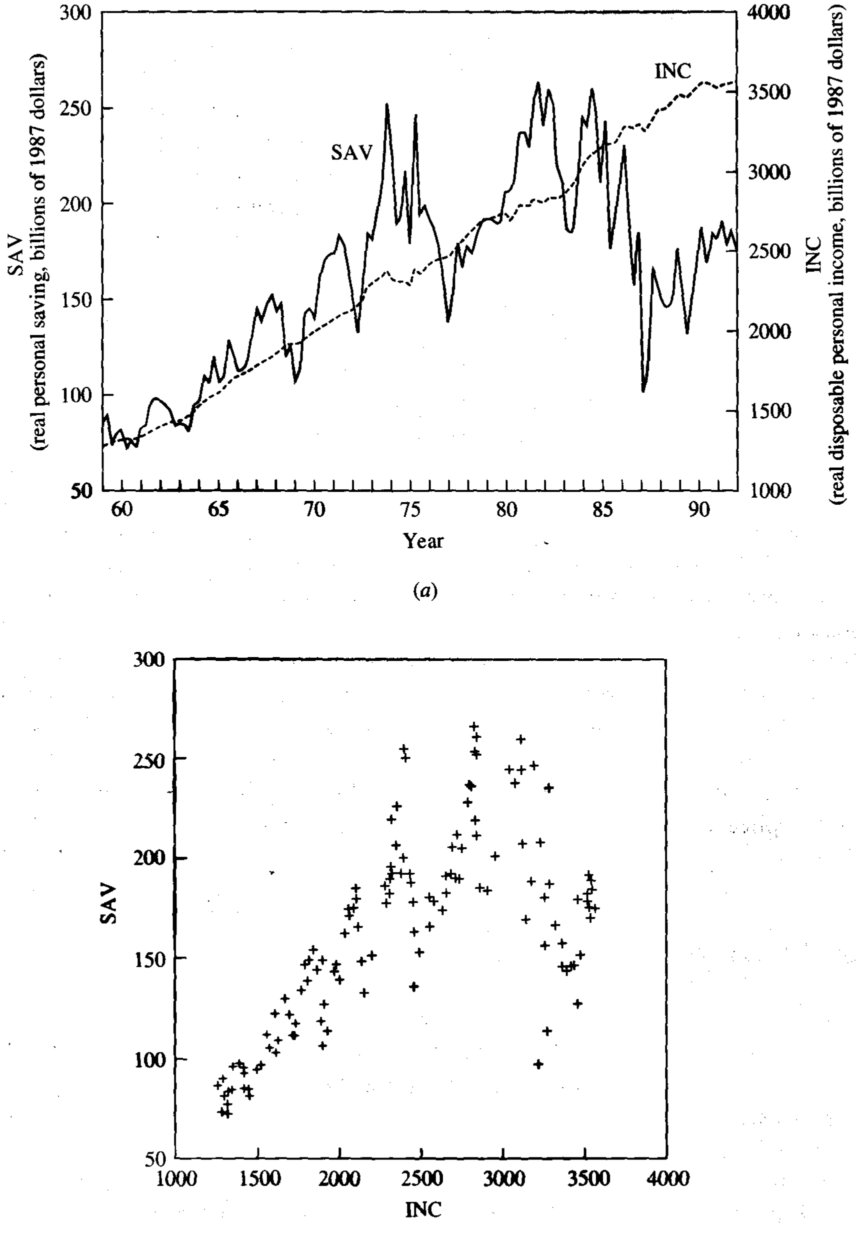
Relationships between Two Variables

The economics literature contains innumerable discussions of relationships between variables in pairs: quantity and price; consumption and income; demand for money and the interest rate; trade balance and the exchange rate; education and income; unemployment and the inflation rate; and many more. This is not to say that economists believe that the world can be analyzed adequately in terms of a collection of bivariate relations. When they leave the two-dimensional diagrams of the textbooks behind and take on the analysis of real problems, *multivariate* relationships abound. Nonetheless, some bivariate relationships are significant in themselves; more importantly for our purposes, the mathematical and statistical tools developed for two-variable relationships are fundamental building blocks for the analysis of more complicated situations.

1.1 EXAMPLES OF BIVARIATE RELATIONSHIPS

Figure 1.1 displays two aspects of the relationship between real personal saving (SAV) and real personal disposable income (INC) in the United States. In Fig. 1.1*a* the value of each series is shown quarterly for the period from 1959.1 to 1992.1. These two series and many of the others in the examples throughout the book come from the DRI Basic Economics Database (formerly Citibase); where relevant, we indicate the correspondence between our labels and the Citibase labels for the variables.[1] Figure 1.1*a* is a typical example of a **time series plot,** in which time is displayed on the horizontal axis and the values of the series are displayed on the vertical axis. Income shows an upward trend throughout the period, and in the early years, saving does likewise. This pattern, however, is not replicated in the middle and later

[1] A definition of all series is given in the data disk, which accompanies this volume. Instructions for accessing the disk are given in Appendix C.



(a)

(b)

FIGURE 1.1
Saving and income.

years. One might be tempted to conclude from Fig. 1.1*a* that saving is much more volatile than income, but that does not necessarily follow, since the series have separate scales.[2]

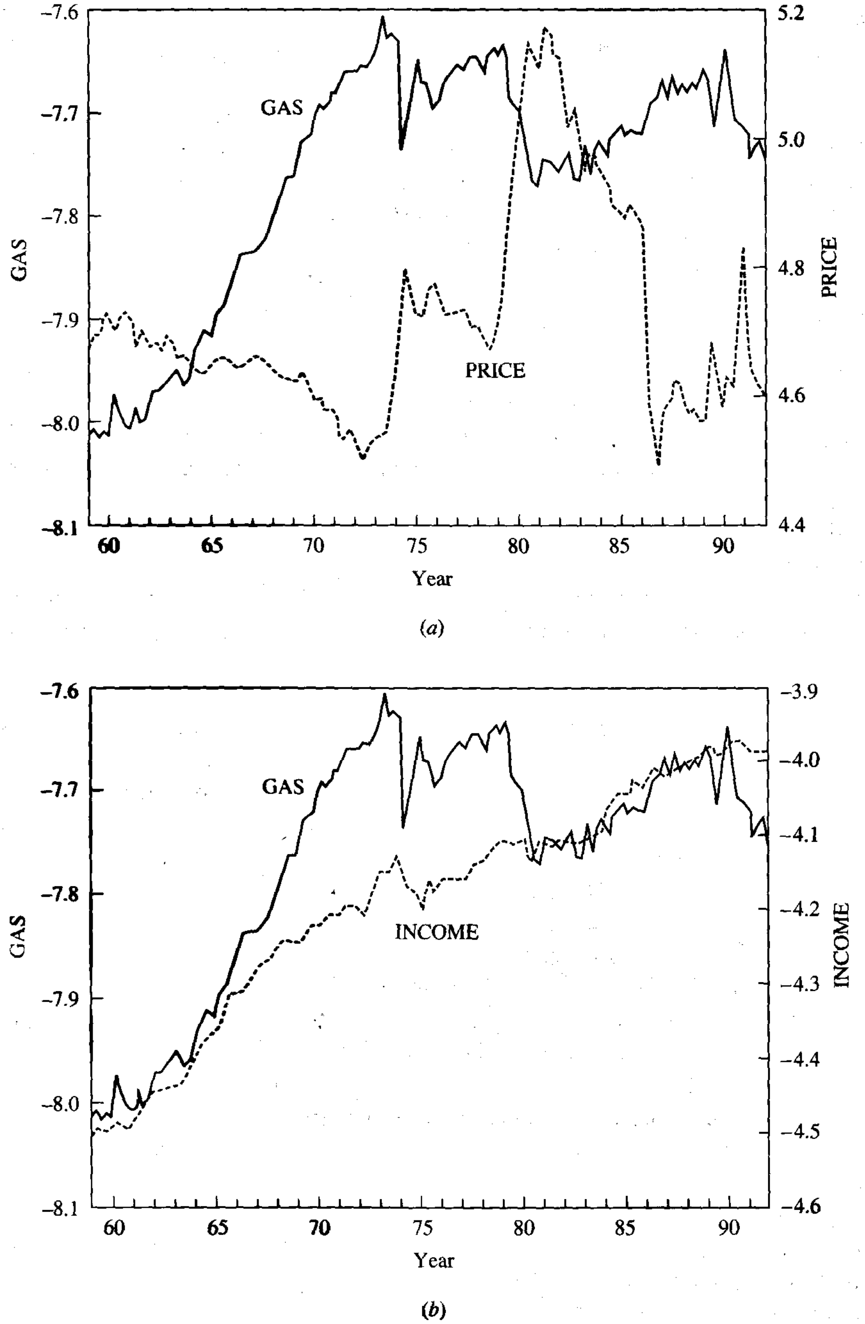
An alternative display of the same information is in terms of a **scatter plot,** shown in Fig. 1.1*b*. Here one series is plotted against the other. The time dimension is no longer shown explicitly, but most software programs allow the option of joining successive points on the scatter so that the evolution of the series over time may still be traced. Both parts of Fig. 1.1 indicate a *positive* association between the variables: increases in one tend to be associated with increases in the other. It is clear that although the association is approximately linear in the early part of the period, it is not so in the second half.

Figures 1.2 and 1.3 illustrate various associations between the natural log of real personal expenditure on gasoline (GAS), the natural log of the real price of gasoline (PRICE), and the natural log of real disposable personal income (INCOME). The derivations of the series are described in the data disk. The rationale for the logarithmic transformations is discussed in Chapter 2. Figure 1.2 gives various time plots of gasoline expenditure, price, and income. The real price series, with 1987 as the base year, shows the two dramatic price hikes of the early and late 1970s, which were subsequently eroded by reductions in the nominal price of oil and by U.S. inflation, so the real price at the end of the period was less than that obtaining at the start. The income and expenditure series are both shown in per capita form, because U.S. population increased by about 44 percent over the period, from 176 million to 254 million. The population series used to deflate the expenditure and income series is the civilian noninstitutional population aged 16 and over, which has increased even faster than the general population. Per capita real expenditure on gasoline increased steadily in the 1960s and early 1970s, as real income grew and real price declined. This steady rise ended,with the price shocks of the 1970s, and per capita gas consumption has never regained the peak levels of the early seventies.

The scatter plots in Fig. 1.3 further illustrate the upheaval in this market. The plot for the whole period in Fig. 1.3*a* shows very different associations between expenditure and price in the earlier and later periods. The scatter for 1959.1 to 1973.3 in Fig. 1.3*b* looks like a conventional negative association between price and quantity. This is shattered in the middle period (1973.4 to 1981.4) and reestablished, though with a very different slope, in the last period (1982.1 to 1992.1). This data set will be analyzed econometrically in this and later chapters.

These illustrative scatter diagrams have three main characteristics. One is the *sign* of the association or covariation—that is, do the variables move together in a *positive* or *negative* fashion? Another is the *strength* of the association. A third characteristic is the *linearity* (or otherwise) of the association—is the general shape of the scatter linear or curvilinear? In Section 1.2 we discuss the extent to which the correlation coefficient measures the first two characteristics for a linear association, and in later chapters we will show how to deal with the linearity question, but first we give an example of a bivariate frequency distribution.

[2] See Problem 1.1.



(*a*)

(*b*)

FIGURE 1.2

Time series plots of natural log of gasoline consumption in 1987 dollars per capita. (*a*) Gasoline consumption vs. natural log of price in 1987 cents per gallon. (*b*) Gasoline consumption vs. natural log of income in 1987 dollars per capita.

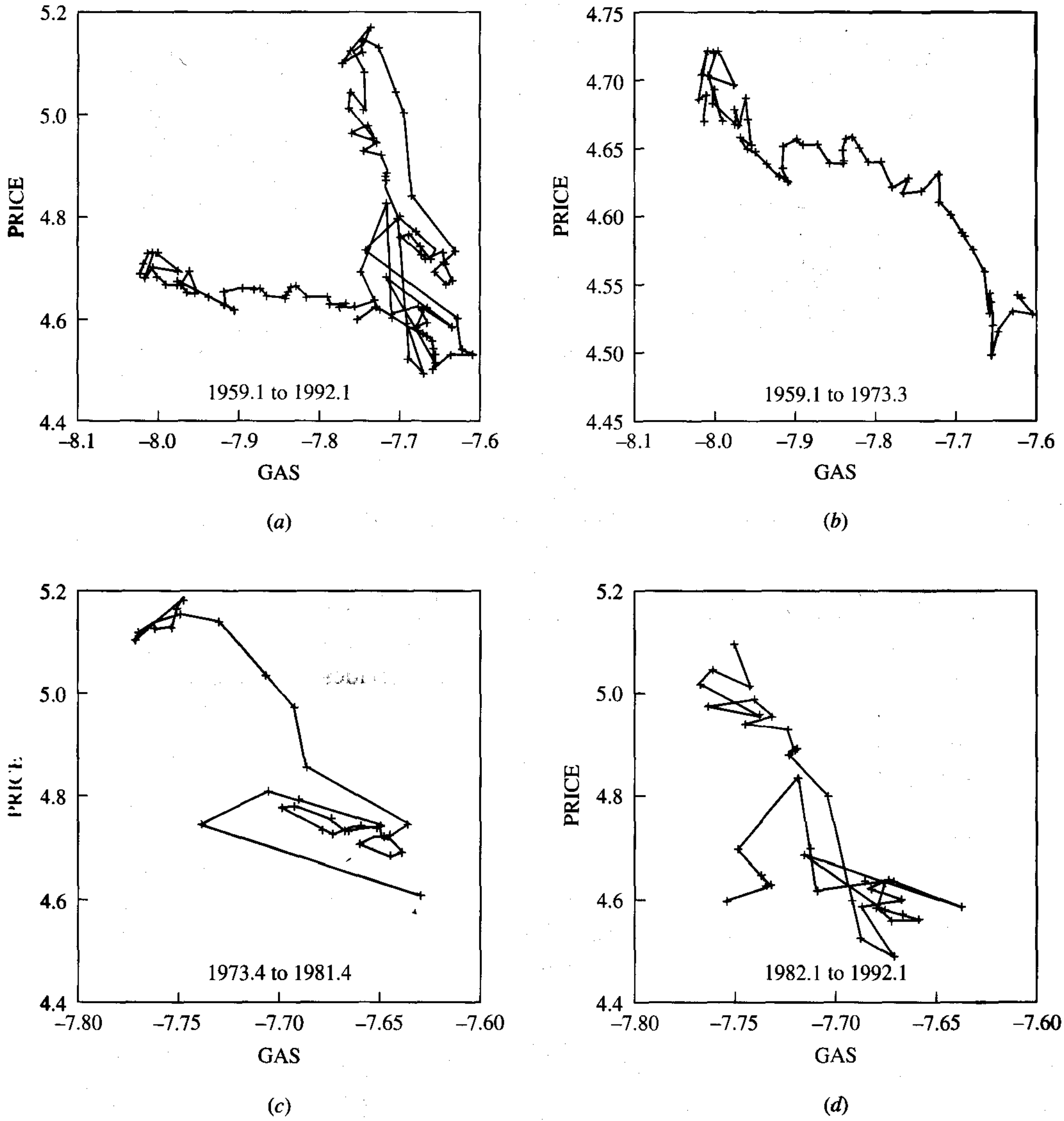


FIGURE 1.3
Scatter plots of price and gasoline consumption.

1.1.1 Bivariate Frequency Distributions

The data underlying Figs. 1.1 to 1.3 come in the form of n pairs of observations of the form (X_i, Y_i), $i = 1, 2, \ldots, n$. When the sample size n is very large, the data are usually printed as a bivariate frequency distribution; the ranges of X and Y are split into subintervals and each cell of the table shows the number of observations

TABLE 1.1
Distribution of heights and chest circumferences of 5732 Scottish militiamen

		Chest circumference (inches)					
		33–35	36–38	39–41	42–44	45 and over	Row totals
	64–65	39	331	326	26	0	722
Height (inches)	66–67	40	591	1010	170	4	1815
	68–69	19	312	1144	488	18	1981
	70–71	5	100	479	290	23	897
	72–73	0	17	120	153	27	317
Column totals		103	1351	3079	1127	72	5732

Source: Edinburgh Medical and Surgical Journal (1817, pp. 260–264).

TABLE 1.2
Conditional means for the data in Table 1.1

Mean of height given chest (inches)	66.31	66.84	67.89	69.16	70.53
Mean of chest given height (inches)	38.41	39.19	40.26	40.76	41.80

in the corresponding pair of subintervals. Table 1.1 provides an example.[3] It is not possible to give a simple, two-dimensional representation of these data. However, inspection of the cell frequencies suggests a positive association between the two measurements. This is confirmed by calculating the *conditional* means. First of all, each of the five central columns of the table gives a distribution of heights for a given chest measurement. These are *conditional* frequency distributions, and traditional statistics such as means and variances may be calculated. Similarly, the rows of the table give distributions of chest measurements, conditional on height. The two sets of conditional means are shown in Table 1.2; each mean series increases monotonically with increases in the conditioning variable, indicating a positive association between the variables.

1.2 THE CORRELATION COEFFICIENT

The direction and closeness of the linear association between two variables are measured by the correlation coefficient.[4] Let the observations be denoted by (X_i, Y_i) with $i = 1, 2, \ldots, n$. Once the sample means have been calculated, the data may be expressed in deviation form as

$$x_i = X_i - \bar{X} \qquad y_i = Y_i - \bar{Y}$$

[3] Condensed from Stephen M. Stigler, *The History of Statistics*, Harvard University Press, 1986, p. 208.

[4] See Stigler, *op. cit.*, for a fascinating and definitive history of the evolution of the correlation coefficient.

where $\bar{X}$ and $\bar{Y}$ denote the sample means of X and Y. Figure 1.4 shows an illustrative point on a scatter diagram with the sample means as new axes, giving four quadrants, which are numbered counterclockwise. The product $x_i y_i$ is positive for all points in quadrants I and III and negative for all points in quadrants II and IV. Since a positive relationship will have points lying for the most part in quadrants I and III, and a negative relationship will have points lying mostly in the other two quadrants, the sign of $\sum_{i=1}^{n} x_i y_i$ will indicate whether the scatter slopes upward or downward. This sum, however, will tend to increase in absolute terms as more data are added to the sample. Thus, it is better to express the sum in *average* terms, giving the sample *covariance,*

$$\begin{aligned} \operatorname{cov}(X, Y) &= \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})/n \\ &= \sum_{i=1}^{n} x_i y_i / n \end{aligned} \tag{1.1}$$

The value of the covariance depends on the units in which the variables are measured. Changing one variable from dollars to cents will give a new covariance 100 times the old. To obtain a measure of association that is invariant with respect to units of measurement, the deviations are expressed in *standard deviation* units. The covariance of the standardized deviations is the *correlation coefficient, r* namely,

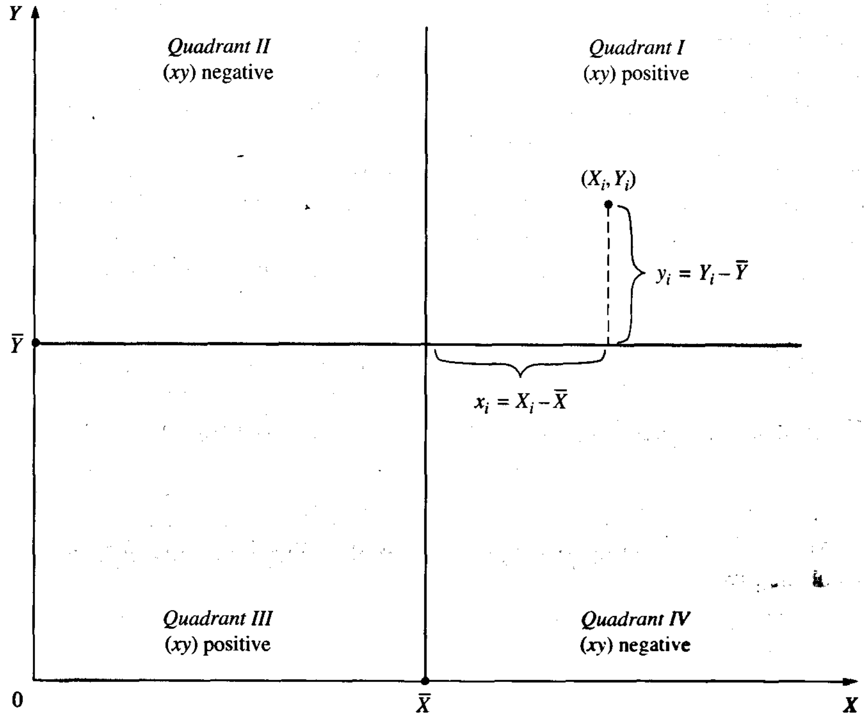


FIGURE 1.4
Coordinates for scatter diagram for paired variables.

$$r = \sum_{i=1}^{n}\left(\frac{x_i}{s_x}\right)\left(\frac{y_i}{s_y}\right)\Big/n = \sum_{i=1}^{n} x_i y_i / n s_x s_y \tag{1.2}$$

where $s_x = \sqrt{\sum_{i=1}^{n} x_i^2/n}$

$$s_y = \sqrt{\sum_{i=1}^{n} y_i^2/n}$$

Omitting subscripts and the limits of summation (since there is no ambiguity) and performing some algebraic manipulations give three equivalent expressions for the correlation coefficient—two in terms of deviations and one in terms of the raw data:

$$\begin{aligned} r &= \frac{\sum xy}{ns_x s_y} \\ &= \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} \\ &= \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}} \end{aligned} \tag{1.3}$$

1.2.1 The Correlation Coefficient for a Bivariate Frequency Distribution

In general, a bivariate distribution such as that shown in Table 1.1 may be represented by the paired values X_i, Y_j with frequency n_{ij} for $i = 1, \ldots, m$ and $j = 1, \ldots, p$. X_i is the midpoint of the ith subinterval on the X axis, and Y_j the midpoint of the jth subinterval on the Y axis. If we use a period for a subscript over which summation has taken place, the marginal frequencies for X are given by $n_{i.} = \sum_{j=1}^{p} n_{ij}$ for $i = 1, \ldots, m$. In conjunction with the X_i values these marginal frequencies will yield the standard deviation of X, that is, s_x. The marginal frequencies for Y are $n_{.j} = \sum_{i=1}^{m} n_{ij}$ for $j = 1, \ldots, p$. Thus, the standard deviation of Y, or s_y, may be obtained. Finally the covariance is obtained from

$$\operatorname{cov}(X, Y) = \sum_{i=1}^{m}\sum_{j=1}^{p} n_{ij}(X_i - \bar{X})(Y_j - \bar{Y})/n \tag{1.4}$$

where n is the total number of observations. Putting the three elements together, one may express the correlation coefficient for the bivariate frequency distribution in terms of the raw data as

$$r = \frac{n\sum_{i=1}^{m}\sum_{j=1}^{p} n_{ij}X_iY_j - (\sum_{i=1}^{m} n_{i.}X_i)(\sum_{j=1}^{p} n_{.j}Y_j)}{\sqrt{n\sum_{i=1}^{m} n_{i.}X_i^2 - (\sum_{i=1}^{m} n_{i.}X_i)^2}\sqrt{n\sum_{j=1}^{p} n_{.j}Y_j^2 - (\sum_{j=1}^{p} n_{.j}Y_j)^2}} \tag{1.5}$$

1.2.2 The Limits of *r*

The correlation coefficient must lie in the range from -1 to $+1$. To see this, let c be any arbitrary constant. Then $\sum(y - cx)^2 \geq 0$. Now let $c = \sum xy / \sum x^2$. Substitution in the inequality gives $(\sum xy)^2 \leq (\sum x^2)(\sum y^2)$, that is, $r^2 \leq 1$. This expression is one form of the **Cauchy-Schwarz inequality.** The equality will only hold if each and every y deviation is a constant multiple of the corresponding x deviation. In such a case the observations all lie on a single straight line, with a positive slope ($r = 1$) or a negative slope ($r = -1$). Figure 1.5 shows two cases in which r is approximately zero. In one case the observations are scattered over all four quadrants; in the other they lie exactly on a quadratic curve, where positive and negative products offset one another. Thus, the correlation coefficient measures the degree of *linear* association. A low value for r does not rule out the possibility of a strong *nonlinear* association, and such an association might give positive or negative values for r if the sample observations happen to be located in particular segments of the nonlinear relation.

1.2.3 Nonsense Correlations and Other Matters

Correlation coefficients must be interpreted with care. Many coefficients that are both numerically large and also adjudged *statistically significant* by tests to be described later may contain no real information. That statistical significance has been achieved does not necessarily imply that a meaningful and useful relationship has been found. The crucial question is, What has caused the observed covariation? If there is a theory about the joint variation of X and Y, the sign and size of the correlation coefficient may lend support to that theory, but if no such theory exists or can be devised, the correlation may be classed as a nonsense correlation.

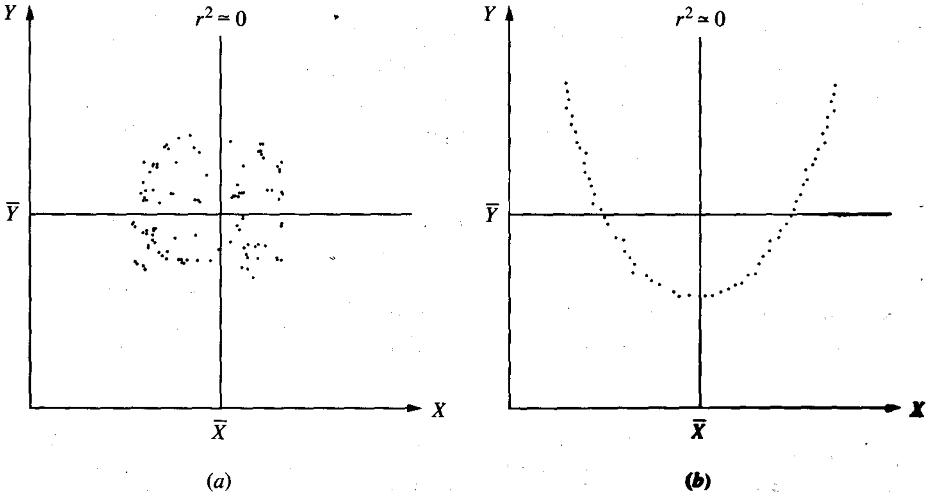


FIGURE 1.5
Paired variables for which $r^2 \simeq 0$.

Our favorite spurious, or nonsense, correlation was given in a beautiful 1926 paper by the statistician G. Udny Yule.[5] Yule took annual data from 1866 to 1911 for the death rate in England and Wales and for the proportion of all marriages solemnized in the Church of England and found the correlation coefficient to be +0.95. However, no British politician proposed closing down the Church of England to confer immortality on the electorate. More recently, using annual data from 1897 to 1958, Plosser and Schwert have found a correlation coefficient of +0.91 between the log of nominal income in the United States and the log of accumulated sunspots.[6] Hendry has noted a very strong, though somewhat nonlinear, positive relationship between the inflation rate and the accumulation of annual rainfall in the United Kingdom.[7] It would be nice if the British could reduce their inflation rate and, as a bonus, enjoy the inestimable side effect of improved weather, but such happy conjunctions are not to be.

In these three examples all of the variables are subject to trend-like movements over time.[8] Presumably some complex set of medical, economic, and social factors contributed to the reduction in the death rate in England and Wales, even as a different set of factors produced a decline in the proportion of marriages in the Church of England. Cumulative sunspots and cumulative rainfall necessarily trend upward, as do the U.S. nominal income and the British inflation rate. Series responding to essentially unrelated generating mechanisms may thus display contemporaneous upward and/or downward movements and thus yield strong correlation coefficients. Trends may be fitted to such series, as will be shown in the next chapter, and the *residuals* from such trends calculated. Correlations between pairs of residuals for such series will be negligible.

An alternative approach to correlating detrended residuals is to correlate the **first differences** of the series. The first differences are simply the changes in the series between adjacent observations. They are usually denoted by the prefix Δ. Thus,

$$\Delta X_t = X_t - X_{t-1} \qquad \Delta Y_t = Y_t - Y_{t-1}$$

Many series that show very high correlations between X and Y (the *levels*) will show very low correlations between ΔX and ΔY (the *first differences*). This result usually indicates a spurious relationship. On the other hand, if there is a causal relationship between the variables, we expect to find correlations between levels and also between first differences. This point has recently been emphasized in an important paper by Stigler and Sherwin.[9] The main thesis of the paper is that if

[5] G. Udny Yule, "Why Do We Sometimes Get Nonsense Correlations between Time Series?", *Journal of the Royal Statistical Society*, Series A, General, **89**, 1926, 1–69.

[6] Charles I. Plosser and G. William Schwert, "Money, Income, and Sunspots: Measuring Economic Relationships and the Effects of Differencing," *Journal of Monetary Economics*, **4**, 1978, 637–660.

[7] David F. Hendry, "Econometrics—Alchemy or Science?", *Economica*, **47**, 1980, 387–406.

[8] Trends, like most economic phenomena, are often fragile and transitory. The point has been made in lyrical style by Sir Alec Cairncross, one of Britain's most distinguished economists and a former chief economic adviser to the British government. "A trend is a trend, is a trend, but the question is, will it bend? Will it alter its course, through some unforeseen force and come to a premature end?"

[9] George J. Stigler and Robert A. Sherwin, "The Extent of the Market," *Journal of Law and Economics*, **28**, 1985, 555–585.

two goods or services are in the same market their prices should be closely related. However, since most prices, like many economic series, show trend-like movements over time, Stigler and Sherwin wish to guard against being misled by spurious correlation. Thus, in addition to correlating price levels they correlate price changes. As one example, the prices of December 1982 silver futures on the New York Commodity Exchange and the Chicago Board of Trade over a 30-day trading period gave $r = 0.997$, and the price changes gave $r = 0.956$. In Minneapolis, Minnesota, and Kansas City, Missouri, two centers of the flour-milling industry, the monthly wholesale prices of flour over 1971–1981 gave correlations of 0.97 for levels and 0.92 for first differences. In these two cases the first difference correlations strongly reinforce the levels correlations and support the thesis of a single market for these goods.

1.2.4 A Case Study

Gasoline is retailed on the West Coast of the United States by the "majors" (Arco, Shell, Texaco, etc.) and by "minors," or "independents." Traditionally the majors have offered a greater variety of products, differentiated in terms of grade of gasoline, method of payment, degree of service, and so forth; whereas the minors have sold for cash and offered a smaller range of products. In the spring of 1983 Arco abolished its credit cards and sold for cash only. By the fall of 1983 the other majors had responded by continuing their credit cards but introducing two prices, a credit price and a lower cash price. Subsequently one of the independents sued Arco under the antitrust laws. The essence of the plaintiff's case was that there were really two separate markets for gasoline, one in which the majors competed with each other, and a second in which the minors competed. They further alleged, though not in this precise language, that Arco was like a shark that had jumped out of the big pool into their little pool with the intention of gobbling them all up. No one questioned that there was competition *within* the majors and competition *within* the minors: the crucial question was whether there was competition between majors and minors.

The problem was a perfect candidate for the Stigler/Sherwin type of analysis. The Lundberg Survey reports detailed information twice a month on the prices of all types and grades of gasoline at a very large sample of stations. These data are also averaged for majors and minors. Twelve differentiated products were defined for the majors and four for the minors. This step allowed the calculation of 66 correlation coefficients for all pairs of products within the majors and 6 correlation coefficients within the minors. Each set of coefficients would be expected to consist of very high numbers, reflecting the intensity of competition inside each group. However, it was also possible to calculate 48 correlation coefficients for all cross-pairs of a major price and a minor price. If the plaintiff's argument were correct, these 48 coefficients would be of negligible size. On the other hand, if there were just a single large market for gasoline, the cross correlations should not be markedly less than correlations within each group. A nice feature of the problem was that the within-group correlations provided a standard of reference for the assessment of the cross correlations. In the cases discussed in the Stigler/Sherwin paper only subjective judgments could be made about the size of correlation coefficient required to establish that two goods were in the same market.

The preceding approach yielded a matrix of 120 correlation coefficients. In order to guard against possible spurious correlation, such a matrix was computed for levels, for first differences, for logs of levels, and for first differences of logs (which measure percent changes in price). In addition, regression analysis was used to adjust for possible common influences from the price of crude oil or from general inflation, and matrices were produced for correlations between the residuals from these regressions. In all cases the matrices showed "forests" of tall trees (that is, high correlation coefficients), and the trees were just as tall in the rectangle of cross correlations as in the triangles of within correlations. The simple correlation coefficients thus provided conclusive evidence for the existence of a single market for retail gasoline.

1.3 PROBABILITY MODELS FOR TWO VARIABLES

Classical statistical inference is based on the presumption that there exists some *population distribution* of all possible observations on the variables of interest. That distribution is characterized by certain crucial parameter values. From a sample of n observations sample statistics are computed and these serve as a basis for inference about the population parameters. Ever since the work of Haavelmo in the 1940s the probability approach has been extensively used in econometrics.[10] Indeed the development of econometrics in the past half century has been driven mainly by the effort to adapt and extend classical inference procedures to deal with the special problems raised by the nature of the data generation process in economics and the general unavailability of controlled economic experiments.

1.3.1 Discrete Bivariate Probability Distribution

To introduce some of the main ideas, consider a *discrete* bivariate probability distribution as shown in Table 1.3. The cell entries indicate the probability of the joint occurrence of the associated X, Y values. Thus, p_{ij} = probability that $X = X_i$ and $Y = Y_j$. The column and row totals, where a period indicates the subscript over which summation has taken place, give the marginal probabilities for X and Y, respectively. There are six important population parameters for the bivariate distribution. The means are defined by

$$\mu_x = E(X) = \sum_i p_{i.} X_i \qquad \text{and} \qquad \mu_y = E(Y) = \sum_j p_{.j} Y_j \tag{1.6}$$

The variances are defined as

$$\begin{aligned} \sigma_x^2 &= \text{var}(X) = E[(X - \mu_x)^2] = \sum_i p_{i.}(X_i - \mu_x)^2 \\ \sigma_y^2 &= \text{var}(Y) = E[(Y - \mu_y)^2] = \sum_j p_{.j}(Y_j - \mu_y)^2 \end{aligned} \tag{1.7}$$

[10] Trygve Haavelmo, *The Probability Approach in Econometrics*, supplement to *Econometrica*, **12**, July, 1944.

TABLE 1.3
A bivariate probability distribution

	X_1	$\cdots$	X_i	$\cdots$	X_m	**Marginal probability**
Y_1	p_{11}	$\cdots$	p_{i1}	$\cdots$	p_{m1}	$p_{\cdot 1}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
Y_j	p_{1j}	$\cdots$	p_{ij}	$\cdots$	p_{mj}	$p_{\cdot j}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
Y_p	p_{1p}	$\cdots$	p_{ip}	$\cdots$	p_{mp}	$p_{\cdot p}$
Marginal probability	$p_{1\cdot}$	$\cdots$	$p_{i\cdot}$	$\cdots$	$p_{m\cdot}$	1

$\sum p_{1\bullet}$

The covariance is

$$\begin{aligned}\sigma_{xy} &= \operatorname{cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)] \\ &= \sum_i \sum_j p_{ij}(X_i - \mu_x)(Y_j - \mu_y)\end{aligned} \tag{1.8}$$

Finally, the population correlation coefficient is defined as

$$\operatorname{corr}(X, Y) = \rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{1.9}$$

In these formulae $\sum_i$ and $\sum_j$ indicate summation over the relevant subscripts.

Conditional probabilities

Consider the X_i column in Table 1.3. Each cell probability may be divided by the column total, $p_{i.}$, to give a *conditional* probability for Y given X_i. Thus,

$$\begin{aligned}\frac{p_{ij}}{p_{i.}} &= \text{probability that } Y = Y_j \text{ given that } X = X_i \\ &= \operatorname{prob}(Y_j \mid X_i)\end{aligned} \tag{1.10}$$

The mean of this distribution is the *conditional expectation* of Y, given X_i, that is,

$$\mu_{y|x_i} = E(Y \mid X_i) = \sum_j \left(\frac{p_{ij}}{p_{i.}}\right) Y_j \tag{1.11}$$

Similarly, the variance of this distribution is a *conditional* variance, or

$$\sigma^2_{y|x_i} = \operatorname{var}(Y \mid X_i) = \sum_j \left(\frac{p_{ij}}{p_{i.}}\right)(Y_j - \mu_{y|x_i})^2 \tag{1.12}$$

The conditional means and variances are both functions of X, so there is a set of m conditional means and variances. In a similar fashion one may use the row probabilities to study the conditional distributions of X given Y.

TABLE 1.4
Bivariate distribution of income (X) and vacation expenditure (Y)

		X ($'000)		
		20	30	40
	1	.28	.03	0
	2	.08	.15	.03
Y	3	.04	.06	.06
($'000)	4	0	.06	.15
	5	0	0	.03
	6	0	0	.03
Marginal probability		.40	.30	.30
Mean (Y | X)		1.4	2.5	3.9
Var (Y X)		.44	.85	1.09

0.7+ 0.4 + 0.3+0+0+0 = 1.4.

TABLE 1.5
Conditional probabilities from Table 1.4

		Y					
		1	2	3	4	5	6
	20	0.7	0.2	0.1	0	0	0
X	30	0.1	0.5	0.2	0.2	0	0
	40	0	0.1	0.2	0.5	0.1	0.1

$\frac{.28}{0.4}$ $\frac{0.08}{0.4}$ $\frac{0.04}{0.4}$

A numerical example

Table 1.4 presents hypothetical data on income and vacation expenditure for an imaginary population. There are just three levels of income and six possible levels of vacation expenditure. Everyone, no matter how humble, gets to spend at least $1,000 on vacation. The marginal probabilities show that 40 percent of this population have incomes of $20,000, 30 percent have incomes of $30,000, and 30 percent have incomes of $40,000. The conditional probabilities derived from these data are shown in Table 1.5. These conditional probabilities are used to calculate the conditional means and variances shown in the last two rows of Table 1.4. Mean vacation expenditure rises with income but the increase is not linear, being greater for the increase from $30,000 to $40,000 than for the increase from $20,000 to $30,000. The conditional variance also increases with income. One could carry out the parallel analysis for X given Y. This might be of interest to a travel agent concerned with the distribution of income for people with a given vacation expenditure.

1.3.2 The Bivariate Normal Distribution

The previous examples have been in terms of discrete variables. For continuous variables the most famous distribution is the bivariate normal. When X and Y follow a bivariate normal distribution, the probability density function (pdf) is given by

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times$$

$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\} \tag{1.13}$$

In this equation we have used x and y to indicate the values taken by the variables X and Y. The lower-case letters here do *not* measure deviations from sample means, as they do in the discussion of the correlation coefficient in Section 1.2. The range of variation for both variables is from minus to plus infinity. Integrating over y in Eq. (1.13) gives the marginal distribution for X, which is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_x}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right] \tag{1.14}$$

Thus, the marginal distribution of X is seen to be normal with mean μ_x and standard deviation σ_x. Likewise, the marginal distribution of Y is normal with mean μ_y and standard deviation σ_y. The remaining parameter in Eq. (1.13) is ρ, which can be shown to be the correlation coefficient between X and Y. Finally, from the joint distribution [Eq. (1.13)] and the marginal distribution [Eq. (1.14)], the conditional distribution of Y given X may be obtained[11] as

$$\begin{aligned} f(y \mid x) &= f(x,y)/f(x) \\ &= \frac{1}{\sqrt{2\pi}\sigma_{y|x}}\exp\left[-\frac{1}{2}\left(\frac{y-\mu_{y|x}}{\sigma_{y|x}}\right)^2\right] \end{aligned} \tag{1.15}$$

The conditional distribution is also seen to be normal. The conditional mean is

$$\mu_{y|x} = \alpha + \beta x \tag{1.16}$$

$$\text{where} \quad \alpha = \mu_y - \beta\mu_x \quad \text{and} \quad \beta = \rho\frac{\sigma_y}{\sigma_x} \tag{1.17}$$

The conditional mean is thus a linear function of the X variable. The conditional variance is invariant with X and is given by

$$\sigma_{y|x}^2 = \sigma_y^2(1-\rho^2) \tag{1.18}$$

This condition of constant variance is referred to as **homoscedasticity.** Finally, the conditional mean and variance for X given Y may be obtained by interchanging x and y in the last three formulae.

1.4 THE TWO-VARIABLE LINEAR REGRESSION MODEL

In many bivariate situations the variables are treated in a symmetrical fashion. For the Scottish soldiers of Table 1.1 the conditional distribution of height, given chest

[11] See Problem 1.4.

size, is just as meaningful and interesting as the conditional distribution of chest size, given height. These are two aspects of the joint variation. However, in the vacation expenditure/income example we have already tended to show more interest in the conditional distribution of expenditure, given income, than in the distribution of income, given expenditure. This example is typical of many economic situations. Economists often have explicit notions, derived from theoretical models, of causality running from X, say, to Y. Thus, the theory of consumer behavior leads one to expect that household income will be a major determinant of household vacation expenditure, but labor economics does not give equal strength to the proposition that household vacation expenditure is a major determinant of household income. Although it is formally true that a joint distribution can always be factored in two different ways into the product of a marginal and a conditional distribution, one factorization will often be of more interest to an economist than the other. Thus, in the expenditure/income case the factorization $f(X, Y) = f(X) \cdot f(Y \mid X)$ will be of greater interest than the factorization $f(X, Y) = f(Y) \cdot f(X \mid Y)$. Moreover, in the first factorization the conditional distribution of expenditure, given income, will usually receive much more attention and analysis than the marginal distribution for income.

1.4.1 A Conditional Model

To formulate a model for vacation expenditure that is conditional on income, let us consider how data on such variables might be obtained. One possibility is that a sample of n households from the N households in the population was taken and the values of Y and X recorded for the year in question.[12] This is an example of **cross-section data.** There will be some—presumably complex and certainly unknown—bivariate distribution for all N households. This bivariate distribution itself will be some marginalization of a multivariate distribution covering income and all categories of expenditure. Concentrating on the conditional distribution, economic theory would suggest

$$E(Y \mid X) = g(X)$$

where $g(X)$ is expected to be an increasing function of X. If the conditional expectation is linear in X, as in the case of a bivariate normal distribution, then

$$E(Y \mid X) = \alpha + \beta X \tag{1.19}$$

For the ith household this expectation gives

$$E(Y \mid X_i) = \alpha + \beta X_i$$

The actual vacation expenditure of the ith household is denoted by Y_i, so we define a discrepancy or disturbance u_i as

$$u_i = Y_i - E(Y \mid X_i) = Y_i - \alpha - \beta X_i \tag{1.20}$$

[12] We now return to the earlier convention of using X and Y to indicate both the label for a variable and the values that it may assume.

The disturbance u_i must therefore represent the *net* influence of everything other than the income of the *i*th household. These other factors might include such things as the number and ages of household members, accumulated savings, and so forth. Such factors might be measured and included in Eq. (1.19), but with any finite number of explanatory factors we still cannot expect perfect agreement between individual observations and expected values. Thus, the need to specify a disturbance term remains. Taking conditional expectations of both sides of Eq. (1.20) gives $E(u_i \mid X_i) = 0$. The variance of u_i is also seen to be the variance of the conditional distribution, $\sigma^2_{y|x_i}$. If we look at the *j*th household, the disturbance u_j will have zero expectation and variance $\sigma^2_{y|x_j}$. These conditional variances may well vary with income. In the hypothetical data of Table 1.4 they are positively associated with income. For the present, however, we will make the homoscedasticity assumption that the disturbance variances are constant and independent of income. Finally, we make the assumption that the disturbances are distributed independently of one another. This rules out such things as "vacation mania," where everyone rushes off to Europe and large positive disturbances become apparent. This assumption implies that the disturbances are pairwise uncorrelated.[13] Collecting these assumptions together gives

$$\begin{aligned} E(u_i) &= 0 && \text{for all } i \\ \operatorname{var}(u_i) &= E(u_i^2) = \sigma^2 && \text{for all } i \\ \operatorname{cov}(u_i, u_j) &= E(u_i u_j) = 0 && \text{for } i \neq j \end{aligned} \tag{1.21}$$

These assumptions are embodied in the simple statement

$$\text{The } u_i \text{ are iid}(0, \sigma^2) \tag{1.22}$$

which reads "the u_i are independently and identically distributed with zero mean and variance σ^2."

Now suppose the available data come in *time series* form and that

$$\begin{aligned} X_t &= \text{aggregate real disposable personal income in year } t \\ Y_t &= \text{aggregate real vacation expenditure in year } t \end{aligned}$$

where $t = 1, 2, \ldots, n$. The series $\{X_t\}$ is no longer a set of sample values from the distribution of all N incomes in any year: it is the *actual sum* of all incomes in each

[13] Two variables are said to be independently distributed, or stochastically independent, if the conditional distributions are equal to the corresponding marginal distributions. This statement is equivalent to the joint probabilities being the product of the marginal probabilities. For the discrete case, the covariance between X and Y is then

$$\begin{aligned} \operatorname{cov}(X, Y) &= \sum_i \sum_j p_{ij}(X_i - \mu_x)(Y_j - \mu_y) \\ &= \sum_i p_{i.}(X_i - \mu_x) \sum_j p_{.j}(Y_j - \mu_y) \qquad \text{using Eq. (1.6)} \\ &= 0 \end{aligned}$$

The converse is not necessarily true since the covariance measures linear association; but substituting $\rho = 0$ in Eq. (1.13) shows that it is true for the bivariate normal distribution, since the bivariate density then collapses into the product of the two marginal densities.

year. It might be regarded as a sample of n observations from the "population" of all possible aggregate income numbers, but this interpretation seems to be putting some strain on the meaning of both *sample* and *population*. Moreover, the usual time series "sample" consists of data for *n adjacent* years. We would be rather suspicious of cross-section samples that always consisted only of n adjacent households. They could be from Millionaires' Row or from Skid Row. Thus, it is difficult to give an unambiguous and useful interpretation of $f(X)$, the marginal distribution of X over time. However, the conditional distribution $f(Y \mid X)$ is still important and must be given a probabilistic formulation. To see this reasoning, return to the cross section formulation and introduce the time subscript. Thus,

$$Y_{it} = \alpha + \beta X_{it} + u_{it} \tag{1.23}$$

where Y_{it} = real vacation expenditure by the ith household in year t
X_{it} = real disposable income of the ith household in year t

Making the (implausible) assumption that the α and β parameters are the same for all households and aggregating Eq. (1.23) over all N households in the economy, we find

$$\sum_i Y_{it} = N\alpha + \beta\left(\sum_i X_{it}\right) + \sum_i u_{it}$$

which may be rewritten as

$$Y_t = N\alpha + \beta X_t + U_t \tag{1.24}$$

where Y and X denote aggregate expenditure and aggregate income and U is an aggregate disturbance. The assumptions made about the household u's imply that U_t is a stochastic variable with zero mean and variance $N\sigma^2$. In the context of time series, one needs to make a further assumption about the independence, or lack thereof, of the U's. If the independence assumption is chosen, then the statement is that the U_t are iid$(0, N\sigma^2)$.

1.4.2 Estimates and Estimators

Whether the sample data are of cross section or time series form, the simplest version of the two-variable model is $Y_i = \alpha + \beta X_i + u_i$, with the u_i being iid$(0, \sigma^2)$. There are thus three parameters to be estimated in the model, namely, α, β, and σ^2. The parameters α and β are taken as a pair, since numerical values of both are required to fit a specific line. Once such a line has been fitted, the residuals from that line may be used to form an estimate of σ^2.

An **estimator** is a formula, method, or recipe for estimating an unknown population parameter; and an **estimate** is the numerical value obtained when sample data are substituted in the formula. The first step in fitting a straight line to sample data is to plot the scatter diagram and make sure from visual inspection that the scatter is approximately linear. The treatment of nonlinear scatters is discussed in the next chapter. Let the straight line fitted to the data be denoted by $\hat{Y}_i = a + bX_i$, where $\hat{Y}_i$ indicates the height of the line at X_i. The actual Y_i value will in general deviate from $\hat{Y}_i$. Many estimators of the pair *a,b* may be devised.

1. Fit a line by eye and read off the implied values for the intercept a and slope b. Different "artists" may, of course, draw different lines, so it is preferable to have an estimator that will yield the same result for a given data set, irrespective of the investigator.
2. Pass a line through the leftmost point and the rightmost point of the scatter. If X_* denotes the smallest value of X in the sample and X_{**} the largest and Y_*, Y_{**} the associated Y values, this estimator is

$$b = (Y_{**} - Y_*)/(X_{**} - X_*)$$
$$a = Y_* - bX_* = Y_{**} - bX_{**}$$

This estimator can hardly be expected to perform very well since it uses only two of the sample points and ignores the rest.

3. The last criticism may be met by averaging the X and Y coordinates of the m left most and the m rightmost points, where m is some integer between 1 and $n/2$, and passing a line through the resultant average points. Such an estimator with m set at $n/3$ or $n/2$ has been proposed in the literature on errors in variables, as will be discussed later. This type of estimator does not easily lend itself to mathematical manipulation, and some of its properties in repeated applications are difficult to determine.

1.4.3 Least-Squares Estimators

The dominant and powerful estimating principle, which emerged in the early years of the nineteenth century for this and other problems, is that of *least squares.*[14] Let the residuals from any fitted straight line be denoted by

$$e_i = Y_i - \hat{Y}_i = Y_i - a - bX_i \qquad i = 1, 2, \ldots, n \tag{1.25}$$

From the definition of $\hat{Y}_i$ and from Fig. 1.6 these residuals are seen to be measured in the vertical (Y) direction. Each pair of a, b values defines a different line and hence a different set of residuals. The residual sum of squares is thus a function of a and b. The least squares principle is

Select a, b to minimize the residual sum of squares.

$$\text{RSS} = \sum e_i^2 = f(a, b)$$

The necessary conditions for a stationary value of RSS are[15]

[14] See again the unfolding story in Stephen M. Stigler, *The History of Statistics*, Harvard University Press, 1986.

[15] In obtaining the derivatives we leave the summation sign in place and differentiate the typical term with respect to a and b in turn, and simply observe the rule that any constant can be moved in front of the summation sign but anything that varies from one sample point to another must be kept to the right of the summation sign. Finally, we have dropped the subscripts and range of summation since there is no ambiguity. Strictly speaking, one should also distinguish between the a and b values that appear in the expression to be minimized and the specific values that actually do minimize the residual sum of squares, but again there is little risk of ambiguity and we have kept the expressions uncluttered.

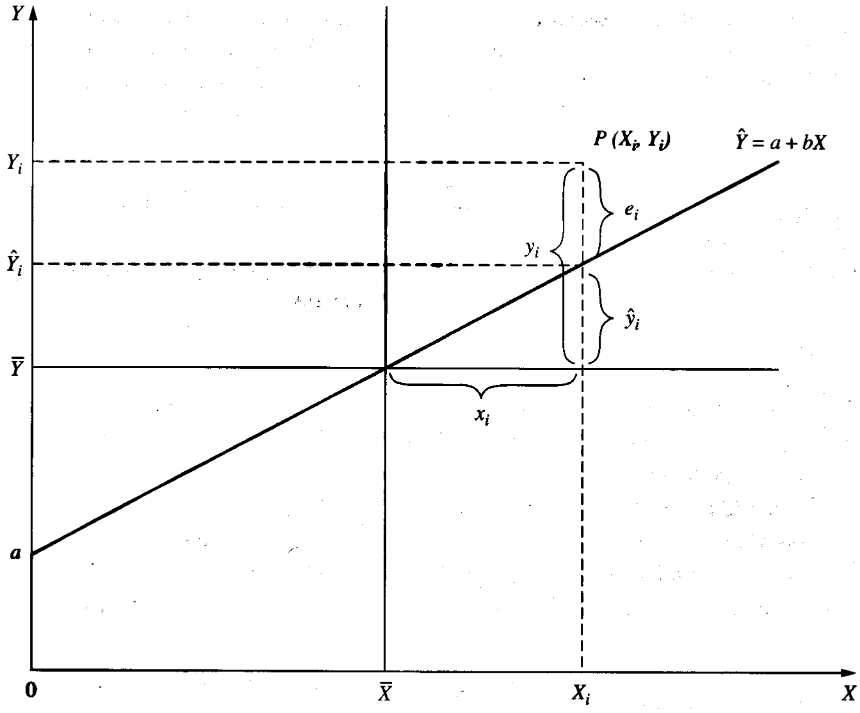


FIGURE 1.6
Residuals from a fitted straight line.

$$\frac{\partial(\sum e^2)}{\partial a} = -2\sum(Y - a - bX) = -2\sum e = 0 \tag{1.26}$$

and
$$\frac{\partial(\sum e^2)}{\partial b} = -2\sum X(Y - a - bX) = -2\sum Xe = 0 \tag{1.27}$$

Simplifying gives the **normal equations** for the linear regression of Y on X. That is,

$$\begin{aligned}\sum Y &= na + b\sum X \\ \sum XY &= a\sum X + b\sum X^2\end{aligned} \tag{1.28}$$

The reason for the adjective *normal* will become clear when we discuss the geometry of least squares later.

The first normal equation may be rewritten as

$$a = \bar{Y} - b\bar{X} \tag{1.29}$$

Substituting for a in the second normal equation gives

$$b = \frac{\sum xy}{\sum x^2} = r\frac{s_y}{s_x} \tag{1.30}$$

Thus, the least-squares slope may first of all be estimated by Eq. (1.30) from the sample deviations, and the intercept then obtained from substituting for b in Eq. (1.29). Notice that these two expressions have exactly the same form as those given in Eq. (1.17) for the intercept and slope of the conditional mean in the bivariate normal distribution. The only difference is that Eqs. (1.29) and (1.30) are in terms of sample statistics, whereas Eq. (1.17) is in terms of population statistics.

To summarize, the least-squares line has three important properties. It minimizes the sum of the squared residuals. It passes through the mean point $(\bar{X}, \bar{Y})$, as shown by Eq. (1.29). Finally, the least-squares residuals have zero correlation in the sample with the values of X.[16]

The disturbance variance σ^2 cannot be estimated from a sample of u values, since these depend on the unknown α and β values and are thus unobservable. An estimate can be based on the calculated residuals (the e_i). Two possibilities are $\sum e^2/n$ or $\sum e^2/(n-2)$. For reasons to be explained in Chapter 3 the usual choice is

$$s^2 = \frac{\sum e^2}{(n-2)} \tag{1.31}$$

1.4.4 Decomposition of the Sum of Squares

Using Eqs. (1.25) and (1.29), one may express the residuals in terms of the x, y deviations, namely

$$e_i = y_i - bx_i \qquad i = 1, 2, \ldots, n \tag{1.32}$$

Squaring both sides, followed by summing over the sample observations, gives

$$\sum e^2 = \sum y^2 - 2b\sum xy + b^2\sum x^2$$

The residual sum of squares is thus seen to be a quadratic function of b. Since $\sum x^2 \geq 0$, and the equality would only hold in the pathological case of zero variation in the X variable, the single stationary point is necessarily a minimum. Substitution from Eq. (1.30) gives

$$\begin{aligned} \sum y^2 &= b^2\sum x^2 + \sum e^2 \\ &= b\sum xy + \sum e^2 \\ &= r^2\sum y^2 + \sum e^2 \end{aligned} \tag{1.33}$$

This famous decomposition of the sum of squares is usually written as

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

[16]
$$\begin{aligned} \sum Xe &= \sum (x + \bar{X})e \\ &= \sum xe + \bar{X}\sum e \\ &= \sum xe && \text{using Eq. (1.26)} \\ \text{Hence, } \operatorname{cov}(X, e) &= 0 && \text{using Eq. (1.27)} \end{aligned}$$

where[17] TSS = total sum of squared deviations in the Y variable
RSS = residual, or unexplained, sum of squares from the regression of Y on X
ESS = explained sum of squares from the regression of Y on X

The last line of Eq. (1.33) may be rearranged to give

$$r^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = \frac{\text{ESS}}{\text{TSS}} \tag{1.34}$$

Thus, r^2 may be interpreted as the proportion of the Y variation attributable to the linear regression on X. Equation (1.34) provides an alternative demonstration that the limits of r are ±1 and that in the limiting case the sample points all lie on a single straight line.

1.4.5 A Numerical Example

Table 1.6 gives some simple data to illustrate the application of these formulae. Substitution in Eq. (1.28) then gives the normal equations

$$\begin{aligned} 40 &= 5a + 20b \\ 230 &= 20a + 120b \end{aligned}$$

with solution

$$\hat{Y} = 1 + 1.75X$$

The same data in deviation form are shown in Table 1.7. The regression coefficients may be obtained from

$$b = \frac{\sum xy}{\sum x^2} = \frac{70}{40} = 1.75$$

and $$a = \bar{Y} - b\bar{X} = 8 - 1.75(4) = 1$$

The explained sum of squares may be calculated as

$$\text{ESS} = b\sum xy = 1.75(70) = 122.5$$

and the residual sum of squares is given by subtraction as

$$\text{RSS} = \text{TSS} - \text{ESS} = 124 - 122.5 = 1.5$$

Finally, the proportion of the Y variation explained by the linear regression is

$$r^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{122.5}{124} = 0.9879$$

[17] Unfortunately there is no uniform notation for sums of squares. Some authors use SSR to indicate the sum of squares due to the regression (our ESS), and SSE to indicate the sum of squares due to error (our RSS).

TABLE 1.6

	X	Y	XY	X^2	$\hat{Y}$	e	Xe
	2	4	8	4	4.50	−0.50	−1
	3	7	21	9	6.25	0.75	2.25
	1	3	3	1	2.75	0.25	0.25
	5	9	45	25	9.75	−0.75	−3.75
	9	17	153	81	16.75	0.25	2.25
Sums	20	40	230	120	40	0	0

TABLE 1.7

	x	y	xy	x^2	y^2	$\hat{y}$	e	xe
	−2	−4	8	4	16	−3.50	−0.50	1.00
	−1	−1	1	1	1	−1.75	0.75	−0.75
	−3	−5	15	9	25	−5.25	0.25	−0.75
	1	1	1	1	1	1.75	−0.75	−0.75
	5	9	45	25	81	8.75	0.25	1.25
Sums	0	0	70	40	124	0	0	0

1.5 INFERENCE IN THE TWO-VARIABLE, LEAST-SQUARES MODEL

The least-squares (LS) estimators of α and β have been defined in Eqs. (1.28) to (1.30). There are now two important questions:

1. What are the properties of these estimators?
2. How may these estimators be used to make inferences about α and β?

1.5.1 Properties of LS Estimators

The answers to both questions depend on the **sampling distribution** of the LS estimators. A sampling distribution describes the behavior of the estimator(s) in *repeated* applications of the estimating formulae. A given sample yields a specific numerical estimate. Another sample from the same population will yield another numerical estimate. A sampling distribution describes the results that will be obtained for the estimator(s) over the potentially infinite set of samples that may be drawn from the population.

The **parameters of interest** are α, β, and σ^2 of the conditional distribution, $f(Y \mid X)$. In that conditional distribution the only source of variation from one hypothetical sample to another is variation in the stochastic disturbance (u), which in conjunction with the given X values will determine the Y values and hence the sample values of a, b, and s^2. Analyzing Y *conditional* on X thus treats the $X_1, X_2, \ldots, X_n$ values as fixed in repeated sampling. This treatment rests on the implicit assumption that the marginal distribution for X, that is, $f(X)$, does not involve the parameters of interest or, in other words, that $f(X)$ contains no information on α, β, and σ^2. This is called the **fixed regressor** case, or the case of **nonstochastic** X. From Eq. (1.30) the LS slope may be written

$$b = \sum w_i y_i$$

where the weights w_i are given by

$$w_i = \frac{x_i}{\sum x_i^2} \tag{1.35}$$

These weights are fixed in repeated sampling and have the following properties:

$$\sum w_i = 0 \qquad \sum w_i^2 = \frac{1}{\sum x_i^2} \qquad \text{and} \qquad \sum w_i x_i = \sum w_i X_i = 1 \tag{1.36}$$

It then follows that

$$b = \sum w_i Y_i \tag{1.37}$$

so that the LS slope is a linear combination of the Y values.

The sampling distribution of b is derived from Eq. (1.37) by substituting $Y_i = \alpha + \beta X_i + u_i$ and using the stochastic properties of u to determine the stochastic properties of b. Thus,

$$\begin{aligned} b &= \alpha\left(\sum w_i\right) + \beta\left(\sum w_i X_i\right) + \sum w_i u_i \\ &= \beta + \sum w_i u_i \end{aligned} \tag{1.38}$$

and so

$$E(b) = \beta \tag{1.39}$$

that is, the LS slope is an *unbiased* estimator of β. From Eq. (1.38) the variance of β is seen to be

$$\text{var}(b) = E[(b-\beta)^2] = E\left[\left(\sum w_i u_i\right)^2\right]$$

From the properties of the w's it may be shown[18] that

$$\text{var}(b) = \frac{\sigma^2}{\sum x^2} \tag{1.40}$$

By similar methods it may be shown[19] that

$$E(a) = \alpha \tag{1.41}$$

and

$$\text{var}(a) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum x^2}\right] \tag{1.42}$$

These four formulae give the means and variances of the *marginal* distributions of a and b. The two estimators, however, are in general not stochastically independent,

[18] See Appendix 1.1.

[19] See Appendix 1.2.

for the covariance is[20]

$$\operatorname{cov}(a, b) = -\frac{\sigma^2\bar{X}}{\sum x^2} \tag{1.43}$$

This covariance only vanishes if $\bar{X} = 0$. One can always rearrange an LS regression to have a zero mean for the right-hand-side variable. By using Eq. (1.29), $Y = a + bX + e$ can be rewritten as $Y = \bar{Y} + bx + e$, which gives $\operatorname{cov}(\bar{Y}, b) = \operatorname{cov}(\bar{u}, b) = 0$.

1.5.2 Gauss–Markov Theorem

The LS estimators are seen to be linear combinations of the Y variable and hence linear combinations of the stochastic u variable. Because they are also unbiased, they belong to the class of linear unbiased estimators. Their great importance in the theory and practice of statistics is that their sampling variances are the smallest that can be achieved by any linear unbiased estimator. Looking at estimators of β, for example, let

$$b^* = \sum c_i Y_i$$

denote any arbitrary linear unbiased estimator of β. The unbiasedness criterion imposes two linear constraints on the weights, (c_i), leaving $(n - 2)$ weights "free." It can be shown[21] that

$$\operatorname{var}(b^*) = \operatorname{var}(b) + \sigma^2 \sum (c_i - w_i)^2$$

Since $\sum (c_i - w_i)^2 \geq 0$, $\operatorname{var}(b^*) \geq \operatorname{var}(b)$. Equality only holds when $c_i = w_i$ for all i, that is, when $b^* = b$. The least-squares estimator thus has minimum variance in the class of linear unbiased estimators and is said to be a **best linear unbiased estimator,** or BLUE.

1.5.3 Inference Procedures

The results established so far have required the assumption that the u_i are iid$(0, \sigma^2)$. The derivation of inference procedures requires a further assumption about the form of the probability distribution of the u's. The standard assumption is that of *normality,* which may be justified by appeal to the Central Limit Theorem, since the u's represent the net effect of many separate but unmeasured influences. Linear combinations of normal variables are themselves normally distributed. Thus, the sampling distribution of a,b is bivariate normal, as in the formula in Eq. (1.13). The marginal distributions are therefore also normal and are determined by the means and variances already obtained. Thus,

$$b \sim N(\beta, \sigma^2 / \sum x^2) \tag{1.44}$$

[20] See Appendix 1.3.

[21] See Appendix 1.4.

to be read, "b is normally distributed with mean β and variance $\sigma^2/\sum x^2$." The square root of this variance, that is, the standard deviation of the sampling distribution, is often referred to as the **standard error** of b and denoted by s.e.(b).

The sampling distribution of the intercept term is

$$a \sim N\left[\alpha, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x^2}\right)\right] \tag{1.45}$$

If σ^2 were known, these results could be put to practical use. For example, a 95 percent confidence interval for β would be provided by

$$b \pm 1.96\sigma/\sqrt{\sum x^2}$$

It also follows from Eq. (1.44) that

$$z = \frac{b - \beta}{\sigma/\sqrt{\sum x^2}} \sim N(0, 1) \tag{1.46}$$

where $N(0, 1)$ denotes the standard normal distribution (a normally distributed variable with zero mean and unit variance). Thus, a test of the hypothesis H_0: $\beta = \beta_0$ is carried out by computing

$$\frac{b - \beta_0}{\sigma/\sqrt{\sum x^2}} = \frac{b - \beta_0}{\text{s.e.}(b)}$$

and contrasting this statistic with a preselected critical value from the standard normal distribution. If, for example, the absolute value of this statistic exceeded 1.96, H_0 would be rejected at the 5 percent level of significance.

When σ^2 is unknown these procedures are not feasible. T● derive an operational procedure we need two further results. They will be stated here and proved for the general case of multiple regression in Chapter 3. The relevant results are

$$\frac{\sum e^2}{\sigma^2} \sim \chi^2(n - 2) \tag{1.47}$$

to be read "$\sum e^2/\sigma^2$ is distributed as χ^2 with $(n - 2)$ degrees of freedom," and

$$\sum e^2 \text{ is distributed independently of } f(a, b) \tag{1.48}$$

As shown in Appendix B the t distribution is defined as a combination of a standard normal variable and an independent χ^2 variable. Thus, Eqs. (1.46) through (1.48) give

$$\frac{b - \beta}{s/\sqrt{\sum x^2}} \sim t(n - 2) \tag{1.49}$$

where $s^2 = \sum e^2/(n - 2)$, the estimator of σ^2 defined in Eq. (1.31). Notice that Eq. (1.49) has the same structure as Eq. (1.46), the only difference being that the unknown σ is replaced by the estimate s. This causes a shift from the normal distribution to the t distribution. For degrees of freedom in excess of about 30, the differences between the critical values of the t distribution and the standard normal distribution are negligible. A 95 percent confidence interval for β is

$$b \pm t_{0.025} s/\sqrt{\textstyle\sum x^2} \tag{1.50}$$

and H_0: $\beta = \beta_0$ would be rejected if

$$\left| \frac{b - \beta_0}{s/\sqrt{\sum x^2}} \right| > t_{0.025}(n-2) \tag{1.51}$$

where $t_{0.025}(n-2)$ indicates the 2.5 percent point of the t distribution with $(n-2)$ degrees of freedom.

The conditions in Eqs. (1.50) and (1.51) are opposite sides of the same coin. If Eq. (1.51) leads to a rejection of H_0, then β_0 lies outside the confidence interval given in Eq. (1.50). Likewise, if β_0 lies inside the 95 percent confidence interval, Eq. (1.51) will not lead to the rejection of H_0 at the 5 percent level. The most commonly used test of significance is that of H_0: $\beta = 0$. The test statistic is then

$$t = \frac{b}{s/\sqrt{\sum x^2}} = \frac{b}{\text{s.e.}(b)} \tag{1.52}$$

and H_0 would be rejected at the 5 percent level of significance if the absolute value of b exceeded $t_{0.025}$ times its standard error. The abbreviation s.e.(b) is used to denote both the true and the estimated standard error of β. The test statistic in Eq. (1.52) is a routine output of most regression packages and usually has some label such as T-STAT. Many programs also report a P-value, which is the probability of obtaining a coefficient as far or farther from zero as the sample value if, in fact, the true value of the coefficient is zero. This number may be labeled P-VALUE or 2-TAIL SIG.

By a similar development, tests on the intercept are based on the t distribution:

$$\frac{a - \alpha}{s\sqrt{1/n + \bar{X}^2/\sum x^2}} \sim t(n-2) \tag{1.53}$$

Thus, a $100(1-\epsilon)$ percent confidence interval for α is given by

$$a \pm t_{\epsilon/2} s\sqrt{1/n + \bar{X}^2/\textstyle\sum x^2} \tag{1.54}$$

and the hypothesis H_0: $\alpha = \alpha_0$ would be rejected at the 100ϵ percent level of significance if

$$\left| \frac{a - \alpha_0}{s\sqrt{1/n + \bar{X}^2/\sum x^2}} \right| > t_{\epsilon/2}$$

Tests on σ^2 may be derived from the result stated in Eq. (1.47). Using that result one may, for example, write

$$\text{prob}\left[\chi^2_{0.025} < \frac{(n-2)s^2}{\sigma^2} < \chi^2_{0.975}\right] = 0.95 \tag{1.55}$$

which states that 95 percent of the values of a χ^2 variable will lie between the values that cut off 2.5 percent in each tail of the distribution. The critical values are read off from the χ^2 distribution with $(n-2)$ degrees of freedom, or accessed through any appropriate software package. The only unknown in Eq. (1.55) is σ^2, and the contents of the probability statement may be rearranged to give a 95 percent confidence interval for σ^2 as

$$\frac{(n-2)s^2}{\chi^2_{0.975}} \quad \text{to} \quad \frac{(n-2)s^2}{\chi^2_{0.025}}$$

1.5.4 Numerical Example (Continued from Section 1.4.5)

From the data in Tables 1.6 and 1.7 we have already calculated

$$n = 5 \qquad a = 1 \qquad b = 1.75$$
$$\text{TSS} = 124 \qquad \text{ESS} = 122.5 \qquad \text{RSS} = 1.5 \qquad r^2 = 0.9879$$

We now obtain

$$s^2 = \text{RSS}/(n-2) = 1.5/3 = 0.5$$
$$\text{var}(b) = s^2/\sum x^2 = 0.5/40 = 0.0125$$
$$\text{var}(a) = 0.5\left(\frac{1}{5} + \frac{16}{40}\right) = 0.3$$

The estimated standard errors of the regression coefficients are thus

$$\text{s.e.}(a) = \sqrt{0.3} = 0.5477 \qquad \text{s.e.}(b) = \sqrt{0.0125} = 0.1118$$

A preselected critical value from the t distribution with 3 degrees of freedom is $t_{0.025} = 3.182$. Thus, a 95 percent confidence interval for α is

$$1 \pm 3.182(0.5477)$$

that is,

$$-0.74 \quad \text{to} \quad 2.74$$

and a 95 percent confidence interval for β is

$$1.75 \pm 3.182(0.1118)$$

that is,

$$1.39 \quad \text{to} \quad 2.11$$

The intercept is not significantly different from zero since

$$\frac{a}{\text{s.e.}(a)} = \frac{1}{0.5477} = 1.826 < 3.182$$

whereas the slope is strongly significant since

$$\frac{b}{\text{s.e.}(b)} = \frac{1.75}{0.1118} = 15.653 > 3.182$$

As indicated earlier, once confidence intervals have been computed, actually computing the significance tests is unnecessary, since a confidence interval that includes zero is equivalent to accepting the hypothesis that the true value of the parameter is zero, and an interval that does not embrace zero is equivalent to rejecting the null hypothesis.

From the χ^2 distribution with 3 degrees of freedom $\chi^2_{0.025} = 0.216$ and $\chi^2_{0.975} = 9.35$. We also have $\sum e^2 = 1.5$. Thus, a 95 percent confidence interval for σ^2 is

$$\frac{1.5}{9.35} \quad \text{to} \quad \frac{1.5}{0.216}$$

that is,

$$0.16 \quad \text{to} \quad 6.34$$

1.6 ANALYSIS OF VARIANCE IN THE TWO-VARIABLE REGRESSION MODEL

The test for the significance of X, (H_0: $\beta = 0$), derived in the previous section may also be set out in an analysis of variance framework, and this alternative approach will be especially helpful later when we treat problems of multiple regression.

We have seen in Eq. (1.46) that the ratio of $(b - \beta)$ to the true standard error of b is a standard normal variable. From the definition of the χ^2 variable in Appendix B it follows that

$$\frac{(b-\beta)^2}{\sigma^2/\sum x^2} \sim \chi^2(1)$$

It has also been stated in Eq. (1.47) that

$$\frac{\sum e^2}{\sigma^2} \sim \chi^2(n-2)$$

and that this statistic is distributed independently of b. Recalling from Appendix B that the ratio of two independent χ^2 variables, each divided by the associated degrees of freedom, follows the F distribution, we then have

$$F = \frac{(b-\beta)^2 \sum x^2}{\sum e^2/(n-2)} \sim F(1, n-2) \tag{1.56}$$

As with the shift from the normal to the t distribution in the preceding section, this development has the felicitous result that the unknown σ^2 disappears from the expression for F. To test the hypothesis H_0: $\beta = 0$, we make this substitution in Eq. (1.56), giving

$$F = \frac{b^2 \sum x^2}{\sum e^2/(n-2)} \sim F(1, n-2) \tag{1.57}$$

By referring to the decomposition of the sum of squares in Eq. (1.33), the F statistic in Eq. (1.57) is seen to be

$$F = \frac{\text{ESS}/1}{\text{RSS}/(n-2)} \tag{1.58}$$

TABLE 1.8
ANOVA for two-variable regression

Source of variation (1)	Sums of squares (2)	Degrees of freedom (3)	Mean squares (4)
X	$\text{ESS} = b^2 \sum x^2$	1	ESS/1
Residual	$\text{RSS} = \sum e^2$	$(n-2)$	$\text{RSS}/(n-2)$
Total	$\text{TSS} = \sum y^2$	$(n-1)$	

Following this approach, we can set out the data in an analysis of variance (ANOVA) table (Table 1.8). The entries in columns 2 and 3 of the table are additive. The mean squares in the final column are obtained by dividing the sum of squares in each row by the corresponding number of degrees of freedom. An intuitive explanation of the degrees of freedom concept is that it is equal to the number of values that may be set arbitrarily. Thus, we may set $n-1$ values of y at will, but the nth is then determined by the condition that $\sum y = 0$. Likewise, we may set $n-2$ values of e at will, but the least-squares fit imposes two conditions on e, namely, $\sum e = \sum Xe = 0$, and finally there is only 1 degree of freedom attached to the explained sum of squares since that depends only on a single parameter β.

The F statistic in Eq. (1.58) is seen to be the ratio of the mean square due to X to the residual mean square. The latter may be regarded as a measure of the "noise" in the system, and thus an X effect is only detected if it is greater than the inherent noise level. The significance of X is thus tested by examining whether the sample F exceeds the appropriate critical value of F taken from the *upper tail* of the F distribution. The test procedure is then as follows: Reject $H_0\colon \beta = 0$ at the 5 percent level of significance if

$$F = \frac{\text{ESS}/1}{\text{RSS}/(n-2)} > F_{0.95}(1, n-2)$$

where $F_{0.95}$ indicates the value of F such that just 5 percent of the distribution lies to the right of the ordinate at $F_{0.95}$. Other levels of significance may be handled in a similar fashion.

For the simple numerical example from Tables 1.6 and 1.7

$$\text{Sample } F = 122.5/0.5 = 245.0$$

and $F_{0.95}(1, 3) = 10.1$. Thus, we reject $H_0\colon \beta = 0$, as before. We now have two ways of testing the significance of X, one based on a statistic that follows a t distribution, and another based on a statistic with an F distribution. The tests, however, are identical since, as shown in Appendix B,

$$t^2(m) = F(1, m) \tag{1.59}$$

A little arithmetic shows that this relation holds for both sample statistics and critical values in the foregoing example. There is also a *third* version of this test, which is sometimes used. By using the decomposition of the sum of squares in Eq. (1.33), the F statistic in Eq. (1.58) may be written

$$F = \frac{r^2(n-2)}{(1-r^2)} \tag{1.60}$$

Taking the square root will give a *t* statistic,

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \tag{1.61}$$

Either statistic may be used and referred to the appropriate *F* or *t* distribution to test the significance of *X*. Thus, we may base a test on the correlation coefficient, on the regression slope, or on the decomposition of the sum of squares; but all three approaches give the same answer to the single question: Does *X* play a statistically significant role in the explanation of *Y*?

1.7 PREDICTION IN THE TWO-VARIABLE REGRESSION MODEL

After having estimated a regression line from the sample of *n* observations, our interest often centers on some specific value X_0 of the regressor variable and we are required to predict the value Y_0 likely to be associated with X_0. For instance, if *Y* denotes the consumption of gasoline and *X* the price, we might be interested in predicting the demand for gasoline at some future higher price. The value of X_0 may lie within the range of sample *X* values or, more frequently, we may be concerned with predicting *Y* for a value of *X* *outside* the sample observations. In either case the prediction involves the assumption that the relationship presumed to have generated the sample data still holds for the new observation, whether it relates to a future time period or to a unit that was not included in a sample cross section. Alternatively, we may have a new observation (X_0, Y_0), where the actual *Y* value is known, and the question arises whether this observation may be presumed to have come from the same population as the sample data. For example, does the introduction of a speed limit reduce the demand for gasoline, conditional on the price being charged? Prediction theory enables us to address both questions. We may make two kinds of predictions, a *point* prediction or an *interval* prediction, in just the same way as we can give a point estimate or an interval estimate of a parameter β. But in practice, a point estimate is of little use without some indication of its precision, so one should always provide an estimate of the prediction error.

The point prediction is given by the regression value corresponding to X_0, that is,

$$\hat{Y}_0 = a + bX_0 = \bar{Y} + bx_0 \tag{1.62}$$

where $x_0 = X_0 - \bar{X}$. The true value of *Y* for the prediction period or observation is

$$Y_0 = \alpha + \beta X_0 + u_0$$

The *average* value of *Y* taken over the *n* sample observations is

$$\bar{Y} = \alpha + \beta\bar{X} + \bar{u}$$

Subtracting gives

$$Y_0 = \bar{Y} + \beta x_0 + u_0 - \bar{u} \tag{1.63}$$

The prediction error is defined as

$$e_0 = Y_0 - \hat{Y}_0 = -(b - \beta)x_0 + u_0 - \bar{u} \tag{1.64}$$

Clearly the expected prediction error is zero, so $\hat{Y}_0$ is a linear unbiased predictor of Y_0. The variance of e_0 may be shown[22] to be

$$\text{var}(e_0) = \sigma^2\left(1 + \frac{1}{n} + \frac{x_0^2}{\sum x^2}\right) \tag{1.65}$$

It will be shown in Chapter 3 that the error variance in Eq. (1.65) is a minimum in the class of linear unbiased predictors. Thus, the optimum property of LS estimates of coefficients carries over to prediction based on the LS regression.

From Eq. (1.64) we see that e_0 is a linear combination of normally distributed variables (b, u_0, and $\bar{u}$). Thus, it is also normally distributed, and so

$$\frac{e_0}{\sigma\sqrt{1 + 1/n + x_0^2/\sum x^2}} \sim N(0, 1)$$

Replacing the unknown σ^2 by its estimate s^2 then gives

$$\frac{Y_0 - \hat{Y}_0}{s\sqrt{1 + 1/n + (X_0 - \bar{X})^2/\sum x^2}} \sim t(n-2) \tag{1.66}$$

Everything in Eq. (1.66) is known except Y_0, and so, in the usual way, we derive a 95 percent confidence interval for Y_0 as

$$(a + bX_0) \pm t_{0.025}s\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2}} \tag{1.67}$$

Even if we know α and β with certainty, there is an inherent element of uncertainty in predicting Y_0, owing to the random drawing u_0 that occurs in the prediction period. Thus, interest often centers more realistically on the prediction of the *mean* value of Y_0, that is,

$$E(Y_0) = \alpha + \beta X_0$$

This eliminates the term u_0 from the prediction error. Following through the same analysis gives a 95 percent confidence interval for $E(Y_0)$ as

$$(a + bX_0) \pm t_{0.025}s\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2}} \tag{1.68}$$

The width of the confidence interval in both Eqs. (1.67) and (1.68) is seen to increase symmetrically the further X_0 is from the sample mean $\bar{X}$.

[22] See Appendix 1.5.

If we continue the previous numerical example, a 95 percent confidence interval for Y conditional on $X = 10$ is

$$1 + 1.75(10) \pm 3.182\sqrt{0.5}\sqrt{1 + \frac{1}{5} + \frac{(10-4)^2}{40}}$$

which is

$$15.24 \quad \text{to} \quad 21.76$$

The 95 percent interval for $E(Y \mid X = 10)$ is

$$18.5 \pm 3.182\sqrt{0.5}\sqrt{\frac{1}{5} + \frac{(10-4)^2}{40}}$$

which is

$$16.14 \quad \text{to} \quad 20.86$$

To test whether a new observation (X_0, Y_0) comes from the structure generating the sample data, one contrasts the observation with the confidence interval for Y_0. For example, the observation (10, 25) gives a Y value that lies outside the interval 15.24 to 21.76; and one would reject at the 5 percent level the hypothesis that it came from the same structure as the sample data.

1.8 GASOLINE CONSUMPTION: A PRELIMINARY ANALYSIS

This preliminary look at gasoline consumption cannot be expected to be economically realistic, since we are currently restricted to using just one explanatory variable. One does not need to be a Nobel laureate in economics to suspect that price and income *both* influence consumption. The main purpose is to illustrate the various descriptive and test statistics in a typical computer printout.

Consider first the price and gasoline consumption (GAS) scatter for the period 1959.1 to 1973.3, shown in Fig. 1.3*b*. Fitting a linear regression to these series gives the results in Table 1.9.[23]

The dependent variable and the sample time span are indicated at the top of the table. Then follows the estimated intercept (coefficient on the C variable) and the estimated coefficient on the regressor PRICE, or (X2), with their standard errors and the t statistics for testing the hypothesis that the true value of each coefficient is zero. *R-squared* is the statistic defined in Eq. (1.34). *Adjusted R-squared* will be explained in Chapter 3. *S.E. of regression* is the square root of the statistic defined in Eq. (1.31), and *Sum squared resid* is the residual sum of squares (RSS). The *log likelihood* will be discussed in Chapter 2, and the *Durbin-Watson statistic* in Chapter 6. The *Akaike* and *Schwarz information criteria* will be discussed in Chapter 3. Finally the *F-statistic*, defined in three equivalent forms in Eqs. (1.56), (1.57), and

[23] This regression output, like most of the empirical results in Chapters 1 to 9, comes from EViews, a Windows software program from Quantitative Micro Software, Irvine, California.

TABLE 1.9

Regression of gasoline consumption on price, 1959.1 to 1973.3

LS // Dependent Variable is GAS
Sample: 1951:1 1973:3
Included observations: 59

Variable	Coefficient	Std. Error	T-Statistic	Prob.
C	2.121645	0.548643	3.867078	0.0003
X2	−2.150563	0.118430	−18.15899	0.0000

R-squared	0.852618	Mean dependent var	−7.840375
Adjusted R-squared	0.850032	S.D. dependent var	0.136145
S.E. of regression	0.052723	Akaike info criterion	−5.852095
Sum squared resid	0.158444	Schwarz criterion	−5.781670
Log likelihood	90.91943	F-statistic	329.7489
Durbin-Watson stat	0.290306	Prob(F-statistic)	0.000000

(1.58), tests the significance of the overall regression, which in the two-variable case is equivalent to testing the significance of the single regressor. The numerical value of the F statistic is the square of the t statistic on the regressor, as may be checked from the regression output. The *Prob(F-statistic)* is the P-value that tests the null hypothesis of no relationship between the two variables, and this hypothesis is seen to be decisively rejected.

A similar regression for the period 1982.1 to 1992.1, for which the scatter is shown in Fig. 1.3*d*, gives the results shown in Table 1.10. The fit is much less good than in the earlier period, and the price coefficient is numerically much smaller. However, these two-variable regressions have no economic significance. A proper analysis requires a *multivariate* approach with attention to the *dynamics* of the demand function, and this will be attempted in Chapter 8.

TABLE 1.10

Regression of gasoline consumption on price, 1982.1 to 1992.1

LS // Dependent Variable is GAS
Sample: 1982:1 1992:1
Included observations: 41

Variable	Coefficient	Std. Error	T-Statistic	Prob.
C	−7.055535	0.091918	−76.75925	0.0000
X2	−0.137704	0.019344	−7.118512	0.0000

R-squared	0.565088	Mean dependent var	−7.709410
Adjusted R-squared	0.553936	S.D. dependent var	0.032420
S.E. of regression	0.021653	Akaike info criterion	−7.617707
Sum squared resid	0.018285	Schwarz criterion	−7.534118
Log likelihood	99.98651	F-statistic	50.67321
Durbin-Watson stat	0.669097	Prob(F-statistic)	0.000000

APPENDIX

APPENDIX 1.1
To prove var(b) = $\sigma^2/\sum x^2$

As shown in the text

$$\operatorname{var}(b) = E\left[\left(\sum w_i u_i\right)^2\right]$$

where the w_i and their properties are defined in Eqs. (1.35) and (1.36). Expanding the right-hand side

$$\begin{aligned}\left(\sum w_i u_i\right)^2 &= (w_1 u_1 + w_2 u_2 + \cdots + w_n u_n)^2 \\ &= w_1^2 u_1^2 + w_2^2 u_2^2 + \cdots + w_n^2 u_n^2 \\ &\quad + 2 w_i w_j u_i u_j + \cdots\end{aligned}$$

where all the remaining terms are cross products of the form $w_i w_j u_i u_j$ $(i < j)$. Taking expectations, we find

$$\begin{aligned}E\left[\left(\sum w_i u_i\right)^2\right] &= \sigma^2 \sum w_i^2 \\ &= \frac{\sigma^2}{\sum x^2}\end{aligned}$$

by virtue of Eqs. (1.21) and (1.36).

APPENDIX 1.2
To derive the mean and variance of the sampling distribution of a

From Eq. (1.29)

$$\begin{aligned}a &= \bar{Y} - b\bar{X} \\ &= \alpha + \beta\bar{X} + \bar{u} - b\bar{X} \\ &= \alpha - (b - \beta)\bar{X} + \bar{u}\end{aligned}$$

Since $E(b) = \beta$ and $E(\bar{u}) = 0$, it follows that

$$E(a) = \alpha$$

Then

$$\begin{aligned}\operatorname{var}(a) &= E[(a - \alpha)^2] \\ &= \bar{X}^2 E[(b - \beta)^2] + E[\bar{u}^2] - 2\bar{X}E[(b - \beta)\bar{u}]\end{aligned}$$

Now

$$E[(b - \beta)^2] = \sigma^2 / \sum x^2$$

and

$$E[\bar{u}^2] = \sigma^2/n$$

since $\bar{u}$ is the mean of a random sample of n drawings from the u distribution, which has zero mean and variance σ^2. Finally,

$$\begin{aligned} E[(b-\beta)\bar{u}] &= E\left[\left(\sum w_i u_i\right)\left(\frac{1}{n}\sum u_i\right)\right] \\ &= E\left[\frac{1}{n}\left(\sum w_i u_i^2 + \text{cross-product terms in } u_i u_j\right)\right] \\ &= \frac{1}{n}\sigma^2 \sum w_i \\ &= 0 \end{aligned}$$

Thus,

$$\operatorname{var}(a) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum x^2}\right]$$

APPENDIX 1.3
To derive cov(a, b)

$$\begin{aligned} \operatorname{cov}(a,b) &= E[(a-\alpha)(b-\beta)] \\ &= E[(\bar{u} - (b-\beta)\bar{X})(b-\beta)] \\ &= E[(b-\beta)\bar{u}] - \bar{X}E[(b-\beta)^2] \\ &= -\frac{\sigma^2\bar{X}}{\sum x^2} \end{aligned}$$

APPENDIX 1.4
Gauss–Markov theorem

A linear estimator of β is $b^* = \sum c_i Y_i$, where the c_i are to be determined. Unbiasedness requires $E(b^*) = \beta$. Now

$$\begin{aligned} b^* &= \sum c_i(\alpha + \beta X_i + u_i) \\ &= \alpha\left(\sum c_i\right) + \beta\left(\sum c_i X_i\right) + \sum c_i u_i \end{aligned}$$

Thus, b^* will be a linear unbiased estimator if and only if

$$\sum c_i = 0 \qquad \text{and} \qquad \sum c_i X_i = \sum c_i x_i = 1$$

When these conditions are satisfied

$$b^* = \beta + \sum c_i u_i$$

and

$$\operatorname{var}(b^*) = E\left[\left(\sum c_i u_i\right)^2\right] = \sigma^2 \sum c_i^2$$

To compare this variance with that of the least squares b, write

$$c_i = w_i + (c_i - w_i)$$

Thus, $$\sum c_i^2 = \sum w_i^2 + \sum (c_i - w_i)^2 + 2\sum w_i(c_i - w_i)$$

The properties of the w_i and the conditions on the c_i ensure that

$$\sum w_i(c_i - w_i) = 0$$

and so $$\text{var}(b^*) = \text{var}(b) + \sigma^2 \sum (c_i - w_i)^2$$

as given in the text, which proves the theorem.

APPENDIX 1.5
To derive var(e_0)

From Eq. (1.64)

$$e_0 = -(b - \beta)x_0 + u_0 - \bar{u}$$

Square both sides and take expectations. The expectations of all three cross-product terms vanish. The independence of the u's means that u_0 is uncorrelated with $\bar{u}$ and also with b, which is a function of u_1 to u_n, and we have seen in 1.3 above that $E[(b - \beta)\bar{u}] = 0$. Thus

$$\begin{aligned} \text{var}(e_0) &= E(u_0^2) + E[\bar{u}^2] + x_0^2 E[(b - \beta)^2] \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{x_0^2}{\sum x^2}\right] \end{aligned}$$

PROBLEMS

1.1. How might the volatility of the savings and income series in Fig. 1.1*a* be measured? A possible measure is the **coefficient of variation,** which is the standard deviation of a series expressed as a percentage of the mean. From your software package compute the coefficient of variation for each series. Also compute the savings ratio (savings/income) and comment on its movement over time.

1.2. Verify the conditional means in Table 1.2.

1.3. Verify the equivalence of the three expressions for the correlation coefficient in Eq. (1.3).

1.4. Verify the results for the conditional normal distribution in Eqs. (1.15), (1.16), and (1.17) by taking the ratio of Eq. (1.13) to Eq. (1.14).

1.5. Compute the correlation coefficients for the various scatter plots in Fig. 1.3 and comment on your results.

1.6. Carry out regression analyses for the data on gasoline consumption and price (illustrated in Fig. 1.3) using any subperiods you consider appropriate and comment on your results.

1.7. A container holds three balls numbered 1, 2, and 3. A ball is drawn at random and the number (X) noted. A second ball is drawn at random and the number (Y) noted. Construct tables showing the bivariate distribution $f(X, Y)$ when the drawing is

(a) with replacement

(b) without replacement

In each case compute the conditional means and correlation coefficient.

1.8. Three discrete variables, X, Y, and Z have possible values:

$$X \quad 1,3 \qquad Y \quad 2,5 \qquad Z \quad 4,8$$

The trivariate distribution is shown in the accompanying table. Construct the bivariate distribution $f(X, Y)$ obtained by "integrating out" Z from the trivariate distribution. In this case the bivariate distribution is itself a marginal distribution, the result of a marginalization with respect to Z.

X	1	1	1	1	3	3	3	3
Y	2	2	5	5	2	2	5	5
Z	4	8	4	8	4	8	4	8
$f(X, Y, Z)$	.2	0	.1	.2	.1	0	.3	.1

1.9. This problem requires the derivation of a sampling distribution from first principles. The postulated relation is

$$Y_i = \beta X_i + u_i \qquad \text{with} \qquad i = 1, 2 \quad \text{and} \quad X_1 = 1, X_2 = 2$$

The sample size is 2, and the bivariate probability distribution for u is given in the accompanying table.

		u_1	
		−1	**1**
u_2	**−2**	.25	.25
	2	.25	.25

Verify that the u's have zero mean and zero covariance. Derive the sampling distribution of the estimator

$$b = \sum X_i Y_i / \sum X_i^2$$

Show that the expected value of the estimator is β, and compute the variance of the sampling distribution. (This problem comes from "A Simple Approach to Teaching Generalized Least Squares Theory," by E. H. Oksanen, *The American Statistician*, **45**, August 1991, 229–233.)

1.10. The fixed values of X in a problem are as follows:

X_1	X_2	X_3	X_4	X_5	X_6
1	2	3	4	5	6

An econometrician with no calculator and an aversion to arithmetic proposes to estimate the slope of the linear relation between Y and X by

$$\tfrac{1}{8}(Y_6 + Y_5 - Y_2 - Y_1)$$

Show that this estimator is unbiased. Deduce its sampling variance and compare this with the sampling variance of the least-squares estimator of the slope.

We know that the least-squares estimator has minimum variance in the class of linear unbiased estimators. The ratio of the two variances, with that of the LS estimator in the numerator, defines the **efficiency** of the alternative estimator. Compute the efficiency of the lazy econometrician's estimator.

1.11. Let us consider the "other" or "reverse" regression. The process of defining residuals in the X direction and minimizing the new RSS gives the regression of X on Y. There are thus at least two possible regressions, but only one r^2 can be computed for a given data set. What is the interpretation of r^2 in the regression of X on Y? Prove that

$$r^2 = b_{yx}b_{xy}$$

where the b's are the LS slopes in the respective regressions. Hence, show that

$$b_{yx} \leq 1/b_{xy}$$

(provided both slopes are positive) and that, viewed in the X, Y plane with Y on the vertical axis and X on the horizontal, the regression of Y on X will have a smaller slope than the regression of X on Y. What can you say if both slopes are negative? Notice that in any specific example, both slopes must have the same sign, since the sign of the slope is given by the sign of the sample covariance.

From a sample of 200 observations the following quantities were calculated:

$$\sum X = 11.34 \qquad \sum Y = 20.72$$

$$\sum X^2 = 12.16 \qquad \sum Y^2 = 84.96 \qquad \sum XY = 22.13$$

Estimate both regression equations, compute r^2, and confirm the foregoing statements.

1.12. Show that if r is the correlation coefficient between n pairs of variables (X_i, Y_i), then the squared correlation between the n pairs $(aX_i + b, cY_i + d)$, where a, b, c, and d are constants, is also r^2.

1.13. Data on aggregate income Y and aggregate consumption C yield the following regressions, expressed in deviation form:

$$\hat{y} = 1.2c$$
$$\hat{c} = 0.6y$$

If Y is identically equal to $C + Z$, where Z is aggregate saving, compute the correlation between Y and Z, the correlation between C and Z, and the ratio of the standard deviations of Z and Y.

1.14. The accompanying table gives the means and standard deviations of two variables X and Y and the correlation between them for each of two samples. Calculate the correlation between X and Y for the composite sample consisting of the two samples taken together. Why is this correlation smaller than either of the correlations in the subsamples?

Sample	Number in sample	$\bar{X}$	$\bar{Y}$	s_x	s_y	r_{xy}
1	600	5	12	2	3	0.6
2	400	7	10	3	4	0.7

1.15. An investigator is interested in the accompanying two series for 1935–1946.

Year	35	36	37	38	39	40	41	42	43	44	45	46
X, deaths of children under 1 year (000)	60	62	61	55	53	60	63	53	52	48	49	43
Y, consumption of beer (bulk barrels)	23	23	25	25	26	26	29	30	30	32	33	31

(*a*) Calculate the coefficient of correlation between X and Y.

(*b*) A linear *time trend* may be fitted to X (or Y) by calculating an LS regression of X (or Y) on time t. This process requires choosing an origin and a unit of measurement for the time variable. For example, if the origin is set at mid-1935 and the unit of measurement is 1 year, then the year 1942 corresponds to $t = 7$, and so forth for the other years. If the origin is set at end-1940 (beginning of 1941) and the unit of measurement is 6 months, then 1937 corresponds to $t = -7$. Show that any computed trend value $\hat{X}_t = a + bt$ is unaffected by the choice of origin and unit of measurement.

(*c*) Let $e_{x.t}$ and $e_{y.t}$ denote the residuals of X and Y from their trend values. Calculate the correlation coefficient between $e_{x.t}$ and $e_{y.t}$. Compare this value with that obtained in part (*a*), and comment on the difference.

1.16. A sample of 20 observations corresponding to the model

$$Y = \alpha + \beta X + u$$

where the u's are normally and independently distributed with zero mean and constant variance, gave the following data:

$$\sum Y = 21.9 \qquad \sum (Y - \bar{Y})^2 = 86.9 \qquad \sum (X - \bar{X})(Y - \bar{Y}) = 106.4$$
$$\sum X = 186.2 \qquad \sum (X - \bar{X})^2 = 215.4$$

Estimate α and β and calculate their standard errors. Estimate the conditional mean value of Y corresponding to $X = 10$ and find a 95 percent confidence interval for this mean.

CHAPTER 2

Further Aspects of Two-Variable Relationships

Chapter 1 presented a set of inference procedures associated with least-squares (LS) estimators in the context of bivariate relationships. The derivation of these procedures was based on two crucial assumptions, one about the form of the conditional expectation $E(Y \mid X)$ and the other about the stochastic properties of the disturbance term u. The specific assumptions were

$$E(Y \mid X) = \alpha + \beta X \tag{2.1}$$

and

$$\begin{aligned} E(u_i) &= 0 \qquad \text{for all } i \\ E(u_i^2) &= \sigma^2 \qquad \text{for all } i \\ E(u_i u_j) &= 0 \qquad \text{for } i \neq j \end{aligned} \tag{2.2}$$

It also follows from Eq. (2.2) and the fixed regressor assumption that

$$E(X_i u_j) = X_i E(u_j) = 0 \qquad \text{for all } i, j \tag{2.3}$$

Adding the assumption of normality to Eq. (2.2) gives

$$\text{The } u_i \text{ are iid } N(0, \sigma^2) \tag{2.4}$$

which reads, "The u_i are independently and identically distributed normal variables with zero mean and variance σ^2." The validity of the inference procedures obviously depends on the correctness of the underpinning assumptions.

Most of this chapter deals with various possible respecifications of the conditional expectation assumption in Eq. (2.1). We will first look at some of the issues raised when the regressor (explanatory) variable is *time*. This leads naturally to the consideration of **constant growth curves,** where the *logarithm* of the dependent variable is expressed as a linear function of time. We then consider cases where transformations of the dependent and/or explanatory variable may be useful. Many relationships that are nonlinear in the original variables may be *linearized* by suitable transformations. In such cases the simple techniques developed in Chapter 1 for the linear model may be applied to the transformed variables.

Next we consider the bivariate model where the explanatory variable is simply the lagged value of the dependent variable. This is the first-order, autoregressive AR(1) scheme. The change seems innocuous enough but it moves us into fundamentally new territory. The least-squares estimators are no longer unbiased; and the exact, finite sample results of Chapter 1 are no longer strictly valid. The least-squares procedures of Chapter 1 can still be applied to the autoregressive equation but they now have only a **large-sample,** or **asymptotic** validity. An appreciation of this result requires an introduction to some basic and very important ideas relating to **large-sample theory,** namely, **asymptotic distributions, asymptotic efficiency, and consistency.** The simple autoregressive equation also raises the issue of the **stationarity** of a time series. These issues will be developed extensively in later chapters, especially Chapters 5, 7, and 8. We hope their introduction in the context of a two-variable relation will keep the initial exposition as simple as possible and serve as a bridge to the later treatment.

2.1 TIME AS A REGRESSOR

In a time series plot, as shown in Chapter 1, a variable Y_t on the vertical axis is plotted against time on the horizontal axis. This may also be regarded as a scatter plot, the only difference from the conventional scatter plot being that the X (time) variable increases monotonically by one unit with each observation. Many economic variables increase or decrease with time. A **linear trend** relationship would be modeled as

$$Y = \alpha + \beta T + u \tag{2.5}$$

where T indicates time. The T variable may be specified in many fashions, but each specification requires one to define the *origin* from which time is measured and the *unit of measurement* that is used. For example, if we had annual observations on some variable for the ($n = 13$) years from 1980 to 1992, possible specifications of the T variable would be

$$T = 1980, 1981, 1982, \ldots, 1992$$
$$T = 1, 2, 3, \ldots, 13$$
$$T = -6, -5, -4, \ldots, 6$$

In all three cases the unit of measurement is a year. The origins are, respectively, the start of the Gregorian calendar, 1979, and 1986. The third scheme is advantageous for small-scale calculations since in this case T has zero mean, so the normal equations for fitting Eq. (2.5) simplify to

$$a = \bar{Y} \qquad \text{and} \qquad b = \sum TY/\sum T^2$$

Many software packages will generate a TREND variable for use in regression analysis. This is the second specification for T above.

2.1.1 Constant Growth Curves

Taking first differences of Eq. (2.5) gives

$$\Delta Y_t = \beta + (u_t - u_{t-1})$$

If we ignore the disturbances, the implication of Eq. (2.5) is that the series increases (decreases) by a constant amount each period. For an increasing series ($\beta > 0$), this implies a *decreasing* growth rate, and for a decreasing series ($\beta < 0$), the specification gives an *increasing* decline rate. For series with an underlying *constant* growth rate, whether positive or negative, Eq. (2.5) is then an inappropriate specification. The appropriate specification expresses the *logarithm* of the series as a linear function of time. This result may be seen as follows.

Without disturbances a constant growth series is given by the equation

$$Y_t = Y_0(1 + g)^t \tag{2.6}$$

where $g = (Y_t - Y_{t-1})/Y_{t-1}$ is the constant proportionate rate of growth per period. Taking logs of both sides of Eq. (2.6) gives[1]

$$\ln Y_t = \alpha + \beta t \tag{2.7}$$

where $$\alpha = \ln Y_0 \qquad \text{and} \qquad \beta = \ln(1 + g) \tag{2.8}$$

If one suspects that a series has a constant growth rate, plotting the log of the series against time provides a quick check. If the scatter is approximately linear, Eq. (2.7) can be fitted by least squares, regressing the log of Y against time. The resultant slope coefficient then provides an estimate $\hat{g}$ of the growth rate, namely,

$$b = \ln(1 + \hat{g}) \qquad \text{giving} \qquad \hat{g} = e^b - 1$$

The β coefficient of Eq. (2.7) represents the *continuous* rate of change $\partial \ln Y_t/\partial t$, whereas g represents the *discrete* rate. Formulating a constant growth series in continuous time gives

$$Y_t = Y_0 e^{\beta t} \qquad \text{or} \qquad \ln Y_t = \alpha + \beta t$$

Finally, note that taking first differences of Eq. (2.7) gives

$$\Delta \ln Y_t = \beta = \ln(1 + g) \approx g \tag{2.9}$$

Thus, taking first differences of logs gives the continuous growth rate, which in turn is an approximation to the discrete growth rate. This approximation is only reasonably accurate for small values of g.

2.1.2 Numerical Example

Table 2.1 gives data on bituminous coal output in the United States by decades from 1841 to 1910. Plotting the log of output against time, we find a linear relationship.

[1] We use ln to indicate logs to the natural base e.

TABLE 2.1
Bituminous coal output in the United States, 1841–1910

Decade	Average annual output (1,000 net tons), Y	$\ln Y$	t	$t(\ln Y)$
1841–1850	1,837	7.5159	−3	−22.5457
1851–1860	4,868	8.4904	−2	−16.9809
1861–1870	12,411	9.4263	−1	−9.4263
1871–1880	32,617	10.3926	0	0
1881–1890	82,770	11.3238	1	11.3238
1891–1900	148,457	11.9081	2	23.8161
1901–1910	322,958	12.6853	3	38.0558
Sum		71.7424	0	24.2408

So we will fit a constant growth curve and estimate the annual growth rate. Setting the origin for time at the center of the 1870s and taking a unit of time to be 10 years, we obtain the t series shown in the table. From the data in the table

$$a = \frac{\sum \ln Y}{n} = \frac{71.7424}{7} = 10.2489$$

$$b = \frac{\sum t \ln Y}{\sum t^2} = \frac{24.2408}{28} = 0.8657$$

The r^2 for this regression is 0.9945, confirming the linearity of the scatter. The estimated growth rate per decade is obtained from

$$\hat{g} = e^b - 1 = 1.3768$$

Thus the constant growth rate is almost 140 percent per decade. The annual growth rate (agr) is then found from

$$(1 + \text{agr})^{10} = 2.3768$$

which gives agr = 0.0904, or just over 9 percent per annum. The equivalent continuous rate is 0.0866.

The time variable may be treated as a fixed regressor, and so the inference procedures of Chapter 1 are applicable to equations like (2.5) and (2.7).[2]

2.2 TRANSFORMATIONS OF VARIABLES

The log transformation of the dependent variable in growth studies leads naturally to the consideration of other transformations. These transformations may be of the

[2] For a very useful discussion of the use of time as a regressor, see Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics*, Oxford University Press, 1993, pp. 115–118.

dependent variable, the regressor variable, or both. Their main purpose is to achieve a **linearizing transformation** so that the simple techniques of Chapter 1 may be applied to suitably transformed variables and thus obviate the need to fit more complicated relations.

2.2.1 Log-Log Transformations

The growth equation has employed a transformation of the dependent variable. Many important econometric applications involve the logs of both variables. The relevant functional specification is

$$Y = AX^{\beta} \qquad \text{or} \qquad \ln Y = \alpha + \beta \ln X \tag{2.10}$$

where $\alpha = \ln A$. The **elasticity** of Y with respect to X is defined as

$$\text{Elasticity} = \frac{dY}{dX}\frac{X}{Y}$$

It measures the percent change in Y for a 1 percent change in X. Applying the elasticity formula to the first expression in Eq. (2.10) shows that the elasticity of this function is simply β, and the second expression in Eq. (2.10) shows that the slope of the log-log specification is the elasticity. Thus Eq. (2.10) specifies a **constant elasticity function.** Such specifications frequently appear in applied work, possibly because of their simplicity and ease of interpretation, since slopes in log-log regressions are direct estimates of (constant) elasticities. Figure 2.1 shows some typical shapes in the Y,X plane for various βs.

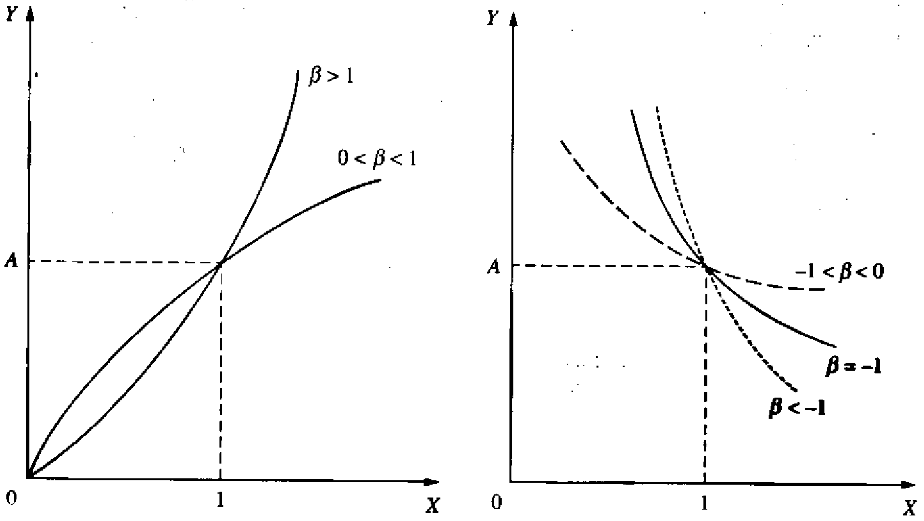


FIGURE 2.1
$Y = AX^{\beta}$.

2.2.2 Semilog Transformations

One example has already been given in the constant growth equation. The general formulation is[3]

$$\ln Y = \alpha + \beta X + u \tag{2.11}$$

This specification is widely used in human capital models, where Y denotes earnings and X years of schooling or work experience.[4] It follows from Eq. (2.11) that

$$\frac{1}{Y}\frac{dY}{dX} = \beta$$

Thus the slope in the semilog regression estimates the *proportionate* change in Y *per unit* change in X. An illustration for positive β is shown in Fig. 2.2*a*. Reversing the

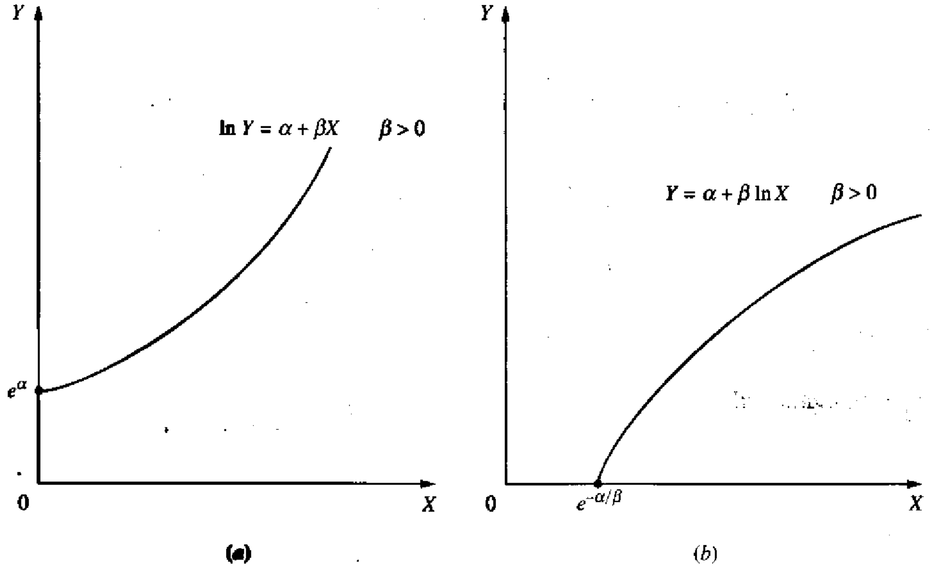


FIGURE 2.2
Semilog model.

[3] The astute reader will have noticed that in discussing various transformations we are playing fast and loose with the disturbance term, inserting it in some equations and not in others in order to simplify the transformations. The only justifications for such a (common) practice are ignorance and convenience. The late Sir Julian Huxley (distinguished biologist and brother of the novelist Aldous Huxley) once described God as a "personified symbol for man's residual ignorance." The disturbance term plays a similar role in econometrics, being a stochastic symbol for the econometrician's residual ignorance. And, just as one often does with God, one ascribes to the inscrutable and unknowable the properties most convenient for the purpose at hand.

[4] This specification is derived from theoretical considerations in J. Mincer, *School, Experience, and Earnings*, Columbia University Press, New York, 1974.

axes gives

$$Y = \alpha + \beta \ln X \tag{2.12}$$

An illustration for positive β appears in Fig. 2.2*b*. In a cross-section study of household budgets such a curve might represent the relation between a class of expenditure Y and income X. A certain threshold level of income $(e^{-\alpha/\beta})$ is needed before anything is spent on this commodity. Expenditure then increases monotonically with income, but at a diminishing rate. The marginal propensity (β/X) to consume this good declines with increasing income, and the elasticity (β/Y) also declines as income increases.

2.2.3 Reciprocal Transformations

Reciprocal transformations are useful in modeling situations where there are **asymptotes** for one or both variables. Consider

$$(Y - \alpha_1)(X - \alpha_2) = \alpha_3 \tag{2.13}$$

This describes a **rectangular hyperbola** with asymptotes at $Y = \alpha_1$ and $X = \alpha_2$. Figure 2.3 shows some typical shapes for positive and negative α_3. Equation (2.13) may be rewritten as

$$Y = \alpha_1 + \frac{\alpha_3}{X - \alpha_2} \tag{2.14}$$

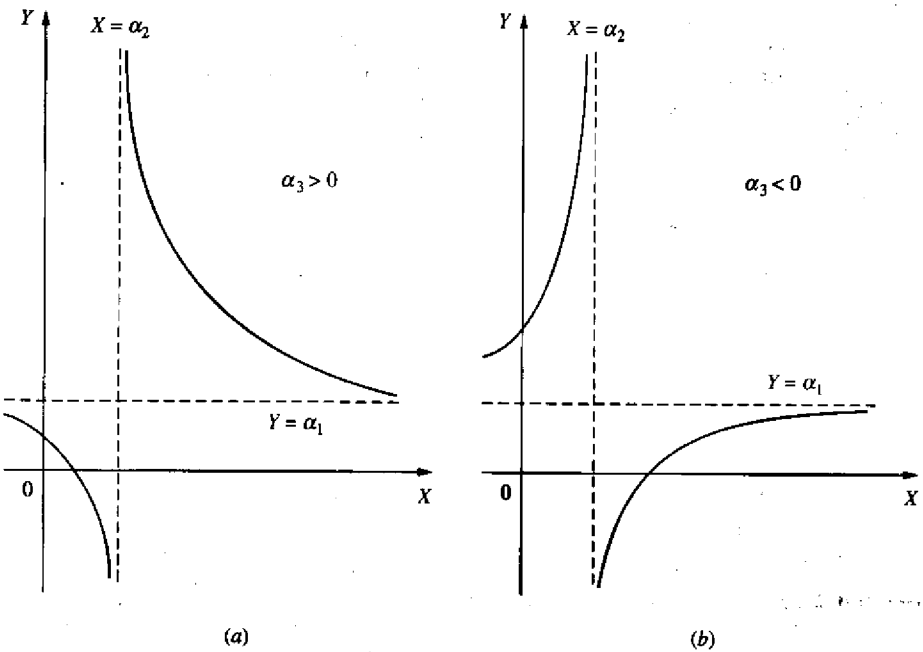


FIGURE 2.3
Rectangular hyperbola.

The result of adding an error term to Eq. (2.14) and attempting to minimize the residual sum of squares gives equations that are nonlinear in the α's. In this case there is no possible linearizing transformation that will take us back to the simple routines of Chapter 1.[5] However, there are two special cases of Eq. (2.14) where linearizing transformations are available. Setting α_2 to zero gives

$$Y = \alpha + \beta\left(\frac{1}{X}\right) \tag{2.15}$$

where $\alpha = \alpha_1$ and $\beta = \alpha_3$. Alternatively, setting α_1 to zero gives

$$\left(\frac{1}{Y}\right) = \alpha + \beta X \tag{2.16}$$

where $\alpha = -\alpha_2/\alpha_3$ and $\beta = 1/\alpha_3$. Illustrations are shown in Figs. 2.4 and 2.5.

Figure 2.4*a* has been fitted frequently in the study of **Phillips curves,** with Y representing the rate of wage or price change and X the unemployment rate. This specification carries the unrealistic implication that the asymptote for the unemployment rate is zero. The alternative simplification in Eq. (2.16) permits a positive minimum unemployment rate, but at the cost of imposing a zero minimum for wage change.

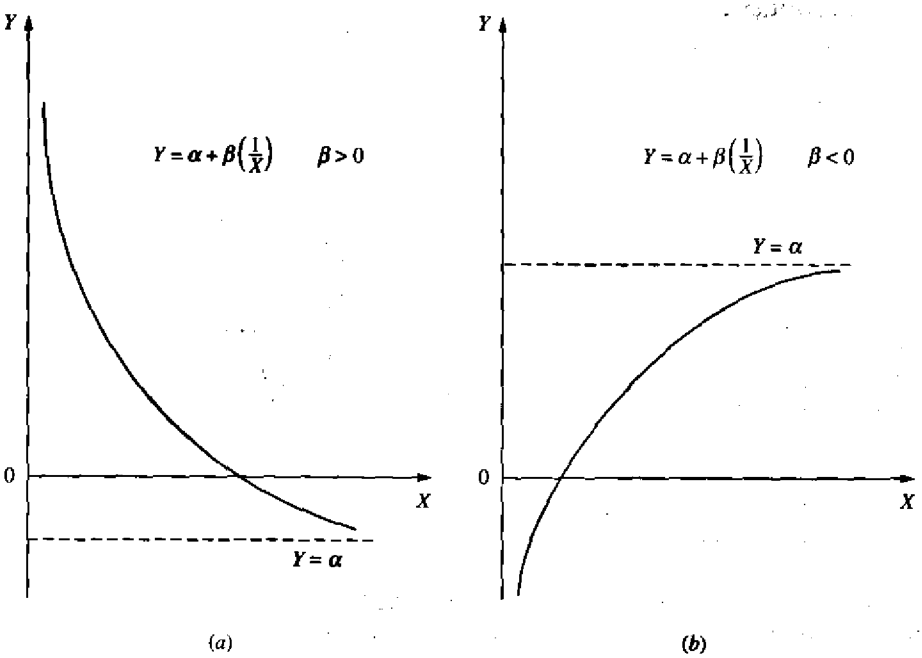


FIGURE 2.4
$Y = \alpha + \beta\left(\frac{1}{X}\right)$

[5] As will be seen later, the equation may be fitted directly by nonlinear least squares.

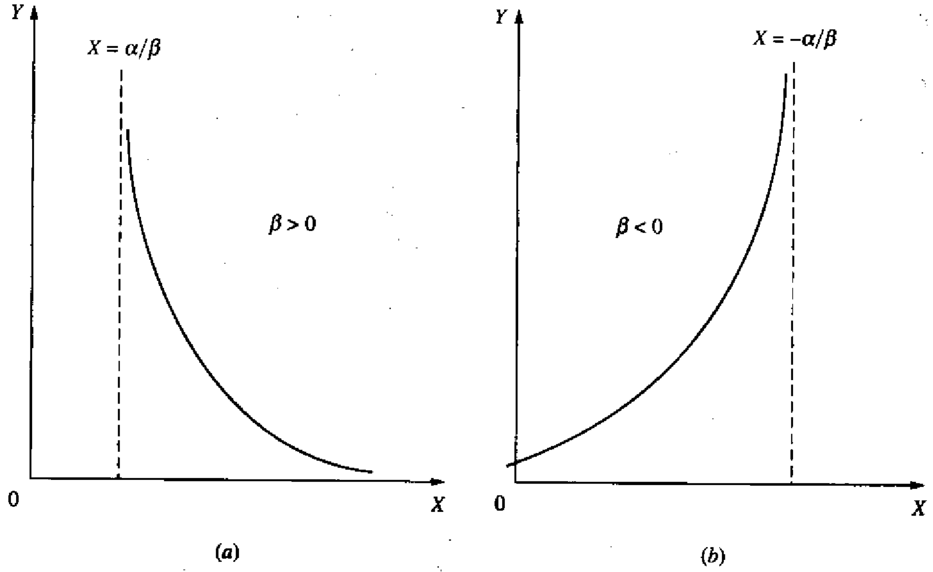


FIGURE 2.5
$\frac{1}{Y} = \alpha + \beta X$

The more general specification illustrated in Fig. 2.3*a* removes both restrictions and allows the possibility of a positive minimum rate of unemployment and a negative wage change. Figure 2.4*b* might represent a cross-section expenditure function. A certain threshold level of income is required before there is any expenditure on, say, restaurant meals, but such expenditure tends toward some upper limit, where the billionaire spends only infinitesimally more than the millionaire.

2.3 AN EMPIRICAL EXAMPLE OF A NONLINEAR RELATION: U.S. INFLATION AND UNEMPLOYMENT

The publication of the "Phillips curve" article in 1958 launched a new growth industry, whose practitioners searched for (and found) Phillips curves in various countries.[6] In the original article Phillips plotted the annual percentage wage change in the United Kingdom against the unemployment rate for the period 1861 to 1913. The scatter revealed a negative nonlinear relation, which Phillips summarized in the form of a curved line. Most remarkably, data for two subsequent periods, 1913–1948, and 1948–1957, lay close to the curve derived from the 1861–1913 data. This simple Phillips curve has not survived the passage of time and has been subject to both statistical and theoretical attack and reformulation. Thus a simple two-variable analysis

[6] A. W. Phillips, "The Relation between Unemployment and the Rate of Change of Money Wages in the United Kingdom, 1861–1957," *Economica*, New Series **25**, 1958, 283–299.

of wage (or price) inflation and unemployment can no longer be regarded as a serious piece of econometrics. However, in this chapter we are still restricted to two-variable relations, and the following example should only be taken as an illustration of the statistical steps in fitting nonlinear relations in two variables.

The data used are annual data for the United States from 1957 to 1970. The inflation variable (INF) is the annual percentage change in the Consumer Price Index (CPI). The unemployment variable (UNR) is the unemployment rate for civilian workers 16 years and over. Inflation ranges from a low of 0.69 percent in 1959 to a high of 5.72 percent in 1970, with a mean of 2.58 percent. The unemployment rate was 4.3 percent in 1957, rising to a peak of 6.7 percent in 1961 and falling steadily

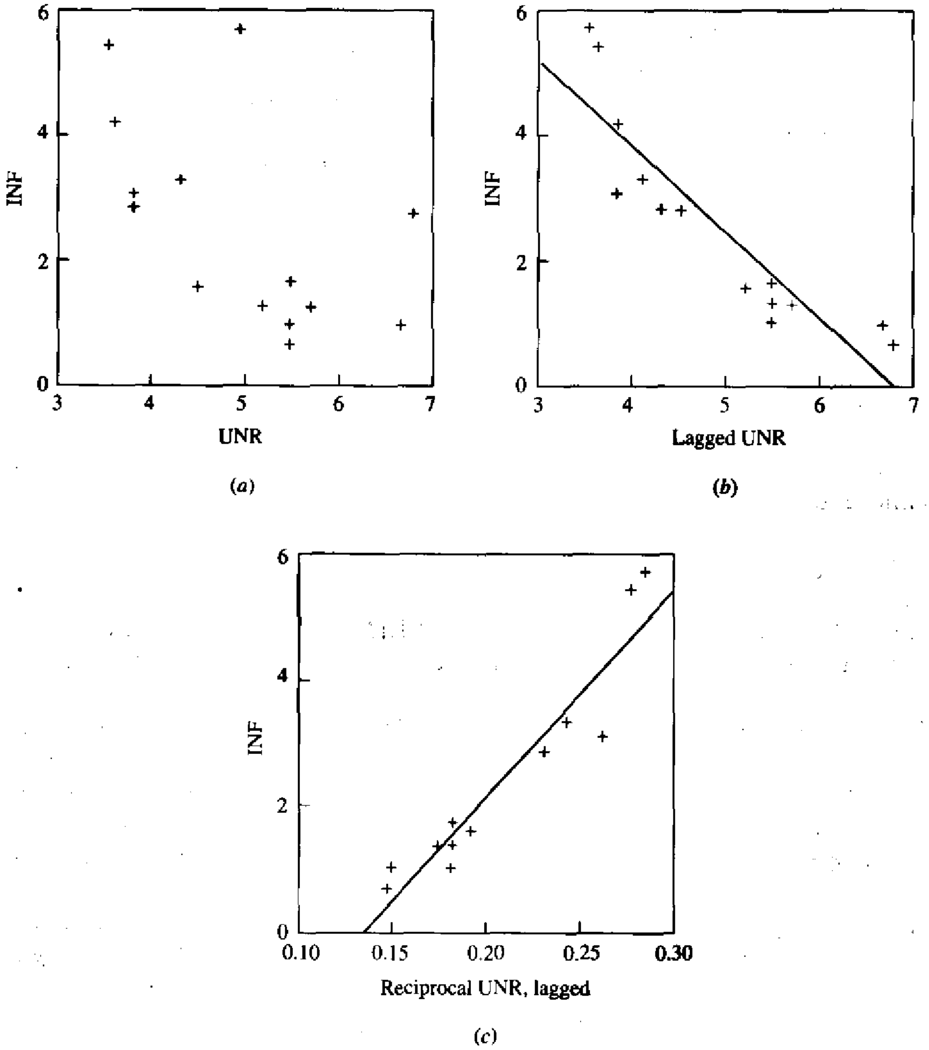


FIGURE 2.6
U.S. inflation and unemployment, 1957–1970.

TABLE 2.2

Various inflation/unemployment regressions, 1957–1970*

Explanatory variable	Constant	Slope	r^2	S.E.R.
UNR	6.92 (3.82)	−0.8764 (−2.45)	0.33	1.40
UNR(−1)	9.13 (9.80)	−1.3386 (−7.19)	0.81	0.74
1/UNR(−1)	−4.48 (−6.51)	32.9772 (10.50)	0.90	0.54

*The t statistics are in parentheses. S.E.R. is the standard error of the regression.

through the rest of the 1960s. Figure 2.6*a* shows the scatter of inflation against the current unemployment rate. The slope is negative but the scatter is dispersed. In Fig. 2.6*b* inflation is plotted against the previous year's unemployment rate. A lagged response is not unreasonable since time is required for unemployment to affect wages and further time for wage changes to filter through to the prices of final goods. The scatter is now much tighter and there is an indication of nonlinearity. The same figure shows the fit of a *linear* regression of inflation on lagged unemployment. The linear specification is clearly an inadequate representation. Of the 14 residuals from the regression, 5 are positive and 9 are negative. The 5 positive residuals occur at the lowest and highest values of the explanatory variable. Inspection of the residuals can thus indicate possible misspecification. Runs of positive or negative residuals suggest misspecification.

Figure 2.6*c* shows the result of fitting the reciprocal relation

$$\text{INF} = \alpha + \gamma\left[\frac{1}{\text{UNR}(-1)}\right] + u \tag{2.17}$$

The residuals are somewhat smaller than in Fig. 2.6*b* and the scatter is more nearly linear, but not totally so. Table 2.2 summarizes the main results from the regressions associated with Fig. 2.6. We notice the substantial jump in r^2 on changing the explanatory variable from current to lagged unemployment, and a still further increase from 0.81 to 0.90 on using the reciprocal transformation.

Finally, we note the result of fitting the nonlinear relation

$$\text{INF} = \alpha_1 + \frac{\alpha_3}{\text{UNR}(-1) - \alpha_2} + u \tag{2.18}$$

This is fitted by nonlinear least squares, which is an iterative estimation process, commencing with some arbitrary values for the unknown parameters, calculating the residual sum of squares, then searching for changes in the parameters to reduce the RSS, and continuing in this way until successive changes in the estimated parameters and in the associated RSS are negligibly small. Standard errors and other test statistics can be produced at the final stage, just as in linear least squares; but, as will be explained shortly, they now have an asymptotic justification rather than exact, finite sample properties. As noted earlier the linearizing transformations obtained by setting α_1 or α_2 to zero impose theoretically inappropriate constraints on the shape of the relationship. Setting α_2 to zero, as in the third regression in Table 2.2, gives a lower asymptote of zero for the unemployment rate, which is implausibly small. On the other hand, setting α_1 to zero gives a lower asymptote of zero for the inflation rate, which implies that the price level could not fall no matter how great the level of unemployment. which again is an implausible restriction. Using nonlinear least squares to estimate the relation without these restrictions gives

$$\widehat{\text{INF}} = -0.32 + \frac{4.8882}{\text{UNR}(-1) - 2.6917}$$

with $r^2 = 0.95$ and S.E.R $= 0.40$. This expression provides the best fit and lowest standard error of all the regressions. The intercept term, which is the estimated asymptote for the inflation rate, is slightly negative but not significantly different from zero. The estimated asymptote for the unemployment rate is 2.69 percent. The contrast between the unemployment asymptotes in fitting Eqs. (2.17) and (2.18) is a striking reminder of the fact that each specification imposes a particular *shape* on the estimated relation. Equation (2.18) implies dramatically increasing inflation rates for unemployment rates just below 3 percent, whereas Eq. (2.17) requires unemployment rates below 1 percent to give similar inflation numbers. The fits to the sample data are not very different, but extrapolations outside the range of the sample data give dramatically different pictures.

2.4 LAGGED DEPENDENT VARIABLE AS REGRESSOR

When variables display trends as in Section 2.1, successive values tend to be fairly close together. Another way of modeling such behavior is by means of an **autoregression.** The simplest autoregressive scheme is

$$Y_t = \alpha + \beta Y_{t-1} + u_t \tag{2.19}$$

This is called a first-order. autoregressive scheme and is frequently denoted by the notation AR(1). The order indicates the (maximum) lag in the equation. If, for example, Y were measured quarterly, $Y_t = \alpha + \beta Y_{t-4} + u_t$ embodies the assumption that the current Y is related to the value in the same quarter of the previous year, and this is a special case of an AR(4) scheme.

The LS equations for fitting Eq. (2.19) are

$$\begin{aligned} \sum Y_t &= na + b\sum Y_{t-1} \\ \sum Y_t Y_{t-1} &= a\sum Y_{t-1} + b\sum Y_{t-1}^2 \end{aligned} \tag{2.20}$$

The range of summation in Eq. (2.20) has not been indicated explicitly. If t ranges from 1 to n, the implementation of Eq. (2.20) requires a value for Y_0. If this is not available, then Y_1 is the starting value and the effective sample size is $n - 1$. LS estimates of the parameters of Eq. (2.19) can thus be computed; but the properties of LS estimators and the associated inference procedures derived in Chapter 1 are not strictly applicable here, even though we continue to make the same assumptions is how a random variable such as $\bar{x}_n$ and its pdf behave as $n \to \infty$. There are two main aspects of this behavior, the first relating to *convergence in probability* and the second to *convergence in distribution.*

2.4.2 Convergence in Probability

The x's are iid(μ, σ^2) by assumption. It follows directly that

$$E(\bar{x}_n) = \mu \qquad \text{and} \qquad \text{var}(\bar{x}_n) = \frac{\sigma^2}{n}$$

Thus $\bar{x}_n$ is an unbiased estimator for any sample size, and the variance tends to zero as n increases indefinitely. It is then intuitively clear that the distribution of $\bar{x}_n$, whatever its precise form, becomes more and more concentrated in the neighborhood of μ as n increases. Formally, if one defines a neighborhood around μ as $\mu \pm \epsilon$, the expression

$$\Pr\{\mu - \epsilon < \bar{x}_n < \mu + \epsilon\} = \Pr\{|\bar{x}_n - \mu| < \epsilon\}$$

indicates the probability that $\bar{x}_n$ lies in the specified interval. The interval may be made arbitrarily small by a suitable choice of ϵ. Since var$(\bar{x}_n)$ declines monotonically with increasing n, there exists a number n^* and a δ $(0 < \delta < 1)$ such that for all $n > n^*$

$$\Pr\{|\bar{x}_n - \mu| < \epsilon\} > 1 - \delta \tag{2.23}$$

The random variable $\bar{x}_n$ is then said to *converge in probability* to the constant μ. An equivalent statement is

$$\lim_{n \to \infty} \Pr\{|\bar{x}_n - \mu| < \epsilon\} = 1 \tag{2.24}$$

In words, the probability of $\bar{x}_n$ lying in an arbitrarily small interval about μ can be made as close to unity as we desire by letting n become sufficiently large. A shorthand way of writing Eq. (2.24) is

$$\text{plim}\ \bar{x}_n = \mu \tag{2.25}$$

where plim is an abbreviation of probability limit. The sample mean is then said to be a **consistent estimator** of μ. The process is called **convergence in probability.**

In this example the estimator is unbiased for *all* sample sizes. Suppose that we have another estimator m_n of μ such that

$$E(m_n) = \mu + \frac{c}{n}$$

where c is some constant. The estimator is biased in finite samples, but

$$\lim_{n \to \infty} E(m_n) = \mu$$

Provided var(m_n) goes to zero with increasing n, m_n is also a consistent estimator of μ. This case is an example of **convergence in mean square,** which occurs when the limit of the expected value of the estimator is the parameter of interest, and the limit of the variance of the estimator is zero. Convergence in mean square is a sufficient condition for consistency and often provides a useful way of establishing a probability limit.

An extremely useful feature of probability limits is the ease with which the probability limits of functions of random variables may be obtained. For example, if we assume that a_n and b_n possess probability limits, then

$$\text{plim}\,(a_n b_n) = \text{plim}\,a_n \cdot \text{plim}\,b_n$$

and

$$\text{plim}\left(\frac{a_n}{b_n}\right) = \frac{\text{plim}\,a_n}{\text{plim}\,b_n}$$

Such relations do not hold for expectations unless a_n and b_n are stochastically independent, but no such condition is required for operating with probability limits.

2.4.3 Convergence in Distribution

The next crucial question is how the pdf of $\bar{x}_n$ behaves with increasing n. The form of the distribution is unknown, since the mean is a linear combination of x's whose distribution is assumed to be unknown. However, since the variance goes to zero in the limit, the distribution collapses on μ. The distribution is then said to be **degenerate.** One seeks then an alternative statistic, some function of $\bar{x}_n$, whose distribution will not degenerate. A suitable alternative statistic is $\sqrt{n}(\bar{x}_n - \mu)/\sigma$, which has zero mean and unit variance. The basic **Central Limit Theorem** states[8]

$$\lim_{n\to\infty} \Pr\left\{\frac{\sqrt{n}(\bar{x}_n - \mu)}{\sigma} \le y\right\} = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz \tag{2.26}$$

The left-hand side of this expression is the limiting value of the probability that the statistic $\sqrt{n}(\bar{x}_n - \mu)/\sigma$ is less than or equal to some value y. The right-hand side is the appropriate area under the **standard normal distribution**, $N(0, 1)$. This is a remarkable and powerful result. Whatever the form of $f(x)$, the limiting distribution of the relevant statistic is standard normal. The process is labeled **convergence in distribution,** and an alternative way of expressing Eq. (2.26) is

$$\sqrt{n}\bar{x}_n \xrightarrow{d} N(\sqrt{n}\mu, \sigma^2) \tag{2.27}$$

to be read, "$\sqrt{n}\bar{x}_n$ tends in distribution to a normal variable with mean $\sqrt{n}\mu$ and variance σ^2." In practice the objective is to use $\bar{x}_n$ to make inferences about μ. This is done by taking the limiting normal form as an approximation for the unknown distribution of $\bar{x}_n$. The relevant statement is

$$\bar{x}_n \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \tag{2.28}$$

[8] See, for example, S. S. Wilks, *Mathematical Statistics*, Wiley, 1962, p. 256.

to be read, "$\bar{x}_n$ is asymptotically normally distributed with mean μ and variance σ^2/n." The unknown σ^2 can be replaced by the sample variance, which will be a consistent estimate, and Eq. (2.28) used for inferences about μ. The closeness of the approximation to the unknown pdf obviously depends on the extent of the departure from normality of the original distribution and on the sample size. However, a striking illustration of the tendency of linear combinations of nonnormal variables to move toward normality is shown in Fig. 2.7. It comes from *Sampling Techniques* by William G. Cochran (Wiley, 1953, pp. 22–23). The initial distribution shows the population of 196 large U.S. cities in 1920. The distribution is highly nonnormal, having in fact a reverse J shape, with very many smaller cities and few very large ones. Two hundred simple random samples were drawn, each of 49 cities, and the total population calculated in each of the 200 cases. The distribution of the sample total population (and, likewise, the sample mean population) was *unimodal* and very much closer in appearance to a normal curve than the original distribution.

2.4.4 The Autoregressive Equation

The autoregressive Eq. (2.19) may be estimated by the LS formulae in Eq. (2.20). If we denote the estimated coefficients by a and b, the results of Mann and Wald establish that $\sqrt{n}(a-\alpha)$ and $\sqrt{n}(b-\beta)$ have a *bivariate normal limiting distribution*

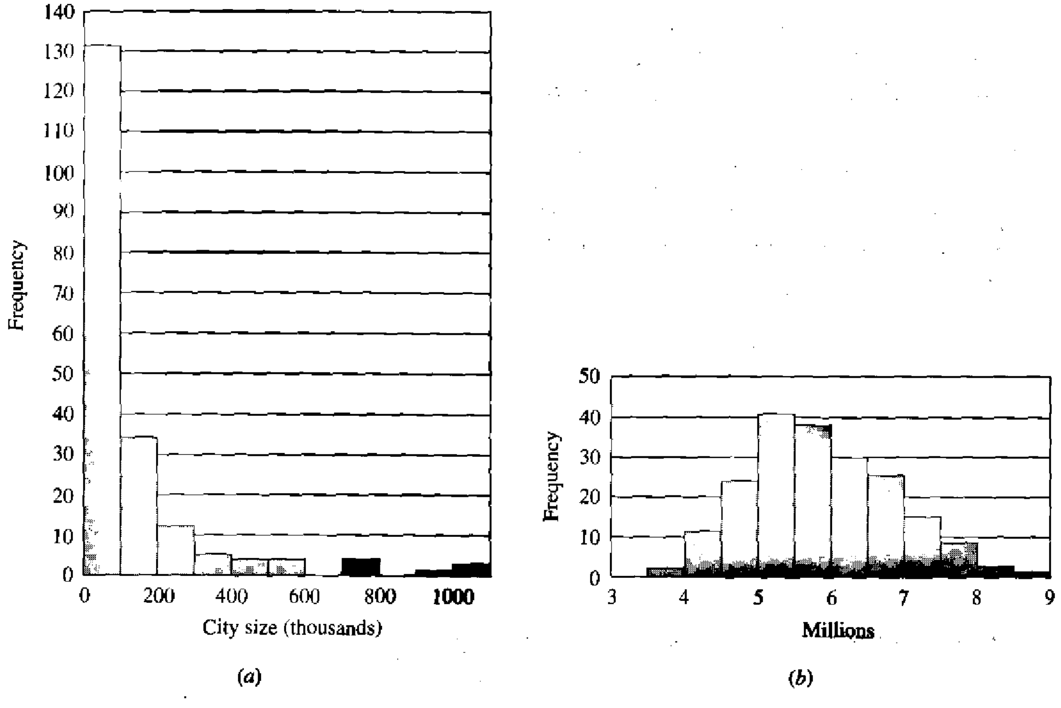


FIGURE 2.7
Influence of original distribution and sample size on approach to normality. (*a*) Frequency distribution of sizes of 196 U.S. cities in 1920; (*b*) frequency distribution of totals of 200 simple random samples with $n = 49$. (Reprinted by permission of John Wiley & Sons, Inc.)

with zero means and finite variances and covariance.[9] Thus the least-squares estimates are consistent for α and β. Moreover the limiting variances and covariance may be consistently estimated by the LS formulae of Chapter 1. Consequently we may apply the LS techniques of Chapter 1 to the autoregressive model; and they now have an asymptotic, or large-sample, justification rather than exact, finite-sample validity.

The Mann–Wald result depends on two crucial assumptions. The first assumption is that the disturbances in the relation are independently and identically distributed with zero mean and finite variance, as in Eq. (2.2). Note, however, that there is no assumption of normality for the disturbances. The second assumption is that the $\{Y_t\}$ series is **stationary.** This raises new considerations, which are explored in the next section.

2.5 STATIONARY AND NONSTATIONARY SERIES

We return to the relationship specified in Eq. (2.19),

$$Y_t = \alpha + \beta Y_{t-1} + u_t$$

and make the assumptions about the u variable stated in Eq. (2.2). These assumptions define a **white noise** series. The crucial question is, how does the Y series behave over time? Equation (2.21) shows Y_t as a function of α, β, Y_0, and the current and previous disturbances. Assuming that the process started a very long time ago, we rewrite Eq. (2.21) as

$$Y_t = \alpha(1 + \beta + \beta^2 + \cdots) + (u_t + \beta u_{t-1} + \beta^2 u_{t-2} + \cdots) \tag{2.29}$$

The stochastic properties of the Y series are determined by the stochastic properties of the u series. Taking expectations of both sides of Eq. (2.29) gives

$$E(Y_t) = \alpha(1 + \beta + \beta^2 + \cdots)$$

This expectation only exists if the infinite geometric series on the right-hand side has a limit. The necessary and sufficient condition is

$$|\beta| < 1 \tag{2.30}$$

The expectation is then

$$E(Y_t) = \mu = \frac{\alpha}{1 - \beta} \tag{2.31}$$

and so the Y series has a constant unconditional mean μ at all points. To determine the variance we can now write

[9] H. B. Mann and A. Wald, "On the Statistical Treatment of Linear Stochastic Difference Equations," *Econometrica*, **11**, 1943, pp. 173–220. The article is long and very technical, but the results are of great practical importance.

$$(Y_t - \mu) = u_t + \beta u_{t-1} + \beta^2 u_{t-2} + \cdots \tag{2.32}$$

Squaring both sides and taking expectations, we find

$$\begin{aligned}\operatorname{var}(Y_t) &= E[(Y_t - \mu)^2] \\ &= E[u_t^2 + \beta^2 u_{t-1}^2 + \beta^4 u_{t-2}^2 + \cdots + 2\beta u_t u_{t-1} + 2\beta^2 u_t u_{t-2} + \cdots]\end{aligned}$$

The assumption in Eq. (2.2) then yields

$$\operatorname{var}(Y) = \sigma_y^2 = \frac{\sigma^2}{1 - \beta^2} \tag{2.33}$$

Thus the Y series has a constant unconditional variance, independent of time.

A new concept is that of **autocovariance,** which is the covariance of Y with a lagged value of itself. The first-lag autocovariance is defined as

$$\begin{aligned}\gamma_1 &= E[(Y_t - \mu)(Y_{t-1} - \mu)] \\ &= \beta\sigma_y^2 \qquad\qquad \text{using Eq. (2.32)}\end{aligned}$$

In a similar fashion the second-lag autocovariance is

$$\begin{aligned}\gamma_2 &= E[(Y_t - \mu)(Y_{t-2} - \mu)] \\ &= \beta^2\sigma_y^2\end{aligned}$$

and, in general,

$$\gamma_s = \beta^s\sigma_y^2 \qquad s = 0, 1, 2, \ldots \tag{2.34}$$

The autocovariances thus depend only on the lag length and are independent of t. Clearly $\gamma_0 (= \sigma_y^2)$ is another symbol for the variance. Dividing the covariances by the variance gives the set of **autocorrelation coefficients,** also known as **serial correlation coefficients,** which we will designate by

$$\rho_s = \gamma_s/\gamma_0 \qquad s = 0, 1, 2, \ldots \tag{2.35}$$

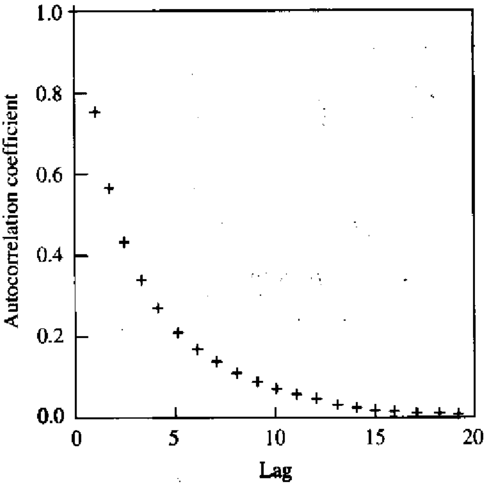


FIGURE 2.8
Correlogram of an AR(1) series (parameter = 0.75).

Plotting the autocorrelation coefficients against the lag lengths gives **the correlogram** of the series. For the first-order AR(1) scheme the autocorrelations decline exponentially from one toward zero, as illustrated in Fig. 2.8.

To summarize, when $|\beta| < 1$ the mean, variance, and covariances of the Y series are constants, independent of time. The Y series is then said to be **weakly** or **covariance stationary**. In particular, it satisfies the stationarity condition required for the asymptotic results of Mann and Wald to hold true.

2.5.1 Unit Root

When $\beta = 1$ the AR(1) process is said to have a unit root. The equation becomes

$$Y_t = \alpha + Y_{t-1} + u_t \tag{2.36}$$

which is called a **random walk with drift.**[10] From Eq. (2.21) the *conditional expectation* is

$$E(Y_t \mid Y_0) = \alpha t + Y_0$$

which increases or decreases without limit as t increases. The *conditional variance* is

$$\begin{aligned} \text{var}(Y_t \mid Y_0) &= E[(Y_t - E(Y_t \mid Y_0))^2] \\ &= E[(u_t + u_{t-1} + \cdots + u_1)^2] \\ &= t\sigma^2 \end{aligned}$$

which increases without limit. Referring to Eq. (2.29) we clearly see that, in the unit root case, the unconditional mean and variance of Y do not exist. The Y series is then said to be **nonstationary**, and the asymptotic results previously described no longer hold. The treatment of nonstationary series will be taken up in Chapters 7 and 8. When $|\beta| > 1$ the Y series will exhibit explosive behavior, as will be illustrated in the following example.

2.5.2 Numerical Illustration

Three series have been generated as follows:

A:	$A_t = 0.05 + 0.95A_{t-1} + u_t$	Autoregressive (stationary) series
R:	$R_t = 0.05 + R_{t-1} + u_t$	Random walk with drift
E:	$E_t = 0.05 + 1.05E_{t-1} + u_t$	Explosive series

All series are started at zero and 500 terms generated by random drawings from a standard normal distribution u_t. Figure 2.9 contrasts the autoregressive and random

[10] When $\alpha = 0$, Eq. (2.36) reduces to a simple random walk, $Y_t = Y_{t-1} + u_t$.

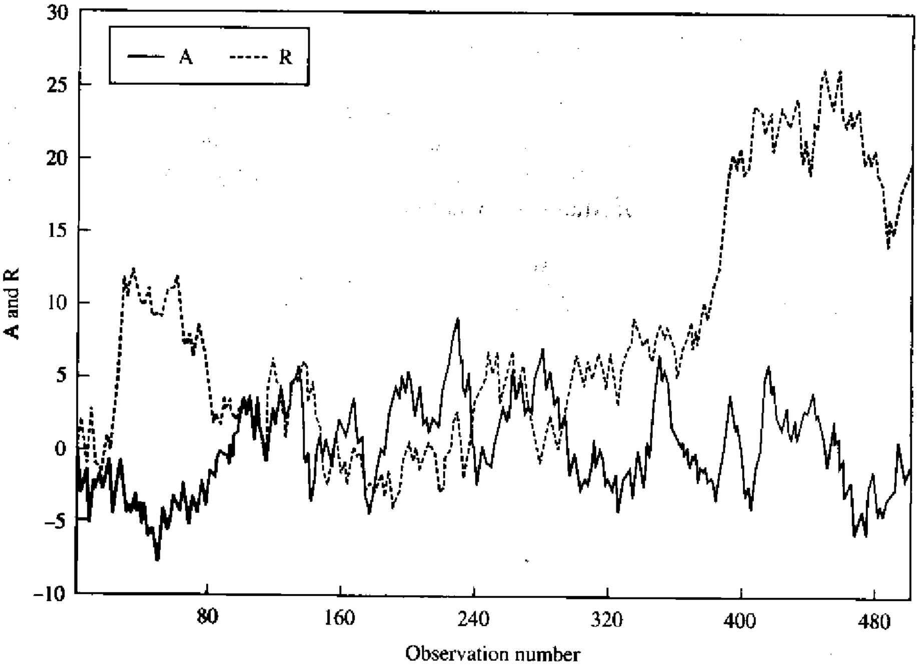


FIGURE 2.9
A stationary AR(1) series and a random walk.

walk series. The theoretical mean of the A series is 1, and its theoretical standard deviation is 3.2. The series displays a fairly steady mean level, and all observations are contained in the range of ±10. The random walk series does not look much different from the stable A series through most of the early observations, but does drift markedly away in the later observations. Figure 2.10 shows the A and R series again, but this time the explosive E series is plotted on the same graph. Notice the dramatic difference in the vertical scale compared with Fig. 2.9. On the scale required to accommodate the E series, the other two appear as a single straight line. The β parameter for E only exceeds unity by 0.05, the same amount as the β parameter for A falls short of unity. *Thus, the unit root case is a watershed.* The A and R series have more the typical look of an economic time series than does the E series. As Fig. 2.9 suggests it may in practice be very difficult to distinguish between stationary series, such as A, and nonstationary series like R.

It might appear simple to test for a unit root by fitting Eq. (2.19), computing the test statistic $(b - 1)/\text{s.e.}(b)$, and referring this to the conventional critical values from the t distribution. Unfortunately this procedure is not valid since, under the null hypothesis, the Y series does not satisfy, even asymptotically, the conditions assumed in deriving the test. The distribution of the test statistic is nonstandard, and critical values can only be obtained by Monte Carlo simulations. Unit root tests will be dealt with in Chapter 7. A more informal test for stationarity is based on inspection of the autocorrelation coefficients. As shown earlier, the correlogram of a stationary AR series should decline exponentially. This will not be true for a nonstationary series. Table 2.3 shows selected autocorrelation coefficients for the A and R

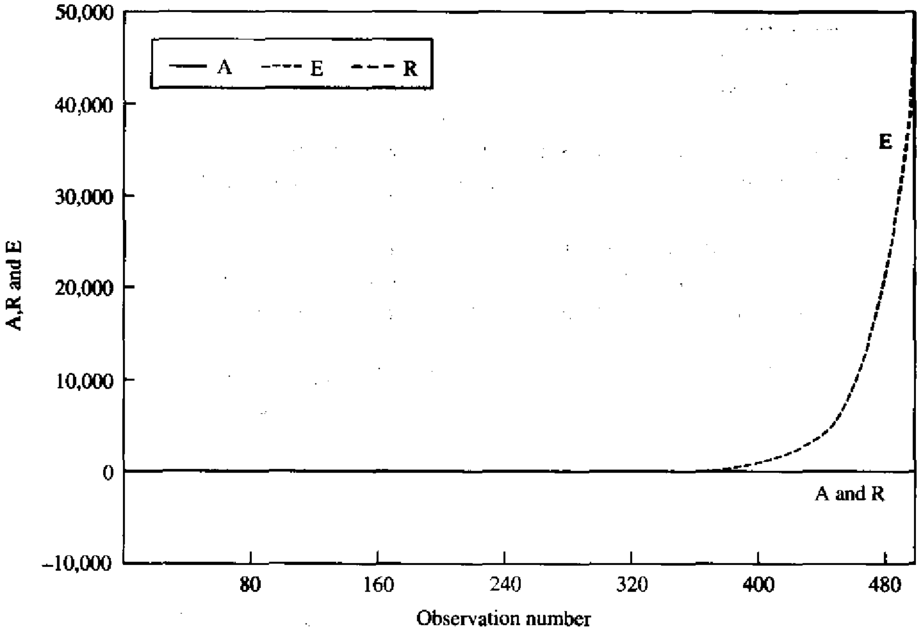


FIGURE 2.10
An explosive series.

TABLE 2.3
Autocorrelations for the A and R series

Lag	A series	R series
1	0.936	0.992
5	0.678	0.958
9	0.395	0.929
13	0.202	0.911
18	0.108	0.882

series. There is a clear difference in the patterns of the two sets of autocorrelations, confirming the stationarity of the first series and the nonstationarity of the second. However, these coefficients have been calculated from observations 101 to 500. The differences would not be so clear in smaller samples. We will return to the issue of stationarity and the validity of standard inference procedures in more realistic *multivariate* situations in later chapters.

2.6 MAXIMUM LIKELIHOOD ESTIMATION OF THE AUTOREGRESSIVE EQUATION

2.6.1 Maximum Likelihood Estimators

The derivation of the Mann–Wald results in Section 2.4 did not require any assumption about the specific form of the pdf for the disturbance term. If such an assumption can be made, it is then possible to derive *maximum likelihood estimators* of the parameters of the autoregressive model. Maximum likelihood estimators, or MLEs, are **consistent** and **asymptotically normal,** as are the Mann–Wald estimators; but the MLEs have the additional property of **asymptotic efficiency,** as will be explained next. A more complete treatment of ML estimation will be given in Chapter 5, to which the present treatment of the simple autoregressive case will serve as an introduction.

The common approach is to assume, as in Eq. (2.4), that the disturbances are identically and independently distributed **normal** variables with zero mean and constant variance. The probability density function for u is then

$$f(u_i) = \frac{1}{\sigma\sqrt{2\pi}}e^{-u_i^2/2\sigma^2} \qquad i = 1, 2, \ldots, n$$

We will further postulate some arbitrary initial value Y_0, the precise value being irrelevant in an asymptotic analysis. Any observed set of sample values $Y_1, Y_2, \ldots, Y_n$ is then generated by some set of u values. Given Eq. (2.4), the probability of a set of u values is

$$\begin{aligned}\Pr(u_1, u_2, \ldots, u_n) &= f(u_1)f(u_2)\cdot\cdots\cdot f(u_n)\\ &= \prod_{t=1}^{n} f(u_t)\\ &= \frac{1}{(2\pi\sigma^2)^{n/2}}e^{-\sum_{t=1}^{n}(u_t^2/2\sigma^2)}\end{aligned}$$

From Eq. (2.19) the joint density of the Y values, conditional on y_0, is then[11]

$$\Pr(Y_1, Y_2, \ldots, Y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(Y_t - \alpha - \beta Y_{t-1})^2\right] \qquad (2.37)$$

This density may be interpreted in two ways. For given α, β, and σ^2 it indicates the probability of a set of sample outcomes. Alternatively it may be interpreted as a function of α, β, and σ^2, *conditional on a set of sample outcomes.* In the latter interpretation it is referred to as a *likelihood function* and written

$$\text{Likelihood function} = L(\alpha, \beta, \sigma^2; Y) \qquad (2.38)$$

with the order of the symbols in the parentheses reflecting the emphasis on the parameters being conditional on the observations. Maximizing the likelihood with respect to the three parameters gives specific values $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma}^2$, which maximize the probability of obtaining the sample values that have actually been observed. These are the **maximum likelihood estimators** (MLEs) of the parameters of Eq. (2.19). They are obtained by solving

$$\frac{\partial L}{\partial \alpha} = \frac{\partial L}{\partial \beta} = \frac{\partial L}{\partial \sigma^2} = 0$$

[11] See Appendix 2.1 on the transformation of variables in pdf's.

In most applications it is simpler to maximize the *logarithm* of the likelihood function. We will denote the log-likelihood by

$$l = \ln L$$

Since l is a monotonic transformation of L the MLEs may equally well be obtained by solving

$$\frac{\partial l}{\partial \alpha} = \frac{\partial l}{\partial \beta} = \frac{\partial l}{\partial \sigma^2} = 0$$

For Eq. (2.19) the log-likelihood (conditional on Y_0) is

$$l = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(Y_t - \alpha - \beta Y_{t-1})^2 \tag{2.39}$$

The formal derivations are given in Appendix 2.2, but it is intuitively clear that the $\hat{\alpha}, \hat{\beta}$ values that *maximize* l are those that *minimize* $\sum_{t=1}^{n}(Y_t - \alpha - \beta Y_{t-1})^2$. Thus, in this case the LS and ML estimates of α and β are identical. The normal equations for the LS estimates appear in Eq. (2.20). The ML estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{\alpha} - \hat{\beta} Y_{t-1})^2 \tag{2.40}$$

2.6.2 Properties of Maximum Likelihood Estimators

The major properties of ML estimators are *large-sample,* or *asymptotic,* ones. They hold under fairly general conditions.

1. **Consistency.** MLEs are consistent. Thus, Eqs. (2.20) and (2.40) yield consistent estimates of α, β, and σ^2.
2. **Asymptotic normality.** The estimators $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma}^2$ have asymptotically normal distributions centered at the true parameter values. The asymptotic variances are derived from the information matrix, which will be explained in Chapter 5. In the present application it may be shown that the asymptotic variance of $\hat{\beta}$ is *estimated* by

$$\text{Est. avar}(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum_{t=1}^{n} Y_{t-1}^2 - \frac{1}{n}(\sum_{t=1}^{n} Y_{t-1})^2}$$

where $\hat{\sigma}^2$ is defined in Eq. (2.40) and the abbreviation *avar* denotes asymptotic variance. If we let $\bar{Y}_{-1} = (1/n)(Y_0 + Y_1 + \cdots + Y_{n-1})$ and $y_{t-1} = Y_{t-1} - \bar{Y}_{-1}$, the estimated asymptotic variance becomes

$$\text{Est. avar}(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum_{t=1}^{n} y_{t-1}^2}$$

which is seen, by comparison with Eq. (1.40), to be the variance that would be calculated by the usual LS procedure, except for the trivial difference that $\hat{\sigma}^2$ has a divisor of n rather than $(n - 2)$.

3. **Asymptotic efficiency.** No other consistent and asymptotically normal estimator can have a smaller asymptotic variance. This property mirrors the finite sample minimum variance property of LS estimators.

To summarize, the lagged dependent variable model of Eq. (2.19) may be estimated by LS and standard errors calculated in the usual way. However, significance levels and confidence coefficients are no longer known precisely. The calculated values are only approximately correct and there is no way of assessing the degree of error in any particular application. Even this somewhat cautious conclusion still rests on two important assumptions. The first is the zero covariance assumption for all pairs of disturbances, which was used in the derivation of Eq. (2.37). The second is that $\sum y_{t-1}^2/n$ has a finite limit as n tends to infinity, which is part of the requirements for the stationarity of the Y series.

APPENDIX

APPENDIX 2.1
Change of variables in density functions

The basic idea may be simply illustrated for the univariate case. Suppose u is a random variable with density function $p(u)$. Let a new variable y be defined by the relation $y = f(u)$. The y variable must then also have a density function, which will obviously depend on both $p(u)$ and $f(u)$. Suppose that the relation between y and u is monotonically increasing, as in Fig. A-2.1. Whenever u lies in the interval Δu, y will be in the corresponding interval Δy. Thus

$$\Pr(y \text{ lies in } \Delta y) = \Pr(u \text{ lies in } \Delta u)$$

or

$$p(y')\Delta y = p(u')\Delta u$$

where u' and y' denote appropriate values of u and y in the intervals Δu and Δy, and $p(y)$ indicates the postulated density function for y. Taking limits as Δu goes to zero gives

$$p(y) = p(u)\frac{du}{dy}$$

If y were a decreasing function of u, the derivative in this last expression would be negative, thus implying an impossible negative value for the density function. Therefore, the *absolute* value of the derivative must be taken and the result reformulated to read

$$p(y) = p(u)\left|\frac{du}{dy}\right|$$

If $y = f(u)$ were not a monotonic function this last result would require amendment, but we are only concerned with monotonic transformations. The relevant relationship

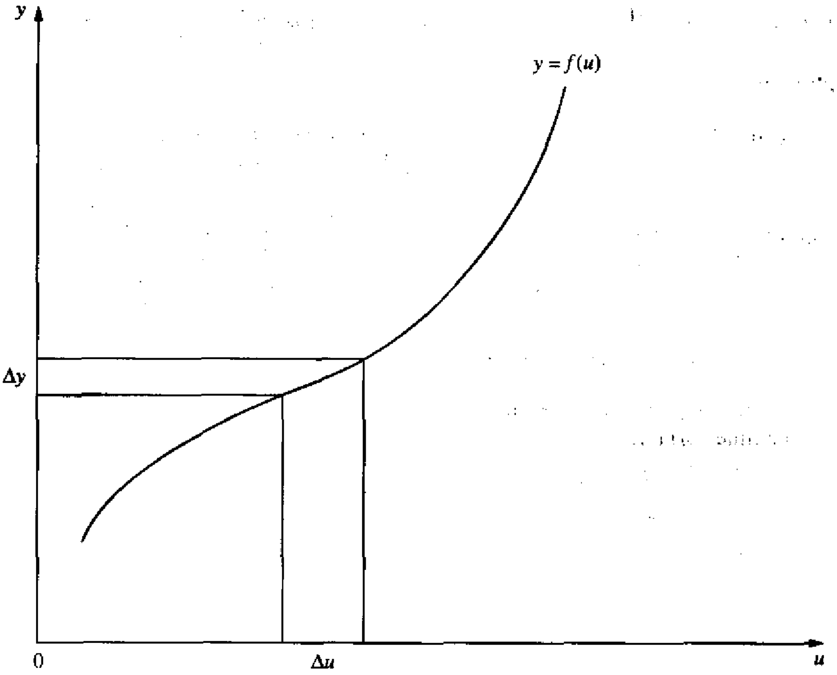


FIGURE A-2.1

in the text is

$$Y_t = \alpha + \beta Y_{t-1} + u_t \tag{2.19}$$

which gives

$$\frac{du_t}{dY_t} = 1 \qquad \text{for all } t$$

which gives the joint density shown in Eq. (2.37).

APPENDIX 2.2
Maximum likelihood estimators for the AR(1) model

The log-likelihood is given in Eq. (2.39) as

$$l = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(Y_t - \alpha - \beta Y_{t-1})^2$$

The partial derivatives with respect to the three parameters are

$$\frac{\partial l}{\partial \alpha} = \frac{1}{\sigma^2}\sum_{t=1}^{n}(Y_t - \alpha - \beta Y_{t-1})$$

$$\frac{\partial l}{\partial \beta} = \frac{1}{\sigma^2}\sum_{t=1}^{n} Y_{t-1}(Y_t - \alpha - \beta Y_{t-1})$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{n}(Y_t - \alpha - \beta Y_{t-1})^2$$

Equating the first two derivatives to zero gives the ML (LS) estimators in Eq. (2.20), and equating the third to zero gives the estimator of the variance in Eq. (2.40).

PROBLEMS

2.1. Fit a constant growth curve to the accompanying data, using two different specifications of the time variable.

Year	Marijuana crop (10,000 tons)
1985	38.1
1986	80.0
1987	170.4
1988	354.5
1989	744.4

Estimate the annual growth rate, and forecast marijuana production in 1995 from each specification.

2.2. Show that $\log Y = \alpha + \beta \log X + u$ gives the same estimate of β whether logs are taken to base 10 or to base e. Is this true of the estimate of α? Do your conclusions need modification if $\log X$ is replaced by t?

2.3. Discuss briefly the advantages and disadvantages of the relation

$$v_i = \alpha + \beta \log v_0$$

as a representation of an Engel curve, where v_i is expenditure per person on commodity i, and v_0 is income per week. Fit such a curve to the following data, and from your results estimate the income elasticity at an income of \$500 per week. Does it make any difference to your estimate of the income elasticity if the logarithms of v_0 are taken to base 10 or to base e?

	\$'00 per week							
v_i	0.8	1.2	1.5	1.8	2.2	2.3	2.6	3.1
v_0	1.7	2.7	3.6	4.6	5.7	6.7	8.1	12.0

2.4. Prove that

$$\frac{dY}{dX}\frac{X}{Y} = \frac{d(\ln Y)}{d(\ln X)} = \frac{d(\log Y)}{d(\log X)}$$

Note that this relationship holds generally and not just for constant elasticity functions.

2.5. A response rate Y to a stimulus X is modeled by the function

$$\frac{100}{100 - Y} = \alpha + \frac{\beta}{X}$$

where Y is measured in percentage terms. Outline the properties of this function and sketch its graph. Fit the function to the accompanying data.

X	3	7	12	17	25	35	45	55	70	120
Y	86	79	76	69	65	62	52	51	51	48

2.6. Discuss the properties of the following functions and sketch their graphs:

$$y = \frac{x}{\alpha x - \beta}$$

$$y = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

Find the transformations that will linearize each function.

2.7. The firmness of cheese depends on the time allowed for a certain process in its manufacture. In an experiment 18 cheeses were taken, and at each of several times, firmness was determined on samples from three of the cheeses. The results were as follows.

Time (hours)	**Firmness**		
0.5	102	105	115
1	110	120	115
1.5	126	128	119
2	132	143	139
3	160	149	147
4	164	166	172

Estimate the parameters of a linear regression of firmness on time. Compute r^2 and the standard errors of the estimates. Calculate the conditional means at each value of time and regress these conditional means on time. Compare the coefficients, standard errors, and r^2 with the previous regression. Can you think of conditions where the results you have obtained in this example would not hold?

2.8. A theorist postulates that the following functional form will provide a good fit to a set of data on Y and X:

$$Y = a + b\left(\frac{1}{1 - X}\right)$$

Sketch the graph of this function when a and b are both positive. Three sample observations give these values:

Y	0	5	6
X	1.2	1.5	2.0

Fit the foregoing function to these data. If Y denotes the per capita consumption of peanuts and X denotes income, give a point estimate of the peanut consumption of a millionaire.

2.9. An economist hypothesizes that the average production cost of an article declines with increasing batch size, tending toward an asymptotic minimum value. Some sample data from the process are as follows.

Batch size	1	5	10	20
Average cost ($)	31	14	12	11

Fit a curve of the form

$$Y = \alpha + \beta\left(\frac{1}{X}\right)$$

to these data, where Y denotes average cost and X indicates batch size. What is the estimated minimum cost level? Estimate the batch size that would be required to get the average cost to within 10 percent of this minimum level.

2.10. A variable x_n has the following pdf:

x_n	1	n
$p(x_n)$	$1 - 1/n$	$1/n$

Determine $E(x_n)$ and $\text{var}(x_n)$, and investigate their limits as n becomes infinitely large. Does plim x_n exist? If so, what is its value?

2.11. Let $x_1, x_2, \ldots, x_n$ be a random sample from each of the following pdf's:

(a) Bernoulli distribution

$$f(x;\theta) = \theta^x(1-\theta)^{1-x} \qquad 0 \leq \theta \leq 1 \qquad x = 0, 1$$

(b) Poisson distribution

$$f(x;\theta) = \frac{\theta^x e^{-\theta}}{x!} \qquad x = 0, 1, 2, \ldots \qquad 0 \leq \theta < \infty$$

(c) Exponential distribution

$$f(x;\theta) = \theta e^{-\theta x} \qquad 0 < x < \infty \qquad 0 < \theta < \infty$$

Derive the maximum likelihood estimator of θ in each case. Verify that the second derivative of the log-likelihood is negative in each case, ensuring a maximum of the likelihood function.

2.12. Use your PC to generate 1,000 observations on two *independent* stationary AR(1) series. Drop the first 50 observations and compute the correlation coefficients for sucessive sets of 50 observations. Repeat this exercise for two independent random walks and for two independent explosive series.

CHAPTER 3

The *k*-Variable Linear Equation

Chapters 1 and 2 have developed the basic statistical tools and procedures for analyzing bivariate relationships. Clearly, however, the bivariate framework is too restrictive for realistic analyses of economic phenomena. Common sense and economic theory alike indicate the need to specify and analyze *multivariate* relations. Economic models generally postulate the joint and simultaneous existence of several relations, each of which typically contains more than two variables. The ultimate objective of econometrics therefore is the analysis of *simultaneous equation systems*. For the present, however, we shall restrict the analysis to a *single equation;* but we shall extend it to include *k* variables, where *k* is, in general, a number larger than two.

The specification of such a relationship is then

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots + \beta_k X_{kt} + u_t \qquad t = 1, \ldots, n \tag{3.1}$$

This equation identifies $k-1$ explanatory variables (regressors), namely, $X_2, X_3, \ldots, X_k$, that are thought to influence the dependent variable (regressand). To keep the notation simple, we shall denote all explanatory variables by X_{jt}, where the first subscript indicates the variable in question and the second subscript indicates the particular observation on that variable. The X's may be various transformations of other variables, as in the examples of Chapter 2, but the relationship is linear in the β coefficients. We will assume that the disturbances are white noise, as in Eq. (1.21). Thus there are $k+1$ parameters in the model, namely, the β's and the disturbance variance σ^2. The multivariate relation gives rise to a richer array of inference questions than the two-variable equation. The simplest way of tackling them is to switch to matrix notation, which eliminates a great mass of summation signs, subscripts, and the rest. The relevant matrix algebra is developed in Appendix A; and sections of that Appendix are keyed into the development in this and succeeding chapters, so that the various matrix operations are treated, as far as possible, in the sequence in which they appear in the main text.

3.1 MATRIX FORMULATION OF THE k-VARIABLE MODEL

Matrices and vectors will be indicated by bold letters, with uppercase bold letters for matrices and lowercase bold letters for vectors. In general, vectors will be taken to be column vectors unless otherwise stated. Thus, for example,

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \boldsymbol{x}_2 = \begin{bmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2n} \end{bmatrix}$$

are $n \times 1$ vectors, also referred to as n-vectors, containing the sample observations on Y and X_2. By using this vector notation, the n sample observations on Eq. (3.1) can be written as

$$\begin{bmatrix} \vdots \\ \boldsymbol{y} \\ \vdots \end{bmatrix} = \beta_1 \begin{bmatrix} \vdots \\ \boldsymbol{x}_1 \\ \vdots \end{bmatrix} + \beta_2 \begin{bmatrix} \vdots \\ \boldsymbol{x}_2 \\ \vdots \end{bmatrix} + \cdots + \beta_k \begin{bmatrix} \vdots \\ \boldsymbol{x}_k \\ \vdots \end{bmatrix} + \begin{bmatrix} \vdots \\ \boldsymbol{u} \\ \vdots \end{bmatrix} \tag{3.2}$$

The $\boldsymbol{y}$ vector is thus expressed as a linear combination of the $\boldsymbol{x}$ vectors plus the disturbance vector $\boldsymbol{u}$. The $\boldsymbol{x}_1$ vector is a column of ones to allow for the intercept term. Collecting all the $\boldsymbol{x}$ vectors into a matrix $\boldsymbol{X}$ and the β coefficients into a vector $\boldsymbol{\beta}$ permits an even simpler representation, namely,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} \tag{3.3}$$

where[1]

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{21} & \cdots & X_{k1} \\ 1 & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & \cdots & X_{kn} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

3.1.1 The Algebra of Least Squares

If the unknown vector $\boldsymbol{\beta}$ in Eq. (3.3) is replaced by some guess or estimate $\boldsymbol{b}$, this defines a vector of residuals $\boldsymbol{e}$,

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb}$$

The least-squares principle is to choose $\boldsymbol{b}$ to minimize the residual sum of squares, $\boldsymbol{e}'\boldsymbol{e}$, namely,

[1] The ordering of the subscripts in the X matrix does not conform with conventional usage, because we prefer to use the first subscript to denote the variable and the second to indicate the observation on that variable. Thus, X_{25}, which denotes the fifth observation on X_2, occurs at the intersection of the fifth row and the second column in $\boldsymbol{X}$, rather than the other way around.

$$\begin{aligned}\text{RSS} &= e'e \\ &= (y - Xb)'(y - Xb) \\ &= y'y - b'X'y - y'Xb + b'X'Xb \\ &= y'y - 2b'X'y + b'X'Xb\end{aligned}$$

where this development uses the fact that the transpose of a scalar is the scalar, so that $y'Xb = (y'Xb)' = b'X'y$. As shown in Appendix A, the first-order conditions are

$$\frac{\partial(\text{RSS})}{\partial b} = -2X'y + 2X'Xb = 0$$

giving the **normal equations**

$$(X'X)b = X'y \tag{3.4}$$

These equations show how the least-squares b vector is related to the data.

EXAMPLE 3.1. NORMAL EQUATIONS FOR THE TWO-VARIABLE CASE. To illustrate the matrix equation, we will specialize Eq. (3.4) to the two-variable case and confirm the normal equations derived in Chapter 1. This process corresponds to $k = 2$ in the present notation, and the equation is written $Y = \beta_1 + \beta_2 X + u$. The X matrix is

$$X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

Thus,

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum X \\ \sum X & \sum X^2 \end{bmatrix}$$

and

$$X'y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix}$$

giving

$$\begin{bmatrix} n & \sum X \\ \sum X & \sum X^2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix}$$

or

$$\begin{aligned} nb_1 + b_2 \sum X &= \sum Y \\ b_1 \sum X + b_2 \sum X^2 &= \sum XY \end{aligned}$$

as in Eq. (1.28).

EXAMPLE 3.2. THREE-VARIABLE EQUATION. In a similar fashion it may be shown that the normal equations for fitting a three-variable equation by least squares are

$$nb_1 + b_2 \sum X_2 + b_3 \sum X_3 = \sum Y$$

$$b_1\sum X_2 + b_2\sum X_2^2 + b_3\sum X_2X_3 = \sum X_2Y$$
$$b_1\sum X_3 + b_2\sum X_2X_3 + b_3\sum X_3^2 = \sum X_3Y$$

If $\boldsymbol{y}$ in Eq. (3.4) is replaced by $\boldsymbol{Xb} + \boldsymbol{e}$ the result is

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{b} = \boldsymbol{X}'(\boldsymbol{Xb} + \boldsymbol{e}) = (\boldsymbol{X}'\boldsymbol{X})\boldsymbol{b} + \boldsymbol{X}'\boldsymbol{e}$$

Thus,
$$\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0} \tag{3.5}$$

which is another fundamental least-squares result. The first element in Eq. (3.5) gives $\sum e_t = 0$, that is,

$$\bar{e} = \bar{Y} - b_1 - b_2\bar{X}_2 - \cdots - b_k\bar{X}_k = 0$$

The residuals have zero mean, and the regression plane passes through the point of means in k dimensional space. The remaining elements in Eq. (3.5) are of the form

$$\sum_t X_{it}e_t = 0 \qquad i = 2, \ldots, k$$

As seen in footnote 16 of Chapter 1 this condition means that each regressor has zero sample correlation with the residuals. This in turn implies that $\hat{\boldsymbol{y}}(= \boldsymbol{Xb})$, the vector of regression values for Y, is uncorrelated with $\boldsymbol{e}$, for

$$\hat{\boldsymbol{y}}'\boldsymbol{e} = (\boldsymbol{Xb})'\boldsymbol{e} = \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}$$

3.1.2 Decomposition of the Sum of Squares

The zero covariances between regressors and the residual underlie the decomposition of the sum of squares. Decomposing the $\boldsymbol{y}$ vector into the part explained by the regression and the unexplained part,

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e} = \boldsymbol{Xb} + \boldsymbol{e}$$

it follows that

$$\boldsymbol{y}'\boldsymbol{y} = (\hat{\boldsymbol{y}} + \boldsymbol{e})'(\hat{\boldsymbol{y}} + \boldsymbol{e}) = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \boldsymbol{e}'\boldsymbol{e} = \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Xb} + \boldsymbol{e}'\boldsymbol{e}$$

However, $\boldsymbol{y}'\boldsymbol{y} = \sum_{t=1}^{n} Y_t^2$ is the sum of squares of the actual Y values. Interest normally centers on analyzing the *variation* in Y, measured by the sum of the squared deviations from the sample mean, namely,

$$\sum_t (Y_t - \bar{Y})^2 = \sum_t Y_t^2 - n\bar{Y}^2$$

Thus, subtracting $n\bar{Y}^2$ from each side of the previous decomposition gives a revised decomposition,

$$\begin{array}{ccccc} (\boldsymbol{y}'\boldsymbol{y} - n\bar{Y}^2) & = & (\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Xb} - n\bar{Y}^2) & + & \boldsymbol{e}'\boldsymbol{e} \\ \text{TSS} & = & \text{ESS} & + & \text{RSS} \end{array} \tag{3.6}$$

where TSS indicates the total sum of squares in Y, and ESS and RSS the explained and residual (unexplained) sum of squares.

3.1.3 Equation in Deviation Form

An alternative approach is to begin by expressing all the data in the form of deviations from the sample means. The least-squares equation is

$$Y_t = b_1 + b_2X_{2t} + b_3X_{3t} + \cdots + b_kX_{kt} + e_t \qquad t = 1, \ldots, n$$

Averaging over the sample observations gives

$$\bar{Y} = b_1 + b_2\bar{X}_2 + b_3\bar{X}_3 + \cdots + b_k\bar{X}_k$$

which contains no term in e, since $\bar{e}$ is zero. Subtracting the second equation from the first gives

$$y_t = b_2x_{2t} + b_3x_{3t} + \cdots + b_kx_{kt} + e_t \qquad t = 1, \ldots, n$$

where, as in Chapter 1, lowercase letters denote deviations from sample means. The intercept b_1 disappears from the deviation form of the equation, but it may be recovered from

$$b_1 = \bar{Y} - b_2\bar{X}_2 - \cdots - b_k\bar{X}_k$$

The least-squares slope coefficients $b_2, \ldots, b_k$ are identical in both forms of the regression equation, as are the residuals.

Collecting all n observations, the deviation form of the equation may be written compactly using a transformation matrix,

$$\boldsymbol{A} = \boldsymbol{I}_n - \left(\frac{1}{n}\right)\boldsymbol{ii}' \tag{3.7}$$

where $\boldsymbol{i}$ is a column vector of n ones. As shown in Appendix A, this is a symmetric, idempotent matrix, which, on premultiplication of a vector of n observations, transforms that vector into deviation form. Thus it follows that $\boldsymbol{Ae} = \boldsymbol{e}$ and $\boldsymbol{Ai} = \boldsymbol{0}$. Write the least-squares equations as

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e} = [\boldsymbol{i} \quad \boldsymbol{X}_2]\begin{bmatrix} b_1 \\ \boldsymbol{b}_2 \end{bmatrix} + \boldsymbol{e}$$

where $\boldsymbol{X}_2$ is the $n \times (k-1)$ matrix of observations on the regressors and $\boldsymbol{b}_2$ is the $k-1$ element vector containing the coefficients, $b_2, b_3, \ldots, b_k$. Premultiplying by $\boldsymbol{A}$ gives

$$\boldsymbol{Ay} = [\boldsymbol{0} \quad \boldsymbol{AX}_2]\begin{bmatrix} b_1 \\ \boldsymbol{b}_2 \end{bmatrix} + \boldsymbol{Ae} = (\boldsymbol{AX}_2)\boldsymbol{b}_2 + \boldsymbol{e}$$

or

$$\boldsymbol{y}_* = \boldsymbol{X}_*\boldsymbol{b}_2 + \boldsymbol{e} \tag{3.8}$$

where $\boldsymbol{y}_* = \boldsymbol{Ay}$ and $\boldsymbol{X}_* = \boldsymbol{AX}_2$ give the data in deviation form. Since $\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}$, it follows that $\boldsymbol{X}_*'\boldsymbol{e} = \boldsymbol{0}$. Thus premultiplying Eq. (3.8) by $\boldsymbol{X}_*'$ gives

$$\boldsymbol{X}_*'\boldsymbol{y}_* = (\boldsymbol{X}_*'\boldsymbol{X}_*)\boldsymbol{b}_2$$

which are the familiar normal equations, as in Eq. (3.4), except that now the data have all been expressed in deviation form and the $\boldsymbol{b}_2$ vector contains the $k-1$ slope coefficients and excludes the intercept term. By using Eq. (3.8), the decomposition of the sum of squares may be expressed as

$$\begin{aligned} y_*'y_* &= b_2'X_*'X_*b_2 + e'e \\ \text{TSS} &= \quad \text{ESS} \quad + \text{RSS} \end{aligned} \tag{3.9}$$

The *coefficient of multiple correlation R* is defined as the positive square root of

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{3.10}$$

Thus R^2 measures the proportion of the total variation in Y explained by the linear combination of the regressors. Most computer programs also routinely produce an *adjusted* R^2, denoted by $\bar{R}^2$. This statistic takes explicit account of the number of regressors used in the equation. It is useful for comparing the fit of specifications that differ in the addition or deletion of explanatory variables. The unadjusted R^2 will never decrease with the addition of any variable to the set of regressors. If the added variable is totally irrelevant the ESS simply remains constant. The adjusted coefficient, however, may decrease with the addition of variables of low explanatory power. It is defined as

$$\bar{R}^2 = 1 - \frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)} \tag{3.11}$$

As will be seen later, the numerator and denominator on the right-hand side of Eq. (3.11) are unbiased estimates of the disturbance variance and the variance of Y. The relation between the adjusted and unadjusted coefficients is

$$\begin{aligned} \bar{R}^2 &= 1 - \frac{n-1}{n-k}(1-R^2) \\ &= \frac{1-k}{n-k} + \frac{n-1}{n-k}R^2 \end{aligned} \tag{3.12}$$

Two other frequently used criteria for comparing the fit of various specifications involving different numbers of regressors are the **Schwarz criterion**,[2]

$$\text{SC} = \ln\frac{e'e}{n} + \frac{k}{n}\ln n$$

and the **Akaike information criterion**,[3]

$$\text{AIC} = \ln\frac{e'e}{n} + \frac{2k}{n}$$

One looks for specifications that will reduce the residual sum of squares, but each criterion adds on a penalty, which increases with the number of regressors.

[2] Schwarz, G., "Estimating the Dimension of a Model," *Annals of Statistics*, **6**, 1978, 461–464.

[3] Akaike, H., "Information Theory and an Extension of the Maximum Likelihood Principle," in B. Petrov and F. Csake, eds., *2nd International Symposium on Information Theory*, Budapest, Akademiai Kiado, 1973.

EXAMPLE 3.3. To illustrate these formulae, here is a brief numerical example. The numbers have been kept very simple so as not to obscure the nature of the operations with cumbersome arithmetic. More realistic computer applications will follow later. The sample data are

$$y = \begin{bmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} 1 & 3 & 5 \\ 1 & 1 & 4 \\ 1 & 5 & 6 \\ 1 & 2 & 4 \\ 1 & 4 & 6 \end{bmatrix}$$

where we have already inserted a column of ones in the first column of X. From these data we readily compute

$$X'X = \begin{bmatrix} 5 & 15 & 25 \\ 15 & 55 & 81 \\ 25 & 81 & 129 \end{bmatrix} \quad \text{and} \quad X'y = \begin{bmatrix} 20 \\ 76 \\ 109 \end{bmatrix}$$

The normal equations, Eq. (3.4), are then

$$\begin{bmatrix} 5 & 15 & 25 \\ 15 & 55 & 81 \\ 25 & 81 & 129 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 20 \\ 76 \\ 109 \end{bmatrix}$$

To solve by Gaussian elimination we first subtract three times the first row from the second row and five times the first row from the third. These steps give the revised system,

$$\begin{bmatrix} 5 & 15 & 25 \\ 0 & 10 & 6 \\ 0 & 6 & 4 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 20 \\ 16 \\ 9 \end{bmatrix}$$

Next, subtract six-tenths of the second row from the third to get

$$\begin{bmatrix} 5 & 15 & 25 \\ 0 & 10 & 6 \\ 0 & 0 & 0.4 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 20 \\ 16 \\ -0.6 \end{bmatrix}$$

The third equation gives $0.4b_3 = -0.6$, that is

$$b_3 = -1.5$$

Substituting for b_3 in the second equation, we find that

$$10b_2 + 6b_3 = 16$$

which gives

$$b_2 = 2.5$$

Finally, the first equation

$$5b_1 + 15b_2 + 25b_3 = 20$$

gives[4]

$$b_1 = 4$$

The regression equation is thus

$$\hat{Y} = 4 + 2.5X_2 - 1.5X_3$$

[4] The sample data have yielded a unique solution for the b vector. The condition required for a unique solution will be examined later in this chapter.

Alternatively, transforming the data into deviation form gives

$$y_* = Ay = \begin{bmatrix} -1 \\ -3 \\ 4 \\ -1 \\ 1 \end{bmatrix} \quad \text{and} \quad X_* = AX_2 = \begin{bmatrix} 0 & 0 \\ -2 & -1 \\ 2 & 1 \\ -1 & -1 \\ 1 & 1 \end{bmatrix}$$

The relevant normal equations are then

$$\begin{bmatrix} 10 & 6 \\ 6 & 4 \end{bmatrix} \begin{bmatrix} b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 16 \\ 9 \end{bmatrix}$$

These are the second and third equations obtained in the first step of the Gaussian elimination above.[5] Thus the solutions for b_2 and b_3 coincide with those already obtained. Likewise b_1 will be the same as before, since the final equation in the foregoing back substitution is readily seen to be

$$b_1 = \bar{Y} - b_2\bar{X}_2 - b_3\bar{X}_3$$

From the y_* vector, TSS is seen to be 28. ESS may be computed from

$$b_*'X_*'X_*b_* = [2.5 \quad -1.5] \begin{bmatrix} 10 & 6 \\ 6 & 4 \end{bmatrix} \begin{bmatrix} 2.5 \\ -1.5 \end{bmatrix} = 26.5$$

or, more simply, from

$$b_*'X_*'y_* = [2.5 \quad -1.5] \begin{bmatrix} 16 \\ 9 \end{bmatrix} = 26.5$$

Then RSS is 1.5, $R^2 = 0.95$ and $\bar{R}^2 = 0.89$.

3.2 PARTIAL CORRELATION COEFFICIENTS

With two or more regressors, partial correlations become relevant. Consider again the Plosser/Schwert example in Chapter 1 of a high positive correlation between the logs of nominal income in the United States and accumulated sunspots. It was suggested there that each variable displayed independent time trends and that the influence of this common variable (time) was basically responsible for the observed correlation between income and sunspots. This supposition can be checked by fitting time trends to each variable separately, computing the unexplained residuals from these time trends, and examining the correlation between the residuals. To simplify the notation let

$$\begin{aligned} Y &= \text{log of nominal income} \\ X_2 &= \text{log of accumulated sunspots} \\ X_3 &= \text{time (measured in years)} \end{aligned}$$

[5] See Problem 3.1.

We will also use the index 1 to refer to Y, 2 for X_2, and 3 for X_3. Then

$$\begin{aligned}
r_{12} &= \text{correlation between } Y \text{ and } X_2 \\
r_{23} &= \text{correlation between } X_2 \text{ and } X_3\text{, etc.} \\
b_{12} &= \text{slope of the regression of } Y \text{ on } X_2 \\
b_{32} &= \text{slope of the regression of } X_3 \text{ on } X_2\text{, etc.} \\
e_{1.2} &= \text{residual from the regression of } Y \text{ on } X_2 \\
e_{3.2} &= \text{residual from the regression of } X_3 \text{ on } X_2\text{, etc.}
\end{aligned}$$

Working with the data in deviation form, the unexplained residual from the regression of log income on time is

$$e_{1.3} = y - b_{13}x_3 \qquad \text{where} \qquad b_{13} = \frac{\sum yx_3}{\sum x_3^2}$$

To keep the equations uncluttered, we omit the observation subscripts. Likewise the unexplained residual from the regression of the log of accumulated sunspots on time is

$$e_{2.3} = x_2 - b_{23}x_3 \qquad \text{where} \qquad b_{23} = \frac{\sum x_2x_3}{\sum x_3^2}$$

The *partial correlation coefficient* between income and sunspots, with the influence of time removed or held constant, is defined as the correlation coefficient between these two sets of residuals. It is denoted by $r_{12.3}$. Thus,

$$r_{12.3} = \frac{\sum e_{1.3}e_{2.3}}{\sqrt{\sum e_{1.3}^2}\sqrt{\sum e_{2.3}^2}} \tag{3.13}$$

Since the e's are least-squares residuals, they have zero means, so there is no need for a correction for means in Eq. (3.13). One could implement Eq. (3.13) directly by computing the two sets of residuals and then calculating the correlation coefficient between them. In practice, however, it is simpler to express $r_{12.3}$ in terms of the three simple correlation coefficients, r_{12}, r_{13}, and r_{23}.[6] That is,

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} \tag{3.14}$$

The simple correlation coefficients are often called *zero-order* coefficients and coefficients like $r_{12.3}$ *first-order* coefficients, since the common influence of one other variable has been taken into account. In a typical spurious correlation case the zero-order correlations tend to be large, whereas $r_{12.3}$ will be negligible. In general, however, first-order correlations may be larger or smaller than the corresponding zero-

[6] See Appendix 3.1.

order coefficients, and they may not even have the same sign. In the three-variable case there are two other first-order coefficients, namely, $r_{13.2}$ and $r_{23.1}$. The first measures the net association between Y and X_3 when any common influence from X_2 has been allowed for, and the second measures the net association between X_2 and X_3 when any common effect from Y has been removed. In a single-equation specification, $r_{12.3}$ and $r_{13.2}$ are usually the first-order coefficients of interest. The formula for $r_{13.2}$ may be derived from first principles or from Eq. (3.14) by interchanging subscripts 2 and 3. Thus,

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}} \tag{3.15}$$

3.2.1 Sequential Buildup of the Explained Sum of Squares

The main objective of the single-equation analysis is to explain the variation $\sum y^2$ in the dependent variable. Running a regression of Y on X_2 gives an explained sum of squares of $r_{12}^2 \sum y^2$, leaving a residual sum of squares of $(1 - r_{12}^2)\sum y^2 = \sum e_{1.2}^2$. The proportion of this residual variation explained by using the *adjusted* X_3 variable, that is, $e_{3.2} = x_3 - b_{32}x_2$, is $r_{13.2}^2$. Thus the increment in the ESS at the second stage is $r_{13.2}^2(1 - r_{12}^2)\sum y^2$. Adding the ESS at each stage gives a total ESS of $[r_{12}^2 + r_{13.2}^2(1 - r_{12}^2)]\sum y^2$. Alternatively, the multiple regression of Y on X_2 and X_3 gives a total ESS, which may be written as $R_{1.23}^2 \sum y^2$, where $R_{1.23}$ is the coefficient of multiple correlation, using the index notation to show explicitly which variables are involved. In this approach the increment in ESS due to the addition of X_3 is $(R_{1.23}^2 - r_{12}^2)\sum y^2$. The two expressions for the incremental ESS are identical, for it may be shown[7] that

$$R_{1.23}^2 = r_{12}^2 + r_{13.2}^2(1 - r_{12}^2) \tag{3.16}$$

The sequence may alternatively be started with X_3, and the increment due to X_2 computed. Both sequences are shown in Table 3.1.

TABLE 3.1
Buildup of the explained sum of squares

Variable	Sum of squares	Variable	Sum of squares
X_2	$r_{12}^2 \sum y^2$	X_3	$r_{13}^2 \sum y^2$
Increment due to X_3	$r_{13.2}^2(1 - r_{12}^2)\sum y^2$	Increment due to X_2	$r_{12.3}^2(1 - r_{13}^2)\sum y^2$
X_2 and X_3	$R_{1.23}^2 \sum y^2$	X_2 and X_3	$R_{1.23}^2 \sum y^2$
Residual	$(1 - R_{1.23}^2)\sum y^2$	Residual	$(1 - R_{1.23}^2)\sum y^2$

[7] See Problem 3.3.

EXAMPLE 3.4. The data of Example 3.3 may be used to illustrate Table 3.1. We have

$$r_{12}^2 = \frac{(16)^2}{(28)(10)} = 0.9143 \qquad r_{12} = 0.9562$$

$$r_{13}^2 = \frac{(9)^2}{(28)(4)} = 0.7232 \qquad r_{13} = 0.8504$$

$$r_{23}^2 = \frac{(6)^2}{(10)(4)} = 0.9000 \qquad r_{23} = 0.9487$$

Remember that the *signs* of the correlation coefficients must be determined from the covariances. If one calculates the squared correlations, one must not simply take the positive square roots in calculating the coefficients. The partial correlations are then

$$r_{12.3} = \frac{0.9562 - (0.8504)(0.9487)}{\sqrt{1 - 0.7232}\sqrt{1 - 0.9000}} = 0.8982$$

$$r_{13.2} = \frac{0.8504 - (0.9562)(0.9487)}{\sqrt{1 - 0.9143}\sqrt{1 - 0.9000}} = -0.6130$$

Thus $r_{12.3}^2 = 0.8067$ and $r_{13.2}^2 = 0.3758$. The various sums of squares for Table 3.1 may then be computed as

$$r_{12}^2 \sum y^2 = 25.6 \qquad r_{13.2}^2(1 - r_{12}^2)\sum y^2 = 0.9$$

$$r_{13}^2 \sum y^2 = 20.25 \qquad r_{12.3}^2(1 - r_{13}^2)\sum y^2 = 6.25$$

Table 3.2 collects the results.

The total ESS was shown in Example 3.3 to be 26.5, and the same result is reached here by use of the simple and partial correlation coefficients.

With two (or more) explanatory variables there is no unambiguous way of determining the relative importance of each variable in explaining the movement in Y, except in the extreme and unlikely case of zero correlation between the explanatory variables. When r_{23} is zero, the variables are said to be *orthogonal,* and it may be shown[8] that $R_{1.23}^2 = r_{12}^2 + r_{13}^2$. Thus in this special case ESS may be split into two components, each of which is unambiguously attributed to an explanatory variable. When the X's are correlated, no such split is possible. Kruskal considers various

TABLE 3.2
Sum of squares from Example 3.3

Variable	Sum of squares	Variable	Sum of squares
X_2	25.6	X_3	20.25
Increment due to X_3	0.9	Increment due to X_2	6.25
X_2 and X_3	26.5	X_2 and X_3	26.5
Residual	1.5	Residual	1.5

[8] See Problem 3.3.

methods of assessing the relative importance of different explanatory variables.[9] He pays most attention to averaging the squared simple and partial correlation coefficients over the various possible orders in which the X's might be introduced. The rationale is that at each stage the relevant squared coefficient indicates the proportion of the remaining variance that is explained by a specific X variable. For this example his method gives

$$\begin{aligned}
\text{Average proportion for } X_2 &= (r_{12}^2 + r_{12.3}^2)/2 \\
&= (0.9143 + 0.8067)/2 = 0.86 \\
\text{Average proportion for } X_3 &= (r_{13}^2 + r_{13.2}^2)/2 \\
&= (0.7232 + 0.3758)/2 = 0.55
\end{aligned}$$

Kruskal has some reservations about applying this technique in a regression framework, but his article does contain an interesting application of the method to data from Friedman and Meiselman on the perennial topic of the relative importance of money and autonomous expenditure in the determination of income.

A different way of illustrating the relative contributions of the explanatory variables is a **Tinbergen diagram.** These diagrams were used extensively in Tinbergen's pioneering study of business cycles.[10] The appropriate diagram for this numerical example is shown in Fig. 3.1. Working with deviations, Fig. 3.1*a* shows

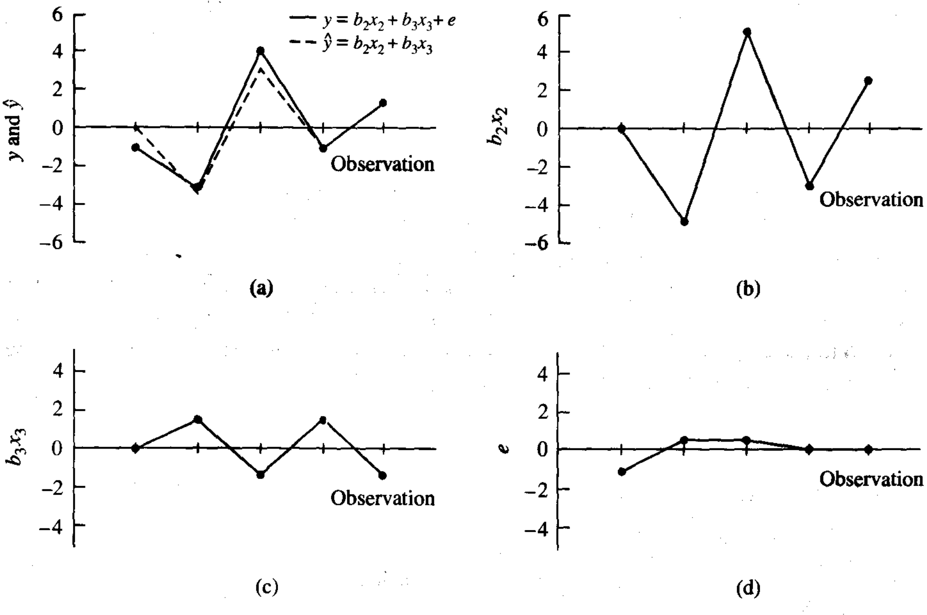


FIGURE 3.1
Tinbergen Diagram for Example 3.3

[9] William Kruskal, "Relative Importance by Averaging over Orderings," *The American Statistician*, 1987, **41**, 6–10.

[10] J. Tinbergen, *Business Cycles in the United States of America, 1919–1932*, League of Nations, 1939.

the actual and calculated y values; parts (b) and (c) show b_2x_2 and b_3x_3, respectively; and part (d) shows the residuals from the regression. All parts should have the same vertical scale to facilitate visual comparisons. This diagram and the Kruskal average coefficients both suggest a larger role for X_2 than for X_3 in the determination of Y in this example.

3.2.2 Partial Correlation Coefficients and Multiple Regression Coefficients

There are two regression slope coefficients in the three-variable equation

$$y = b_2x_2 + b_3x_3 + e$$

Alternatively, one could obtain a slope coefficient from the regression of $e_{1.3}$ on $e_{2.3}$, and another slope coefficient from the regression of $e_{1.2}$ on $e_{3.2}$. Let us denote these two coefficients by $b_{12.3}$ and $b_{13.2}$, respectively, since they come from the series used to calculate the corresponding partial correlation coefficients. What is the relation of these latter regression coefficients to b_2 and b_3 from the multiple regression? The answer is, they are identical: $b_{12.3} = b_2$ and $b_{13.2} = b_3$. The multiple regression coefficients come from the normal equations

$$\begin{bmatrix} \sum x_2^2 & \sum x_2x_3 \\ \sum x_2x_3 & \sum x_3^2 \end{bmatrix} \begin{bmatrix} b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} \sum yx_2 \\ \sum yx_3 \end{bmatrix}$$

Solving for b_2 shows

$$b_2 = \frac{\sum x_3^2 \sum yx_2 - \sum x_2x_3 \sum yx_3}{\sum x_2^2 \sum x_3^2 - (\sum x_2x_3)^2}$$

From the first pair of residuals, we find

$$b_{12.3} = \frac{\sum e_{1.3}e_{2.3}}{\sum e_{2.3}^2} = \frac{\sum(y - b_{13}x_3)(x_2 - b_{23}x_3)}{\sum(x_2 - b_{23}x_3)^2}$$

Algebraic simplification of this last expression gives $b_{12.3} = b_2$. The equality of the other two coefficients may be shown in a similar fashion.

In the early days of econometrics there was some confusion over the use of time in regression analysis. The preceding result shows that it does not matter whether time is included among the explanatory variables or whether the variables are "detrended" before being put into the regression. Suppose, for example, that a demand function is specified as

$$Q = AP^{\beta_2}e^{\beta_3 T}$$

where Q measures the quantity demanded, P indicates price, and T denotes time. The price elasticity is β_2, and the rate of shift in the quantity demanded per unit of time is β_3. Taking logs gives

$$\ln Q = \beta_1 + \beta_2 \ln P + \beta_3 T$$

The price elasticity may be estimated directly by fitting this multiple regression, or by removing a linear time trend from both $\ln Q$ and $\ln P$ and estimating the slope of the regression of the first residual on the second. Notice, however, that neither of the coefficients of time in the separate trend analyses is an estimate of the β_3 shift parameter.[11]

3.2.3 General Treatment of Partial Correlation and Multiple Regression Coefficients

Under the conditions shown in the next section, the normal equations solve for $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. The residuals from the LS regression may then be expressed as

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y} \tag{3.17}$$

where $$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

It is readily seen that $\boldsymbol{M}$ is a symmetric, idempotent matrix. It also has the properties that $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}$ and $\boldsymbol{M}\boldsymbol{e} = \boldsymbol{e}$. Now write the general regression in partitioned form as

$$\boldsymbol{y} = [\boldsymbol{x}_2 \quad \boldsymbol{X}_*]\begin{bmatrix} b_2 \\ \boldsymbol{b}_{(2)} \end{bmatrix} + \boldsymbol{e}$$

In this partitioning $\boldsymbol{x}_2$ is the $n \times 1$ vector of observations on X_2, with coefficient b_2, and $\boldsymbol{X}_*$ is the $n \times (k-1)$ matrix of all the other variables (including the column of ones) with coefficient vector $\boldsymbol{b}_{(2)}$.[12] The normal equations for this setup are

$$\begin{bmatrix} \boldsymbol{x}_2'\boldsymbol{x}_2 & \boldsymbol{x}_2'\boldsymbol{X}_* \\ \boldsymbol{X}_*'\boldsymbol{x}_2 & \boldsymbol{X}_*'\boldsymbol{X}_* \end{bmatrix}\begin{bmatrix} b_2 \\ \boldsymbol{b}_{(2)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_2'\boldsymbol{y} \\ \boldsymbol{X}_*'\boldsymbol{y} \end{bmatrix}$$

We wish to solve for b_2 and interpret the result in terms of a simple regression slope. The solution is[13]

$$b_2 = (\boldsymbol{x}_2'\boldsymbol{M}_*\boldsymbol{x}_2)^{-1}(\boldsymbol{x}_2'\boldsymbol{M}_*\boldsymbol{y}) \tag{3.18}$$

where $$\boldsymbol{M}_* = \boldsymbol{I} - \boldsymbol{X}_*(\boldsymbol{X}_*'\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*'$$

$\boldsymbol{M}_*$ is a symmetric, idempotent matrix with the properties $\boldsymbol{M}_*\boldsymbol{X}_* = \boldsymbol{0}$ and $\boldsymbol{M}_*\boldsymbol{e} = \boldsymbol{e}$. Now by analogy with Eq. (3.17) it follows that

$\boldsymbol{M}_*\boldsymbol{y}$ is the vector of residuals when $\boldsymbol{y}$ is regressed on $\boldsymbol{X}_*$

and $\boldsymbol{M}_*\boldsymbol{x}_2$ is the vector of residuals when $\boldsymbol{x}_2$ is regressed on $\boldsymbol{X}_*$

Regressing the first vector on the second gives a slope coefficient, which, using the symmetry and idempotency of $\boldsymbol{M}_*$, gives the b_2 coefficient already defined in Eq. (3.18). This general result has already been illustrated for the three-variable case.

A simpler and elegant way of proving the same result is as follows. Write the partitioned regression as

$$\boldsymbol{y} = \boldsymbol{x}_2 b_2 + \boldsymbol{X}_*\boldsymbol{b}_{(2)} + \boldsymbol{e}$$

[11] See Problem 3.4.

[12] Note that this is a different use of the star subscript than in an earlier section where it was used to indicate data in deviation form.

[13] See Appendix 3.2.

Premultiply by $\boldsymbol{M}_*$ to obtain

$$\boldsymbol{M}_*\boldsymbol{y} = (\boldsymbol{M}_*\boldsymbol{x}_2)b_2 + \boldsymbol{e}$$

Finally, premultiply by $\boldsymbol{x}_2'$, which gives

$$\boldsymbol{x}_2'\boldsymbol{M}_*\boldsymbol{y} = (\boldsymbol{x}_2'\boldsymbol{M}_*\boldsymbol{x}_2)b_2$$

which replicates Eq. (3.18).

The partial correlation coefficient between Y and X_2, conditional on $X_3, \ldots, X_k$ is defined as

$$\begin{aligned} r_{12.3\ldots k} &= \text{correlation coefficient between } (\boldsymbol{M}_*\boldsymbol{y}) \text{ and } (\boldsymbol{M}_*\boldsymbol{x}_2) \\ &= \frac{\boldsymbol{x}_2'\boldsymbol{M}_*\boldsymbol{y}}{\sqrt{\boldsymbol{x}_2'\boldsymbol{M}_*\boldsymbol{x}_2}\sqrt{\boldsymbol{y}'\boldsymbol{M}_*\boldsymbol{y}}} \end{aligned} \tag{3.19}$$

Comparison of Eq. (3.19) with Eq. (3.18) gives

$$\begin{aligned} b_2 &= r_{12.34\ldots k}\frac{\sqrt{\boldsymbol{y}'\boldsymbol{M}_*\boldsymbol{y}}}{\sqrt{\boldsymbol{x}_2'\boldsymbol{M}_*\boldsymbol{x}_2}} \\ &= r_{12.34\ldots k}\frac{s_{1.34\ldots k}}{s_{2.34\ldots k}} \end{aligned} \tag{3.20}$$

where $s_{1.34\ldots k}$ is the standard deviation of the residuals from the regression of Y on a constant and $X_3, \ldots, X_k$; and $s_{2.34\ldots k}$ is the standard deviation of the residuals from the regression of X_2 on the same variables. Equation (3.20) is the multivariate version of Eq. (1.30) for the two-variable model. The other partial correlation coefficients and multiple regression coefficients may be derived by replacing x_2 by x_i $(i = 3, \ldots, k)$ in Eqs. (3.19) and (3.20) and making the corresponding changes in $\boldsymbol{M}_*$.

3.3 THE GEOMETRY OF LEAST SQUARES

The simplest case is shown in Fig. 3.2, with a $\boldsymbol{y}$ vector and an $\boldsymbol{x}$ vector for a single explanatory variable. The two vectors lie in $\mathbf{E}^n$, that is, n-dimensional Euclidean space.[14] The $\boldsymbol{y}$ vector may be expressed as $\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e}$, where $\hat{\boldsymbol{y}} = b\boldsymbol{x}$ is some multiple

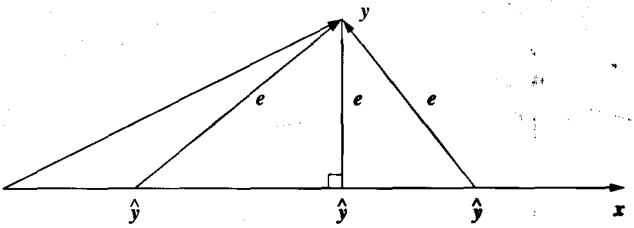


FIGURE 3.2

[14] To distinguish between the use of the same letter for Euclidean space and for the expectation operator we will use bold $\mathbf{E}^n$ for the former and italic, nonbold E for the latter.

of the $\boldsymbol{x}$ vector. Three possibilities are shown in the figure. The least-squares principle is to choose b to make $\hat{y}$ as close as possible to y. This is achieved by making the length of $\boldsymbol{e}$ a minimum, that is, by dropping a perpendicular from y to $\boldsymbol{x}$. As shown in Appendix A, the condition for $\boldsymbol{x}$ and $\boldsymbol{e}$ to be orthogonal is $\boldsymbol{x}'\boldsymbol{e} = 0$. This gives $\boldsymbol{x}'(y - b\boldsymbol{x}) = 0$, or $b = \boldsymbol{x}'y/\boldsymbol{x}'\boldsymbol{x}$. Then

$$\begin{aligned} \hat{y} &= xb = x\left(\frac{x'y}{x'x}\right) \\ &= \left(\frac{xx'}{x'x}\right)y = (x(x'x)^{-1}x')y \\ &= Py \end{aligned}$$

where
$$P = x(x'x)^{-1}x'$$

Notice that $\boldsymbol{xx}'$ is an $n \times n$ matrix, whereas $\boldsymbol{x}'\boldsymbol{x}$ is a scalar. The matrix $\boldsymbol{P}$ is seen to be symmetric, idempotent. It is called a **projection matrix,** since postmultiplication by $\boldsymbol{y}$ gives the projection of the $\boldsymbol{y}$ vector onto the $\boldsymbol{x}$ vector.

Figure 3.3 shows the case where there are two explanatory variables, with vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. All linear combinations of these two vectors define a two-dimensional subspace of E^n. This is the **column space** of $\boldsymbol{X}$. The residual vector must be perpendicular to this column space, which requires $\boldsymbol{e}$ to be orthogonal to both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, which may be stated compactly as

$$X'e = \mathbf{0}$$

This condition was derived algebraically in Eq. (3.5), which in turn gives the normal equations $(X'X)b = X'y$. From the parallelogram law for the addition of vectors it is clear from Fig. 3.3 that $\hat{y}$ may be expressed as a unique linear combination of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The equivalent algebraic condition is that the normal equations solve for a unique $\boldsymbol{b}$ vector. The $\boldsymbol{x}$ vectors in Fig. 3.3 are linearly independent; thus the column space of $\boldsymbol{X}$ has dimension two, which is the rank of $\boldsymbol{X}$. As shown in Appendix A,

$$\text{Rank of } X = \text{rank of } (X'X) = \text{rank of } (XX')$$

Thus $(X'X)$ is nonsingular, and the normal equations solve for $b = (X'X)^{-1}X'y$. Figure 3.4 shows a case where the two $\boldsymbol{x}$ vectors are linearly dependent. The $\hat{y}$ point is still uniquely determined by the perpendicular from y to the line through the $\boldsymbol{x}$ vectors, but there is no unique representation of $\hat{y}$ in terms of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

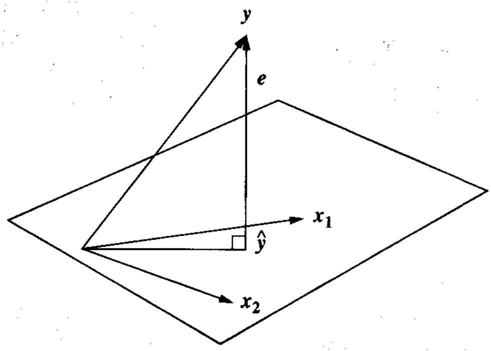


FIGURE 3.3

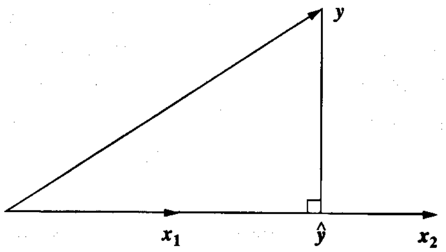


FIGURE 3.4

In the general case the X matrix is a set of k column vectors

$$X = \begin{bmatrix} \vdots & \vdots & & \vdots \\ x_1 & x_2 & \cdots & x_k \\ \vdots & \vdots & & \vdots \end{bmatrix}$$

Any linear combination of these vectors lies in the column space of X. If y, the vector of observations on the dependent variable, were also in the column space, then there would be at least one $k \times 1$ vector b satisfying $y = Xb$. The linear combination of the explanatory variables would account for all of the variation in Y, leaving zero unexplained variation. Such an outcome is totally unlikely in practice; the typical situation is shown in Fig. 3.5, where the y vector lies outside the column space of X. By denoting any arbitrary linear combination of the columns of X by $\hat{y} = Xb$, the y vector can be expressed as $y = \hat{y} + e$. The least-squares principle is to choose $\hat{y}$ to minimize the length of the e vector. This is achieved when the $\hat{y}$ and e vectors are orthogonal. Since $\hat{y}$ is a linear combination of the columns of X, this requires that e be orthogonal to each x_i vector, giving

$$x_i'e = 0 \qquad i = 1, 2, \ldots, k$$

or, written more compactly,

$$X'e = 0$$

which is the same equation as that derived earlier for the case of just two explanatory variables. If the columns of X are linearly independent (i.e., X has full column rank) then $\hat{y}$ can be expressed as a unique linear combination of the x_i vectors (the

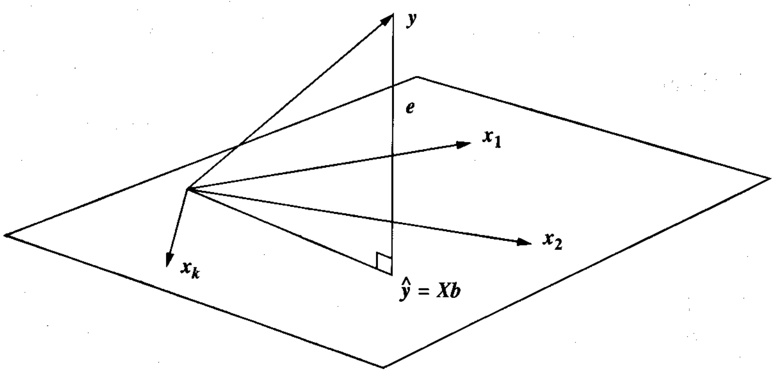


FIGURE 3.5

normal equations solve for a unique $\boldsymbol{b}$). If, however, the column vectors are linearly dependent, $\hat{\boldsymbol{y}}$ is still given uniquely by the perpendicular from $\boldsymbol{y}$ to the column space, but one can find one or more $\boldsymbol{c}$ vectors satisfying $\boldsymbol{Xc} = \boldsymbol{0}$. Then

$$\hat{\boldsymbol{y}} = \boldsymbol{Xb} = \boldsymbol{Xb} + \boldsymbol{Xc} = \boldsymbol{X}(\boldsymbol{b} + \boldsymbol{c})$$

which says that $\hat{\boldsymbol{y}}$ cannot be expressed as a unique linear combination of the $\boldsymbol{x}_i$'s, and so the normal equations do not solve for a unique $\boldsymbol{b}$.

To sum up, least squares requires that X has rank k, so that $(X'X)$ is nonsingular and the normal equations solve for a unique $\boldsymbol{b}$.

3.4 INFERENCE IN THE k-VARIABLE EQUATION

Next we need to establish the statistical properties of the least-squares estimator and to derive appropriate inference procedures. These depend on what assumptions are made in the specification of the relationship.

3.4.1 Assumptions

1. X is nonstochastic and has full column rank k.
   Inferences will be conditional on the sample values of the X variables, so the elements of the X matrix are treated as fixed in repeated sampling. As shown in Section 3.3, linear independence of the columns of X is required for a unique determination of the $\boldsymbol{b}$ vector.
2. The disturbances have the properties

$$E(\boldsymbol{u}) = \boldsymbol{0} \tag{3.21}$$

and

$$\text{var}(\boldsymbol{u}) = E(\boldsymbol{uu}') = \sigma^2\boldsymbol{I} \tag{3.22}$$

When the expectation operator E is applied to a vector or matrix, it is applied to every element in that vector or matrix. Thus, Eq. (3.21) gives

$$E(\boldsymbol{u}) = E\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} E(u_1) \\ E(u_2) \\ \vdots \\ E(u_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \boldsymbol{0}$$

and Eq. (3.22) gives

$$E(\boldsymbol{uu}') = E\left(\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}\begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}\right) = \begin{pmatrix} E(u_1^2) & E(u_1u_2) & \cdots & E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) & \cdots & E(u_2u_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(u_nu_1) & E(u_nu_2) & \cdots & E(u_n^2) \end{pmatrix}$$

$$= \begin{pmatrix} \text{var}(u_1) & \text{cov}(u_1, u_2) & \cdots & \text{cov}(u_1, u_n) \\ \text{cov}(u_2, u_1) & \text{var}(u_2) & \cdots & \text{cov}(u_2, u_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(u_n, u_1) & \text{cov}(u_n, u_2) & \cdots & \text{var}(u_n) \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2\boldsymbol{I}$$

This matrix is the variance-covariance matrix of the disturbance term. The variances are displayed on the main diagonal and the covariances in the off-diagonal positions. We will denote it by the shorthand notation var($\boldsymbol{u}$). It is sometimes indicated by $V(\boldsymbol{u})$ or by cov($\boldsymbol{u}$).

This variance matrix embodies two strong assumptions. The first is that the disturbance variance is constant at each sample point. This condition is termed **homoscedasticity,** and the obverse condition, where the disturbance variances are not the same at all points, is termed **heteroscedasticity.** The second assumption is that the disturbances are *pairwise uncorrelated.* In a cross-section study of, say, the expenditure patterns of households, this implies zero covariances between the disturbances of different households. With time series data, the implication is zero covariances between disturbances at different time periods. When this condition fails, the disturbances are said to be **autocorrelated** or **serially correlated.**

3.4.2 Mean and Variance of *b*

For theoretical derivations it is simpler to rewrite the normal equations as

$$\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y}$$

Substituting for $\boldsymbol{y}$ gives

$$\boldsymbol{b} = (X'X)^{-1}X'(X\boldsymbol{\beta} + \boldsymbol{u}) = \boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{u}$$

from which

$$\boldsymbol{b} - \boldsymbol{\beta} = (X'X)^{-1}X'\boldsymbol{u} \tag{3.23}$$

In taking expectations the expectation operator may be moved to the right past nonstochastic terms such as X, but must be applied to any stochastic variable. Thus,

$$E(\boldsymbol{b} - \boldsymbol{\beta}) = (X'X)^{-1}X'E(\boldsymbol{u}) = \boldsymbol{0}$$

giving

$$E(\boldsymbol{b}) = \boldsymbol{\beta} \tag{3.24}$$

Thus, *under the assumptions of this model,* the LS coefficients are unbiased estimates of the $\boldsymbol{\beta}$ parameters. The variance-covariance matrix of the LS estimates is established as follows. From first principles, as in the development of Eq. (3.22),

$$\text{var}(\boldsymbol{b}) = E[(\boldsymbol{b} - \boldsymbol{\beta})(\boldsymbol{b} - \boldsymbol{\beta})']$$

If we substitute from Eq. (3.23),

$$\begin{aligned} E[(\boldsymbol{b} - \boldsymbol{\beta})(\boldsymbol{b} - \boldsymbol{\beta})'] &= E[(X'X)^{-1}X'\boldsymbol{uu}'X(X'X)^{-1}] \\ &= (X'X)^{-1}X'E[\boldsymbol{uu}']X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

Thus,

$$\text{var}(\boldsymbol{b}) = \sigma^2(X'X)^{-1} \tag{3.25}$$

This expression is a $k \times k$ matrix with the sampling variances of the b_i displayed on the main diagonal and the covariances in the off-diagonal positions.

EXAMPLE 3.5. STANDARD ERRORS IN THE TWO-VARIABLE EQUATION. As already shown in Example 3.1, $X'X$ in this case is

$$X'X = \begin{bmatrix} n & \sum X \\ \sum X & \sum X^2 \end{bmatrix}$$

Thus,
$$(X'X)^{-1} = \frac{1}{D}\begin{bmatrix} \sum X^2 & -\sum X \\ -\sum X & n \end{bmatrix}$$

where D is the determinant of $(X'X)$,

$$D = n\sum X^2 - \left(\sum X\right)^2 = n\sum x^2$$

Then, denoting the LS intercept and slope by a and b, respectively, we have

$$\text{var}(b) = \frac{\sigma^2}{\sum x^2}$$

which confirms Eq. (1.40). Similarly,

$$\begin{aligned} \text{var}(a) &= \frac{\sigma^2 \sum X^2}{n\sum x^2} \\ &= \frac{\sigma^2(\sum x^2 + n\bar{X}^2)}{n\sum x^2} \\ &= \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x^2}\right) \end{aligned}$$

which confirms Eq. (1.42). The square roots of these variances are frequently referred to as **standard errors.** They are the standard deviations of the marginal distributions of a and b.[15] Finally it may be seen that

$$\text{cov}(a, b) = -\sigma^2 \frac{\bar{X}}{\sum x^2}$$

which confirms Equation (1.43).

EXAMPLE 3.6. A THREE-VARIABLE EQUATION. In most economic applications interest centers on the regression *slope* coefficients rather than on the intercept term. One can then work with the data in deviation form. Expressions like Eq. (3.25) still hold, and the problem becomes two-dimensional rather than three-dimensional. Thus,

$$\text{var}(\boldsymbol{b}) = \begin{bmatrix} \text{var}(b_2) & \text{cov}(b_2, b_3) \\ \text{cov}(b_2, b_3) & \text{var}(b_3) \end{bmatrix} = \sigma^2 \begin{bmatrix} \sum x_2^2 & \sum x_2 x_3 \\ \sum x_2 x_3 & \sum x_3^2 \end{bmatrix}^{-1}$$

Some algebra then shows that

$$\text{var}(b_2) = \frac{\sigma^2}{\sum x_2^2(1 - r_{23}^2)} \quad \text{and} \quad \text{var}(b_3) = \frac{\sigma^2}{\sum x_3^2(1 - r_{23}^2)}$$

where r_{23} is the correlation between X_2 and X_3. If the explanatory variables are uncorrelated, these sampling variances reduce to those for the simple regressions of Y on X_2 and Y on X_3. However, increasing correlation between the explanatory variables inflates the standard errors beyond those that would pertain in the orthogonal case. The more the

[15] These formulae are nonoperational since σ^2 is unknown. When it is replaced by the estimator, s^2, derived in Section 3.4.3, we have *estimated* standard errors. The term standard error is thus used interchangeably to refer to true or estimated standard errors, the choice usually being obvious from the context.

X's look alike, the more imprecise is the attempt to estimate their relative effects. This situation is referred to as *multicollinearity* or *collinearity*. With perfect or exact collinearity the standard errors go to infinity. Exact collinearity means that the columns of X are linearly dependent, and so the LS vector cannot be estimated.

3.4.3 Estimation of σ^2

The variance-covariance matrix in Eq. (3.25) involves the disturbance variance σ^2, which is unknown. It is reasonable to base an estimate on the residual sum of squares from the fitted regression. Following Eq. (3.17), we write $\boldsymbol{e} = \boldsymbol{My} = \boldsymbol{M}(\boldsymbol{X\beta} + \boldsymbol{u}) = \boldsymbol{Mu}$, since $\boldsymbol{MX} = \boldsymbol{0}$.

Thus,
$$E(\boldsymbol{e}'\boldsymbol{e}) = E(\boldsymbol{u}'\boldsymbol{M}'\boldsymbol{Mu}) = E(\boldsymbol{u}'\boldsymbol{Mu})$$

Utilizing the fact that the trace of a scalar is the scalar, we write

$$\begin{aligned}
E(\boldsymbol{u}'\boldsymbol{Mu}) &= E[\mathrm{tr}(\boldsymbol{u}'\boldsymbol{Mu})] \\
&= E[\mathrm{tr}(\boldsymbol{uu}'\boldsymbol{M})] \\
&= \sigma^2\mathrm{tr}(\boldsymbol{M}) \\
&= \sigma^2\mathrm{tr}\boldsymbol{I} - \sigma^2\mathrm{tr}[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'] \\
&= \sigma^2\mathrm{tr}\boldsymbol{I} - \sigma^2\mathrm{tr}[(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})] \\
&= \sigma^2(n-k)
\end{aligned}$$

Thus,
$$s^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-k} \tag{3.26}$$

defines an unbiased estimator of σ^2. The square root s is the standard deviation of the Y values about the regression plane. It is often referred to as the **standard error of estimate** or the **standard error of the regression** (SER).

3.4.4 Gauss–Markov Theorem

This is the fundamental least-squares theorem. It states that, conditional on the assumptions made, no other linear, unbiased estimator of the β coefficients can have smaller sampling variances than those of the least-squares estimator, given in Eq. (3.25). We prove a more general result relating to any linear combination of the β coefficients. Let $\boldsymbol{c}$ denote an arbitrary k-element vector of known constants, and define a scalar quantity μ as

$$\mu = \boldsymbol{c}'\boldsymbol{\beta}$$

If we choose $\boldsymbol{c}' = [0 \quad 1 \quad 0 \quad \cdots \quad 0]$, then $\mu = \beta_2$. Thus, we can pick out any single element in $\boldsymbol{\beta}$. If we choose

$$\boldsymbol{c}' = [1 \quad X_{2,n+1} \quad X_{3,n+1} \quad \cdots \quad X_{k,n+1}]$$

then
$$\mu = E(Y_{n+1})$$

which is the *expected value* of the dependent variable Y in period $n + 1$, conditional on the X values in that period. In general, μ represents any linear combination of the elements of $\boldsymbol{\beta}$. We wish to consider the class of linear unbiased estimators of μ. Thus, we define a scalar m that will serve as a linear estimator of μ, that is,

$$m = \boldsymbol{a}'\boldsymbol{y} = \boldsymbol{a}'X\boldsymbol{\beta} + \boldsymbol{a}'\boldsymbol{u}$$

where $\boldsymbol{a}$ is some n-element column vector. The definition ensures linearity. To ensure unbiasedness we have

$$\begin{aligned} E(m) &= \boldsymbol{a}'X\boldsymbol{\beta} + \boldsymbol{a}'E(\boldsymbol{u}) \\ &= \boldsymbol{a}'X\boldsymbol{\beta} \\ &= \boldsymbol{c}'\boldsymbol{\beta} \end{aligned}$$

only if

$$\boldsymbol{a}'X = \boldsymbol{c}' \tag{3.27}$$

The problem is to find an n-vector $\boldsymbol{a}$ that will minimize the variance of m, subject to the k side conditions given by Eq. (3.27). The variance of m is seen to be

$$\text{var}(m) = E(\boldsymbol{a}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{a}) = \sigma^2\boldsymbol{a}'\boldsymbol{a}$$

where the derivation uses the fact that, since $\boldsymbol{a}'\boldsymbol{u}$ is a scalar, its square may be written as the product of its transpose and itself. The problem is then to find $\boldsymbol{a}$ to minimize $\boldsymbol{a}'\boldsymbol{a}$ subject to $X'\boldsymbol{a} = \boldsymbol{c}$. The solution is[16]

$$\boldsymbol{a} = X(X'X)^{-1}\boldsymbol{c}$$

which yields

$$\begin{aligned} m &= \boldsymbol{a}'\boldsymbol{y} \\ &= \boldsymbol{c}'(X'X)^{-1}X'\boldsymbol{y} \\ &= \boldsymbol{c}'\boldsymbol{b} \end{aligned}$$

This result specifically means the following:

1. Each LS coefficient b_i is a best linear unbiased estimator of the corresponding population parameter β_i.
2. The best linear unbiased estimate (BLUE) of any linear combination of β's is that same linear combination of the b's.
3. The BLUE of $E(Y_s)$ is

$$\hat{Y}_s = b_1 + b_2X_{2s} + b_3X_{3s} + \cdots + b_kX_{ks}$$

which is the value found by inserting a relevant vector of X values into the regression equation.

3.4.5 Testing Linear Hypotheses about $\boldsymbol{\beta}$

We have established the properties of the LS estimator of $\boldsymbol{\beta}$. It remains to show how to use this estimator to test various hypotheses about $\boldsymbol{\beta}$. Consider the following examples of typical hypotheses.

[16] See Appendix 3.3.

(*i*) H_0: $\beta_i = 0$. This sets up the hypothesis that the regressor X_i has no influence on Y. This type of test is very common and is often referred to simply as a *significance test.*

(*ii*) H_0: $\beta_i = \beta_{i0}$. Here β_{i0} is some specified value. If, for instance, β_i denotes a price elasticity one might wish to test that the elasticity is -1.

(*iii*) H_0: $\beta_2 + \beta_3 = 1$. If the β's indicate labor and capital elasticities in a production function, this formulation hypothesizes constant returns to scale.

(*iv*) H_0: $\beta_3 = \beta_4$, or $\beta_3 - \beta_4 = 0$. This hypothesizes that X_3 and X_4 have the same coefficient.

(*v*) H_0:

$$\begin{bmatrix} \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

This sets up the hypothesis that the complete set of regressors has no effect on Y. It tests the significance of the overall relation. The intercept term does not enter into the hypothesis, since interest centers on the variation of Y around its mean and the level of the series is usually of no specific relevance.

(*vi*) H_0: $\boldsymbol{\beta}_2 = \mathbf{0}$. Here the $\boldsymbol{\beta}$ vector is partitioned into two subvectors, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, containing, respectively, k_1 and $k_2(= k - k_1)$ elements. This sets up the hypothesis that a specified subset of regressors plays no role in the determination of Y.

All six examples fit into the general linear framework

$$\boldsymbol{R\beta} = \boldsymbol{r} \tag{3.28}$$

where $\boldsymbol{R}$ is a $q \times k$ matrix of known constants, with $q < k$, and $\boldsymbol{r}$ is a q-vector of known constants. Each null hypothesis determines the relevant elements in $\boldsymbol{R}$ and $\boldsymbol{r}$. For the foregoing examples we have

(*i*)	$\boldsymbol{R} = [0 \ \cdots \ 0 \ 1 \ 0 \ \cdots \ 0]$ with 1 in the ith position.	$\boldsymbol{r} = 0$	$q = 1$
(*ii*)	$\boldsymbol{R} = [0 \ \cdots \ 0 \ 1 \ 0 \ \cdots \ 0]$ with 1 in the ith position.	$\boldsymbol{r} = \beta_{i0}$	$q = 1$
(*iii*)	$\boldsymbol{R} = [0 \ 1 \ 1 \ 0 \ \cdots \ 0]$	$\boldsymbol{r} = 1$	$q = 1$
(*iv*)	$\boldsymbol{R} = [0 \ 0 \ 1 \ -1 \ 0 \ \cdots \ 0]$	$\boldsymbol{r} = 0$	$q = 1$
(*v*)	$\boldsymbol{R} = [\mathbf{0} \ \ \boldsymbol{I}_{k-1}]$ where $\mathbf{0}$ is a vector of $k-1$ zeros.	$\boldsymbol{r} = \mathbf{0}$	$q = k - 1$
(*vi*)	$\boldsymbol{R} = [\mathbf{0}_{k_2 \times k_1} \ \ \boldsymbol{I}_{k_2}]$	$\boldsymbol{r} = \mathbf{0}$	$q = k_2$

The efficient way to proceed is to derive a testing procedure for the general linear hypothesis

$$H_0\text{: } \boldsymbol{R\beta} - \boldsymbol{r} = \mathbf{0}$$

The general test may then be specialized to deal with any specific application. Given the LS estimator, an obvious step is to compute the vector $(\boldsymbol{Rb} - \boldsymbol{r})$. This vector measures the discrepancy between expectation and observation. If this vector is, in some sense, "large," it casts doubt on the null hypothesis, and conversely, if it is "small" it tends not to contradict the null. As in all conventional testing procedures, the distinction between large and small is determined from the relevant sampling distribution under the null, in this case, the distribution of $\boldsymbol{Rb}$ when $\boldsymbol{R\beta} = \boldsymbol{r}$.

From Eq. (3.24) it follows directly that

$$E(\boldsymbol{Rb}) = \boldsymbol{R\beta} \tag{3.29}$$

Therefore, from Eq. (3.25)

$$\begin{aligned}\operatorname{var}(\boldsymbol{Rb}) &= E[\boldsymbol{R}(\boldsymbol{b}-\boldsymbol{\beta})(\boldsymbol{b}-\boldsymbol{\beta})'\boldsymbol{R}'] \\ &= \boldsymbol{R}\operatorname{var}(\boldsymbol{b})\boldsymbol{R}' \\ &= \sigma^2\boldsymbol{R}(X'X)^{-1}\boldsymbol{R}'\end{aligned} \tag{3.30}$$

We thus know the mean and variance of the $\boldsymbol{Rb}$ vector. One further assumption is required to determine the form of the sampling distribution. Since $\boldsymbol{b}$ is a function of the $\boldsymbol{u}$ vector, the sampling distribution of $\boldsymbol{Rb}$ will be determined by the distribution of $\boldsymbol{u}$. The assumptions made so far about $\boldsymbol{u}$ are given in Eqs. (3.21) and (3.22). By making the additional assumption that the u_i are normally distributed, all three assumptions may be combined in the single statement

$$\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{3.31}$$

Since linear combinations of normal variables are also normally distributed, it follows directly that

$$\boldsymbol{b} \sim N[\boldsymbol{\beta}, \sigma^2(X'X)^{-1}] \tag{3.32}$$

Then

$$\boldsymbol{Rb} \sim N[\boldsymbol{R\beta}, \sigma^2\boldsymbol{R}(X'X)^{-1}\boldsymbol{R}'] \tag{3.33}$$

and so

$$\boldsymbol{R}(\boldsymbol{b}-\boldsymbol{\beta}) \sim N[\boldsymbol{0}, \sigma^2\boldsymbol{R}(X'X)^{-1}\boldsymbol{R}'] \tag{3.34}$$

If the null hypothesis $\boldsymbol{R\beta} = \boldsymbol{r}$ is true, then

$$(\boldsymbol{Rb}-\boldsymbol{r}) \sim N[\boldsymbol{0}, \sigma^2\boldsymbol{R}(X'X)^{-1}\boldsymbol{R}'] \tag{3.35}$$

This equation gives us the sampling distribution of $\boldsymbol{Rb}$; and, as seen in Appendix B, we may then derive a χ^2 variable, namely,

$$(\boldsymbol{Rb}-\boldsymbol{r})'[\sigma^2\boldsymbol{R}(X'X)^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{Rb}-\boldsymbol{r}) \sim \chi^2(q) \tag{3.36}$$

The only problem hindering practical application of Eq. (3.36) is the presence of the unknown σ^2. However, it is shown in Appendix B that

$$\frac{\boldsymbol{e}'\boldsymbol{e}}{\sigma^2} \sim \chi^2(n-k) \tag{3.37}$$

and that this statistic is distributed independently of $\boldsymbol{b}$. Thus, Eqs. (3.36) and (3.37) may be combined to form a computable statistic, which has an F distribution under the null hypothesis, namely,

$$\frac{(\boldsymbol{Rb}-\boldsymbol{r})'[\boldsymbol{R}(X'X)^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{Rb}-\boldsymbol{r})/q}{\boldsymbol{e}'\boldsymbol{e}/(n-k)} \sim F(q, n-k) \tag{3.38}$$

The test procedure is then to reject the hypothesis $\boldsymbol{R\beta} = \boldsymbol{r}$ if the computed F value exceeds a preselected critical value. Now we must see what this test procedure amounts to in the specific applications indicated previously.

For some cases it is helpful to rewrite Eq. (3.38) as

$$(\boldsymbol{Rb} - \boldsymbol{r})'[s^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{Rb} - \boldsymbol{r})/q \sim F(q, n-k) \tag{3.39}$$

where s^2 was defined in Eq. (3.26). Thus, $s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is the *estimated* variance-covariance matrix of $\boldsymbol{b}$. If we let c_{ij} denote the i, jth element in $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ then

$$s^2c_{ii} = \text{var}(b_i) \qquad \text{and} \qquad s^2c_{ij} = \text{cov}(b_i, b_j) \qquad i, j = 1, 2, \ldots, k$$

In each application the specific forms of $\boldsymbol{R}$ and $\boldsymbol{r}$ are substituted in Eq. (3.38) or (3.39).

(*i*) H_0: $\beta_i = 0$. $\boldsymbol{Rb}$ picks out b_i and $\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'$ picks out c_{ii}, the ith diagonal element in $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Thus Eq. (3.39) becomes

$$F = \frac{b_i^2}{s^2c_{ii}} = \frac{b_i^2}{\text{var}(b_i)} \sim F(1, n-k)$$

or, taking the square root,

$$t = \frac{b_i}{s\sqrt{c_{ii}}} = \frac{b_i}{\text{s.e.}(b_i)} \sim t(n-k)$$

Thus the null hypothesis that X_i has no association with Y is tested by dividing the ith estimated coefficient by its estimated standard error and referring the ratio to the t distribution.

(*ii*) H_0: $\beta_i = \beta_{i0}$. In a similar fashion this hypothesis is tested by

$$t = \frac{b_i - \beta_{i0}}{\text{s.e.}(b_i)} \sim t(n-k)$$

Instead of testing specific hypotheses about β_i one may compute, say, a 95 percent confidence interval for β_i. It is given by

$$b_i \pm t_{0.025}\text{s.e.}(b_i)$$

(*iii*) H_0: $\beta_2 + \beta_3 = 1$. $\boldsymbol{Rb}$ gives the sum of the two estimated coefficients, $b_2 + b_3$. Premultiplying $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ by $\boldsymbol{R}$ gives a row vector whose elements are the sum of the corresponding elements in the second and third rows of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Forming the inner product with $\boldsymbol{R}'$ gives the sum of the second and third elements of the row vector, that is, $c_{22} + 2c_{23} + c_{33}$, noting that $c_{23} = c_{32}$. Thus,

$$\begin{aligned} \boldsymbol{R}s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' &= s^2(c_{22} + 2c_{23} + c_{33}) \\ &= \text{var}(b_2) + 2\text{cov}(b_2, b_3) + \text{var}(b_3) \\ &= \text{var}(b_2 + b_3) \end{aligned}$$

The test statistic is then

$$t = \frac{(b_2 + b_3 - 1)}{\sqrt{\text{var}(b_2 + b_3)}} \sim t(n-k)$$

Alternatively one may compute, say, a 95 percent confidence interval for the sum $(\beta_2 + \beta_3)$ as

$$(b_2 + b_3) \pm t_{0.025}\sqrt{\text{var}(b_2 + b_3)}$$

(*iv*) $H_0\colon \beta_3 = \beta_4$. In a similar fashion to the derivation in (*iii*), the test statistic here is

$$t = \frac{b_3 - b_4}{\sqrt{\text{var}(b_3 - b_4)}} \sim t(n-k)$$

(*v*) $H_0\colon \beta_2 = \beta_3 = \cdots = \beta_k = 0$. The first four examples have each involved just a single hypothesis, which is why we had a choice of equivalent F and t tests. This example involves a composite hypothesis about all $k-1$ regressor coefficients. Now $\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'$ picks out the square submatrix of order $k-1$ in the bottom right-hand corner of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. To evaluate this submatrix, partition the $\boldsymbol{X}$ matrix as $[\boldsymbol{i} \quad X_2]$ where X_2 is the matrix of observations on all $k-1$ regressors. Then

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} \boldsymbol{i}' \\ X_2' \end{bmatrix} [\boldsymbol{i} \quad X_2] = \begin{bmatrix} n & \boldsymbol{i}'X_2 \\ X_2'\boldsymbol{i} & X_2'X_2 \end{bmatrix}$$

From the formula in Appendix A for the inverse of a partitioned matrix, we can then express the required submatrix as

$$[X_2'X_2 - X_2'\boldsymbol{i}n^{-1}\boldsymbol{i}'X_2]^{-1} = [X_2'AX_2]^{-1} = [X_*'X_*]^{-1}$$

where $\boldsymbol{A}$ is the matrix, already defined in Eq. (3.7), which transforms observations into deviation form, and $X_* = AX_2$. $\boldsymbol{Rb} = \boldsymbol{b}_2$, which is the vector of LS estimates of the coefficients of the $k-1$ regressors. Apart from the divisor q, the numerator in Eq. (3.38) is then $\boldsymbol{b}_2'X_*'X_*\boldsymbol{b}_2$, which has already been shown in Eq. (3.9) to be the ESS from the regression. Thus the F statistic for testing the joint significance of the complete set of regressors is

$$F = \frac{\text{ESS}/(k-1)}{\text{RSS}/(n-k)} \sim F(k-1, n-k) \tag{3.40}$$

By using Eq. (3.10), this statistic may also be expressed as

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F(k-1, n-k) \tag{3.41}$$

The test essentially asks whether the mean square due to the regression is significantly larger than the residual mean square.

(*vi*) $H_0\colon \boldsymbol{\beta}_2 = \boldsymbol{0}$. This hypothesis postulates that a *subset* of regressor coefficients is a zero vector, in contrast with the previous example, where *all* regressor coefficients were hypothesized to be zero. Partition the regression equation as follows:

$$\boldsymbol{y} = [X_1 \quad X_2]\begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix} + \boldsymbol{e} = X_1\boldsymbol{b}_1 + X_2\boldsymbol{b}_2 + \boldsymbol{e}$$

where X_1 has k_1 columns, including a column of ones, X_2 has k_2 $(= k - k_1)$ columns, and $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are the corresponding subvectors of regression coefficients. The partitioning of the X matrix gives

$$X'X = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}$$

$R(X'X)^{-1}R'$ picks out the square matrix of order k_2 in the bottom right-hand corner of $(X'X)^{-1}$. In a similar fashion to Example (v), this may be shown to be $(X_2'M_1X_2)^{-1}$ where $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$. This is a symmetric, idempotent matrix of the type already defined in Eq. (3.17). It also has the properties that $M_1X_1 = \mathbf{0}$ and $M_1\boldsymbol{e} = \boldsymbol{e}$. Further, $M_1\boldsymbol{y}$ gives the vector of residuals when $\boldsymbol{y}$ is regressed on X_1. The numerator in Eq. (3.38) is then

$$\boldsymbol{b}_2'(X_2'M_1X_2)\boldsymbol{b}_2/k_2$$

To appreciate what is measured by this numerator, premultiply the partitioned regression equation by M_1 to obtain

$$M_1\boldsymbol{y} = M_1X_2\boldsymbol{b}_2 + \boldsymbol{e}$$

Squaring both sides gives

$$\boldsymbol{y}'M_1\boldsymbol{y} = \boldsymbol{b}_2'(X_2'M_1X_2)\boldsymbol{b}_2 + \boldsymbol{e}'\boldsymbol{e}$$

The term on the left of this equation is the RSS when $\boldsymbol{y}$ is regressed just on X_1. The last term, $\boldsymbol{e}'\boldsymbol{e}$, is the RSS when $\boldsymbol{y}$ is regressed on $[X_1 \quad X_2]$. Thus the middle term measures the *increment* in ESS (or equivalently, the *reduction* in RSS) when X_2 is added to the set of regressors. The hypothesis may thus be tested by running two separate regressions. First regress $\boldsymbol{y}$ on X_1 and denote the RSS by $\boldsymbol{e}_*'\boldsymbol{e}_*$. Then run the regression on all the Xs, obtaining the RSS, denoted as usual by $\boldsymbol{e}'\boldsymbol{e}$. From Eq. (3.38) the test statistic is

$$F = \frac{(\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e})/k_2}{\boldsymbol{e}'\boldsymbol{e}/(n-k)} \sim F(k_2, n-k) \tag{3.42}$$

3.4.6 Restricted and Unrestricted Regressions

Clearly, (v) is a special case of (vi), and so (v) may also be interpreted as the outcome of two separate regressions. Recall from Eq. (3.9) that ESS may be expressed as ESS $= \boldsymbol{y}_*'\boldsymbol{y}_* - \boldsymbol{e}'\boldsymbol{e}$, where $\boldsymbol{y}_* = A\boldsymbol{y}$ with A defined in Eq. (3.7). It may be shown that $\boldsymbol{y}_*'\boldsymbol{y}_*$ is the RSS when $\boldsymbol{y}_*$ is regressed on $\boldsymbol{x}_1$ $(= \boldsymbol{i})$.[17] With this substitution for ESS in Eq. (3.40), that test statistic has the same form as Eq. (3.42).

In both cases (v) and (vi) the first regression may be regarded as a ***restricted regression*** and the second as an ***unrestricted regression.*** Likewise $\boldsymbol{e}_*'\boldsymbol{e}_*$ is the *restricted RSS* and $\boldsymbol{e}'\boldsymbol{e}$ is the *unrestricted RSS.* In the restricted regression the restrictions in H_0

[17] See Problem 3.5.

are actually imposed on the *estimated* equation. Thus, the restricted regression in (v) omits $X_2, X_3, \ldots, X_k$ from the regression, or equivalently, $b_2, b_3, \ldots, b_k$ are set to zero. In (vi) the restricted regression uses only the variables in $\boldsymbol{X}_1$. The unrestricted regression in each case uses all the variables in the $\boldsymbol{X}$ matrix. By the same argument, Example (i) is also a special case of (vi). In the restricted regression all variables except X_i are used. Thus the significance test for β_i asks whether there is a significant reduction in RSS (increase in ESS) upon adding X_i to the set of regressors.

Students sometimes have difficulty in determining the correct value for q in these tests. It may be calculated in several equivalent ways:

1. The number of rows in the $\boldsymbol{R}$ matrix
2. The number of elements in the null hypothesis
3. The difference between the number of β coefficients estimated in the unrestricted and restricted equations
4. The difference in the degrees of freedom attaching to $\boldsymbol{e}_*'\boldsymbol{e}_*$ and $\boldsymbol{e}'\boldsymbol{e}$

In all six examples, test statistics have been derived involving the b_i coefficients from the unrestricted regression. However, we have seen that in Examples (i), (v), and (vi) the test statistics may also be expressed in terms of the difference between the restricted and unrestricted RSS. In all three cases the restricted regression was easily obtained by excluding the relevant variables from the regression. The question naturally arises as to whether the tests in Examples (ii), (iii), and (iv) have a similar interpretation in terms of the difference between two residual sums of squares. This requires an examination of how to fit the restricted regression in these cases.

3.4.7 Fitting the Restricted Regression

This may be done in two ways. One is to work out each specific case from first principles; the other is to derive a general formula into which specific cases can then be fitted. As an example of the first approach consider Example (iii) with the regression in deviation form,

$$y = b_2x_2 + b_3x_3 + e$$

We wish to impose the restriction that $b_2 + b_3 = 1$. Substituting the restriction in the regression gives a reformulated equation as

$$y = b_2x_2 + (1 - b_2)x_3 + e_*$$

or $$(y - x_3) = b_2(x_2 - x_3) + e_*$$

Thus two new variables, $(y - x_3)$ and $(x_2 - x_3)$, are formed; and the simple regression of the first on the second (without an intercept term) gives the restricted estimate of b_2. The RSS from this regression is the restricted RSS, $\boldsymbol{e}_*'\boldsymbol{e}_*$.

The general approach requires a $\boldsymbol{b}_*$ vector that minimizes the RSS *subject to the restrictions* $\boldsymbol{Rb}_* = \boldsymbol{r}$. To do this we set up the function

$$\phi = (\boldsymbol{y} - \boldsymbol{Xb}_*)'(\boldsymbol{y} - \boldsymbol{Xb}_*) - 2\boldsymbol{\lambda}'(\boldsymbol{Rb}_* - \boldsymbol{r})$$

where $\boldsymbol{\lambda}$ is a q-vector of Lagrange multipliers. The first-order conditions are

$$\frac{\partial \phi}{\partial \boldsymbol{b}_*} = -2\boldsymbol{X}'\boldsymbol{y} + 2(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{b}_* - 2\boldsymbol{R}'\boldsymbol{\lambda} = \boldsymbol{0}$$

$$\frac{\partial \phi}{\partial \boldsymbol{\lambda}} = -2(\boldsymbol{R}\boldsymbol{b}_* - \boldsymbol{r}) = \boldsymbol{0}$$

The solution for $\boldsymbol{b}_*$ is[18]

$$\boldsymbol{b}_* = \boldsymbol{b} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{r} - \boldsymbol{R}\boldsymbol{b}) \tag{3.43}$$

where $\boldsymbol{b}$ is the usual, unrestricted LS estimator $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. The residuals from the restricted regression are

$$\begin{aligned} \boldsymbol{e}_* &= \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_* \\ &= \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} - \boldsymbol{X}(\boldsymbol{b}_* - \boldsymbol{b}) \\ &= \boldsymbol{e} - \boldsymbol{X}(\boldsymbol{b}_* - \boldsymbol{b}) \end{aligned}$$

Transposing and multiplying, we obtain

$$\boldsymbol{e}_*'\boldsymbol{e}_* = \boldsymbol{e}'\boldsymbol{e} + (\boldsymbol{b}_* - \boldsymbol{b})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b}_* - \boldsymbol{b})$$

The process of substituting for $(\boldsymbol{b}_* - \boldsymbol{b})$ from Eq. (3.43) and simplifying gives

$$\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{r} - \boldsymbol{R}\boldsymbol{b})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{r} - \boldsymbol{R}\boldsymbol{b})$$

where, apart from q, the expression on the right-hand side is the same as the numerator in the F statistic in Eq. (3.38). Thus an alternative expression of the test statistic for H_0: $\boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$ is

$$F = \frac{(\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e})/q}{\boldsymbol{e}'\boldsymbol{e}/(n-k)} \sim F(q, n-k) \tag{3.44}$$

Consequently all six examples fit into the same framework.

In summary, tests of $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$ may be implemented by fitting the unrestricted regression and substituting the resultant $\boldsymbol{b}$ vector in Eq. (3.38). Alternatively, a restricted regression may be fitted as well, and tests based on the difference $(\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e})$ between the restricted and unrestricted RSS, as in Eq. (3.44). The following numerical examples illustrate these procedures.

EXAMPLE 3.7. We will continue with the data of Example 3.3.

(*i*) H_0: $\beta_3 = 0$. The appropriate test statistic is $t = b_3/s\sqrt{c_{33}}$, where c_{33} is the bottom right-hand element in $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. From the results already obtained, $b_3 = -1.5$, and $s = \sqrt{\text{RSS}/(n-k)} = \sqrt{1.5/2} = \sqrt{0.75}$. The term c_{33} may be obtained by calculating the determinant of the 3×3 matrix $(\boldsymbol{X}'\boldsymbol{X})$ directly and dividing the relevant cofactor by this determinant. Evaluating the determinant directly is tedious. Since adding multiples of rows (columns) to a row (column) of a matrix does not alter the determinant, it is simpler to find the determinant of the echelon matrix already obtained in the Gaussian elimination in Example 3.3, namely,

[18] See Appendix 3.4.

$$|X'X| = \begin{vmatrix} 5 & 15 & 25 \\ 0 & 10 & 6 \\ 0 & 0 & 0.4 \end{vmatrix} = 20$$

where the determinant is simply the product of the elements on the principal diagonal. The relevant cofactor is

$$\begin{vmatrix} 5 & 15 \\ 15 & 55 \end{vmatrix} = 50$$

Thus $c_{33} = 50/20 = 2.5$ and

$$t = \frac{-1.5}{\sqrt{0.75}\sqrt{2.5}} = -\sqrt{1.2} = -1.095$$

which falls well short of any conventional critical value.

When a hypothesis does not involve the intercept term, it is often simpler to work with the data in deviation form, since doing so reduces the dimensionality of the problem. In this case c_{33} will be the bottom right-hand element in the inverse of $(X_*'X_*)$. Referring to Example 3.3 gives

$$[X_*'X_*]^{-1} = \begin{bmatrix} 10 & 6 \\ 6 & 4 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -1.5 \\ -1.5 & 2.5 \end{bmatrix}$$

which is the same result as before.

Alternatively, one may examine the problem in terms of the reduction in RSS when X_3 is added to the set of regressors. From Table 3.2, $\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e} = 0.9$. Thus,

$$F = \frac{\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{e}'\boldsymbol{e}/(n-k)} = \frac{0.9}{1.5/2} = 1.2$$

and $t = \sqrt{F} = \sqrt{1.2}$ as before.

(*ii*) H_0: $\beta_3 = -1$. The relevant test statistic is now

$$t = \frac{-1.5-(-1)}{s\sqrt{c_{33}}} = \frac{-0.5}{\sqrt{0.75}\sqrt{2.5}} = -0.365$$

which is insignificant. Alternatively one might compute, say, a 95 percent confidence interval for β_3. From the t distribution, $t_{0.025}(2) = 4.303$. The interval is then

$$b_3 \pm t_{0.025}\text{s.e.}(b_3)$$

which is $\quad -1.5 \pm 4.303\sqrt{0.75}\sqrt{2.5}$

that is $\quad -7.39 \quad$ to $\quad 4.39$

The wide interval confirms the nonrejection of hypotheses (*i*) and (*ii*). The example is, of course, extremely simplistic, and the small sample size does not permit any sharp inferences.

(*iii*) H_0: $\beta_2 + \beta_3 = 0$. From $(X_*'X_*)^{-1}$ in case (*i*)

$$\text{var}(b_2 + b_3) = 0.75[1 + 2.5 - 2(1.5)] = 0.375$$

giving $\quad t = \dfrac{1}{\sqrt{0.375}} = 1.63$

which is insignificant.

(*iv*) H_0: $\beta_2 = \beta_3 = 0$. Notice carefully the distinction between this and the previous hypothesis. Substitution in Eq. (3.40) gives

$$F = \frac{26.5/2}{1.5/2} = 17.67$$

However, $F_{0.05}(2, 2) = 19.00$, so the hypothesis would not be rejected at the 5 percent level, despite the impressive R^2.

3.5 PREDICTION

Suppose that we have fitted a regression equation, and we now consider some specific vector of regressor values,

$$\boldsymbol{c}' = [1 \quad X_{2f} \quad \cdots \quad X_{kf}]$$

The Xs may be hypothetical if an investigator is exploring possible effects of different scenarios, or they may be newly observed values. In either case we wish to predict the value of *Y conditional on* $\boldsymbol{c}$. Any such prediction is based on the assumption that the fitted model still holds in the prediction period. When a new value Y_f is also observed it is possible to test this stability assumption. An appealing **point prediction** is obtained by inserting the given X values into the regression equation, giving

$$\hat{Y}_f = b_1 + b_2X_{2f} + \cdots + b_kX_{kf} = \boldsymbol{c}'\boldsymbol{b} \tag{3.45}$$

In the discussion of the Gauss–Markov theorem it was shown that $\boldsymbol{c}'\boldsymbol{b}$ is a best linear unbiased estimator of $\boldsymbol{c}'\boldsymbol{\beta}$. In the present context $\boldsymbol{c}'\boldsymbol{\beta} = E(Y_f)$. Thus $\hat{Y}_f$ is an optimal predictor of $E(Y_f)$. Moreover, it was shown in Eq. (3.30) that $\text{var}(\boldsymbol{Rb}) = \boldsymbol{R}\text{var}(\boldsymbol{b})\boldsymbol{R}'$. Replacing $\boldsymbol{R}$ by $\boldsymbol{c}'$ gives

$$\text{var}(\boldsymbol{c}'\boldsymbol{b}) = \boldsymbol{c}'\,\text{var}(\boldsymbol{b})\boldsymbol{c}$$

If we assume normality for the disturbance term, it follows that

$$\frac{\boldsymbol{c}'\boldsymbol{b} - \boldsymbol{c}'\boldsymbol{\beta}}{\sqrt{\text{var}(\boldsymbol{c}'\boldsymbol{b})}} \sim N(0, 1)$$

When the unknown σ^2 in $\text{var}(\boldsymbol{b})$ is replaced by s^2, the usual shift to the t distribution occurs, giving

$$\frac{\hat{Y}_f - E(Y_f)}{s\sqrt{\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim t(n-k) \tag{3.46}$$

from which a 95 percent confidence interval for $E(Y_f)$ is

$$\hat{Y}_f \pm t_{0.025}s\sqrt{\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}} \tag{3.47}$$

EXAMPLE 3.8. Let us continue with the data of Example 3.3. We wish to predict $E(Y)$ if $X_2 = 10$ and $X_3 = 10$. The point prediction is

$$\hat{Y}_f = 4 + 2.5(10) - 1.5(10) = 14$$

Inverting $(\boldsymbol{X}'\boldsymbol{X})$ gives $(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 26.7 & 4.5 & -8.0 \\ 4.5 & 1.0 & -1.5 \\ -8.0 & -1.5 & 2.5 \end{bmatrix}$

Thus, $$c'(X'X)^{-1}c = [1 \quad 10 \quad 10]\begin{bmatrix} 26.7 & 4.5 & -8.0 \\ 4.5 & 1.0 & -1.5 \\ -8.0 & -1.5 & 2.5 \end{bmatrix}\begin{bmatrix} 1 \\ 10 \\ 10 \end{bmatrix} = 6.7$$

and $s^2 = 0.75$ as before. Thus the 95 percent confidence interval for $E(Y_f)$ is

$$14 \pm 4.303\sqrt{0.75}\sqrt{6.7}$$

or $$4.35 \quad \text{to} \quad 23.65$$

When separated from a PC one prefers not to have to invert 3 × 3 matrices. Example 3.8 may be reworked in terms of deviations, which lowers the dimensionality by one.[19] Sometimes one wishes to obtain a confidence interval for Y_f rather than $E(Y_f)$. The two differ only by the disturbance u_f that happens to appear in the prediction period. This is unpredictable, so the point prediction remains as in Eq. (3.45). It is still a best linear unbiased predictor, but the uncertainty of the prediction is increased. We have $\hat{Y}_f = c'b$ as before, and now $Y_f = c'\beta + u_f$. The prediction error is thus

$$e_f = Y_f - \hat{Y}_f = u_f - c'(b - \beta)$$

The process of squaring both sides and taking expectations gives the variance of the prediction error as

$$\begin{aligned} \text{var}(e_f) &= \sigma^2 + c'\text{var}(b)c \\ &= \sigma^2(1 + c'(X'X)^{-1}c) \end{aligned}$$

from which we derive a t statistic

$$\frac{\hat{Y}_f - Y_f}{s\sqrt{1 + c'(X'X)^{-1}c}} \sim t(n-k) \tag{3.48}$$

Comparison with Eq. (3.46) shows that the only difference is an increase of one in the square root term. Thus the revised confidence interval for the data in Example 3.8 is

$$14 \pm 4.303\sqrt{0.75}\sqrt{7.7}$$

or $$3.66 \quad \text{to} \quad 24.34$$

APPENDIX

APPENDIX 3.1
To prove $r_{12.3} = (r_{12} - r_{13}r_{23})/\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}$

Recall from Chapter 1 that for a two-variable regression, $y = bx + e$, in deviation form, $b = rs_y/s_x$ and $\sum e^2 = \sum y^2(1 - r^2) = ns_y^2(1 - r^2)$, where s denotes the

[19] See Problem 3.6.

sample standard deviation of the subscripted variable. Thus,

$$\sum e_{1.3}^2 = ns_1^2(1 - r_{13}^2) \quad \text{and} \quad \sum e_{2.3}^2 = ns_2^2(1 - r_{23}^2)$$

Also $$e_{1.3} = y - r_{13}\frac{s_1}{s_3}x_3 \quad \text{and} \quad e_{2.3} = x_2 - r_{23}\frac{s_2}{s_3}x_3$$

After some simplification, we obtain

$$\sum e_{1.3}e_{2.3} = ns_1s_2(r_{12} - r_{13}r_{23})$$

and so $$r_{12.3} = \frac{\sum e_{1.3}e_{2.3}}{\sqrt{\sum e_{1.3}^2}\sqrt{\sum e_{2.3}^2}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

APPENDIX 3.2
Solving for a single regression coefficient in a multiple regression.

The normal equations are

$$\begin{bmatrix} x_2'x_2 & x_2'X_* \\ X_*'x_2 & X_*'X_* \end{bmatrix}\begin{bmatrix} b_2 \\ b_{(2)} \end{bmatrix} = \begin{bmatrix} x_2'y \\ X_*'y \end{bmatrix}$$

where x_2 is the n-vector of observations on X_2 and X_* is the $n \times (k - 1)$ matrix of observations on all the other right-hand side variables, including a column of ones. The scalar b_2 is the LS coefficient of X_2, and $b_{(2)}$ denotes the coefficients of the remaining $k - 1$ variables. Inverting the matrix in the normal equations gives b_2 as

$$b_2 = c_{11}(x_2'y) + c_{12}(X_*'y)$$

where c_{11} is the first element in the top row of the inverse matrix and c_{12} contains the remaining $k - 1$ elements of the first row. From the formulae for the inverse of a partitioned matrix

$$c_{11} = (x_2'x_2 - x_2'X_*(X_*'X_*)^{-1}X_*'x_2)^{-1} = (x_2'M_*x_2)^{-1}$$

where $$M_* = I - X_*(X_*'X_*)^{-1}X_*'$$

Also $$c_{12} = -(x_2'M_*x_2)^{-1}x_2'X_*(X_*'X_*)^{-1}$$

Thus, $$\begin{aligned} b_2 &= (x_2'M_*x_2)^{-1}x_2'y - (x_2'M_*x_2)^{-1}x_2'X_*(X_*'X_*)^{-1}X_*'y \\ &= (x_2'M_*x_2)^{-1}x_2'M_*y \end{aligned}$$

Notice that this coefficient has two possible interpretations. The vector M_*x_2 denotes the residuals in x_2 when the linear influence of all the variables in X_* has been removed. Similarly M_*y denotes the vector of residuals in y when X_* has been allowed for. The formula then shows that b_2 is the slope in the regression of the latter residuals on the former. However, the idempotency of M_* ensures that the same coefficient would also be obtained by regressing y on M_*x_2.

This is an example of a general result, the **Frisch–Waugh–Lovell Theorem.**[20] Suppose the regressors are partitioned into two submatrices, $X = [X_1 \quad X_2]$. The regression may then be written

$$y = [X_1 \quad X_2]\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + e = X_1 b_1 + X_2 b_2 + e \tag{A3.1}$$

Premultiply this equation by $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$, where M_1 is the type of symmetric, idempotent matrix developed in Eq. (3.17). This multiplication gives

$$M_1 y = M_1 X_2 b_2 + e \tag{A3.2}$$

since $M_1 X_1 = 0$ and $M_1 e = e$. Premultiplying this equation by X_2' yields

$$X_2' M_1 y = (X_2' M_1 X_2) b_2$$

or, equivalently, $$(M_1X_2)'(M_1y) = (M_1X_2)'(M_1X_2)b_2 \tag{A3.3}$$

This last equation shows that the subvector b_2 from the regression in Eq. (A3.1) may also be obtained from the regression of $(M_1 y)$ on $(M_1 X_2)$. From the properties of M_1 it follows that

$M_1 y$ = vector of residuals when y is regressed on X_1

$M_1 X_2$ = matrix of residuals when each variable in X_2 is regressed on X_1

A comparison of Eqs. (A3.1) and (A3.2) shows that the residual vector is the same in each regression. Thus, the RSS may be obtained from either regression.

In a similar fashion one may define $M_2 = I - X_2(X_2'X_2)^{-1}X_2'$ and obtain the regression equation $M_2 y = M_2 X_1 b_1 + e$. In this case y and X_1 have been adjusted for the linear influence of the variables in X_2, and the coefficient vector from the regression of the residuals coincides with the b_1 subvector in Eq. (A3.1).

Earlier results, which have been separately derived from first principles, follow simply from the general theorem. If $X_1 = i$, a column of ones, and X_2 is the $n \times (k - 1)$ matrix of regressors, then

$$M_1 = I - i(i'i)^{-1}i' = I - \frac{1}{n}(ii') = A$$

where A is the deviation-producing matrix defined in Eq. (3.7). Thus the slope coefficients may be obtained from the regression using the variables in deviation form, and this regression also yields the same RSS as the regression with an intercept and the regressors in raw form. If X_1 represents time, the coefficients on the other regressors may be obtained by first removing a linear trend from all variables and regressing the trend-corrected variables, or by using the regressors in raw form and including

[20] R. Frisch and F. V. Waugh, "Partial Time Regressions as Compared with Individual Trends," *Econometrica,* 1933, **1**, 387–401; and M. C. Lovell, "Seasonal Adjustment of Economic Time Series," *Journal of the American Statistical Association, 1963,* **58**, 993–1010. The theorem is discussed and applied extensively in Russell Davidson and James MacKinnon, *Estimation and Inference in Econometrics,* Oxford University Press, 1993.

time explicitly as one of the regressors. Finally, the general theorem may be used to illustrate the sequential buildup of ESS or, equivalently, the sequential reduction in RSS as more variables are added to the regression. Squaring Eq. (A3.2) gives

$$\boldsymbol{y}'\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{b}_2'(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2)\boldsymbol{b}_2 + \boldsymbol{e}'\boldsymbol{e}$$

Now $\boldsymbol{y}'\boldsymbol{M}_1\boldsymbol{y} = (\boldsymbol{M}_1\boldsymbol{y})'(\boldsymbol{M}_1\boldsymbol{y})$ = RSS from the regression of $\boldsymbol{y}$ on X_1, and $\boldsymbol{e}'\boldsymbol{e}$ is the RSS from the regression of $\boldsymbol{y}$ on $[X_1 \quad X_2]$. Thus $\boldsymbol{b}_2'(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2)\boldsymbol{b}_2$ measures the reduction in RSS due to the addition of $\boldsymbol{X}_2$ to the set of regressors.

APPENDIX 3.3
To show that minimizing $a'a$ subject to $X'a = c$ gives $a = X(X'X)^{-1}c$.

In this problem $\boldsymbol{a}$ is an unknown n-vector, $\boldsymbol{c}$ is a known k-vector, and $\boldsymbol{X}$ is a known $n \times k$ matrix. Let

$$\phi = \boldsymbol{a}'\boldsymbol{a} - 2\boldsymbol{\lambda}'(X'\boldsymbol{a} - \boldsymbol{c})$$

The partial derivatives are

$$\frac{\partial\phi}{\partial \boldsymbol{a}} = 2\boldsymbol{a} - 2X\boldsymbol{\lambda} \qquad \text{and} \qquad \frac{\partial\phi}{\partial \boldsymbol{\lambda}} = -2(\boldsymbol{X}'\boldsymbol{a} - \boldsymbol{c})$$

Equating these derivatives to zero and premultiplying the first equation by $\boldsymbol{X}'$ gives

$$\boldsymbol{\lambda} = (X'X)^{-1}X'\boldsymbol{a} = (X'X)^{-1}\boldsymbol{c}$$

Thus $$\boldsymbol{a} = X\boldsymbol{\lambda} = X(X'X)^{-1}\boldsymbol{c}$$

If a scalar m is defined as $m = \boldsymbol{a}'\boldsymbol{y}$, then $m = \boldsymbol{c}'\boldsymbol{b}$, which is a best linear unbiased estimator of $\boldsymbol{c}'\boldsymbol{\beta}$.

APPENDIX 3.4
Derivation of the restricted estimator b_*.

Define

$$\phi = (y - X\boldsymbol{b}_*)'(y - X\boldsymbol{b}_*) - 2\boldsymbol{\lambda}'(\boldsymbol{R}\boldsymbol{b}_* - \boldsymbol{r})$$

where $\boldsymbol{\lambda}$ is a q-vector of Lagrange multipliers. Equating the partial derivatives of ϕ to zero gives the equations

$$(X'X)\boldsymbol{b}_* = X'y + \boldsymbol{R}'\boldsymbol{\lambda}$$

$$\boldsymbol{R}\boldsymbol{b}_* = \boldsymbol{r}$$

The first equation yields

$$\boldsymbol{b}_* = \boldsymbol{b} + (X'X)^{-1}\boldsymbol{R}'\boldsymbol{\lambda}$$

where $\boldsymbol{b}$ is the unrestricted LS estimator. Premultiplying by $\boldsymbol{R}$ gives

$$Rb_* = Rb + \left[R(X'X)^{-1}R'\right]\lambda$$

whence $$\lambda = \left[R(X'X)^{-1}R'\right]^{-1}(r - Rb)$$

and so $$b_* = b + (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(r - Rb)$$

PROBLEMS

3.1. Show algebraically that in general the two equations in terms of deviations in Example 3.3 must be identical with the second and third equations obtained in the first step of Gaussian elimination applied to the normal equations in terms of raw data.

3.2. Derive Eq. (3.15) for $r_{13.2}$ from first principles.

3.3. Prove Eq. (3.16) that $R^2_{1.23} - r^2_{12} = r^2_{13.2}(1 - r^2_{12})$. Also prove that $R^2_{1.23} - r^2_{13} = r^2_{12.3}(1 - r^2_{13})$. Corollary: Show that if $r_{23} = 0$, then $R^2_{1.23} = r^2_{12} + r^2_{13}$.

3.4. Consider the hypothetical demand function $\ln Q = \beta_1 + \beta_2 \ln P + \beta_3 T$. Denote three of the simple regression coefficients by

$$b_{13} = \frac{\sum qt}{\sum t^2} \qquad b_{23} = \frac{\sum pt}{\sum t^2} \qquad b_{32} = \frac{\sum pt}{\sum p^2}$$

where q, p, and t denote $\ln Q$, $\ln P$, and T in deviation form. Prove that the LS estimate of the shift parameter β_3 in the multiple regression may be written

$$b_3 = b_{13.2} = \frac{b_{13} - b_{12}b_{23}}{1 - b_{23}b_{32}}$$

which, in general, is not equal to either of the coefficients of time, b_{13} and b_{23}, in the two-variable trend analyses. What happens when $b_{23} = 0$?

3.5. Prove that $y_*'y_*$, where $y_* = Ay$ with A defined in Eq. (3.7), is the RSS when y is regressed on $x_1 = i$. Show also that the estimated coefficient of the regression is $\bar{Y}$.

3.6. Consider an alternative setup for the regression equation where the right-hand side regressors are expressed in deviation form, that is, $X = [i\ X_*]$, where X_* is the $n \times (k-1)$ matrix of deviations. Show that when y is regressed on X the LS vector is

$$b = \begin{bmatrix} \bar{Y} \\ b_2 \end{bmatrix}$$

where b_2 is the $(k-1)$ element vector,

$$b_2 = (X_*'X_*)^{-1}X_*'y = (X_*'X_*)^{-1}X_*'y_*$$

A point prediction may then be obtained from

$$\hat{Y}_f = \bar{Y} + b_2 x_{2f} + \cdots + b_k x_{kf}$$

Show that

$$\text{var}(\hat{Y}_f) = s^2\left[\frac{1}{n} + x_*'(X_*'X_*)^{-1}x_*\right]$$

where $x_*' = [x_{2f} \cdots x_{kf}]$, and hence rework the prediction problem in Example 3.8.

3.7. Your research assistant reports the following results in several different regression problems. In which cases could you be certain that an error had been committed? Explain.

(*a*) $R^2_{1.23} = 0.89$ and $R^2_{1.234} = 0.86$

(*b*) $r^2_{12} = 0.227$, $r^2_{13} = 0.126$, and $R^2_{1.23} = 0.701$

(*c*) $(\sum x^2)(\sum y^2) - (\sum xy)^2 = -1732.86$

(University of Michigan, 1980)

3.8. Sometimes variables are *standardized* before the computation of regression coefficients. Standardization is achieved by dividing each observation on a variable by its standard deviation, so that the standard deviation of the transformed variable is unity. If the original relation is, say,

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

and the corresponding relation between the transformed variables is

$$Y^* = \beta_1^* + \beta_2^* X_2^* + \beta_3^* X_3^* + u^*$$

where $Y^* = Y/s_y$
$X_i^* = X_i/s_i \qquad i = 2, 3$

what is the relationship between β_2^*, β_3^* and β_2, β_3? Show that the partial correlation coefficients are unaffected by the transformation. The β^* coefficients are often referred to in the statistical literature as **beta coefficients.** They measure the effect of a one-standard-deviation change in a regressor on the dependent variable (also measured in standard deviation units).

3.9. Test each of the hypotheses $\beta_1 = 1, \beta_2 = 1, \beta_3 = -2$, in the regression model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

given the following sums of squares and products of deviations from means for 24 observations:

$$\sum y^2 = 60 \qquad \sum x_1^2 = 10 \qquad \sum x_2^2 = 30 \qquad \sum x_3^2 = 20$$

$$\sum yx_1 = 7 \qquad \sum yx_2 = -7 \qquad \sum yx_3 = -26$$

$$\sum x_1x_2 = 10 \qquad \sum x_1x_3 = 5 \qquad \sum x_2x_3 = 15$$

Test the hypothesis that $\beta_1 + \beta_2 + \beta_3 = 0$. How does this differ from the hypothesis that $[\beta_1 \quad \beta_2 \quad \beta_3] = [1 \quad 1 \quad -2]$? Test the latter hypothesis.

3.10. The following sums were obtained from 10 sets of observations on Y, X_1, and X_2:

$$\sum Y = 20 \qquad \sum X_1 = 30 \qquad \sum X_2 = 40$$

$$\sum Y^2 = 88.2 \qquad \sum X_1^2 = 92 \qquad \sum X_2^2 = 163$$

$$\sum YX_1 = 59 \qquad \sum YX_2 = 88 \qquad \sum X_1X_2 = 119$$

Estimate the regression of Y on X_1 and X_2, including an intercept term, and test the hypothesis that the coefficient of X_2 is zero.

3.11. The following regression equation is estimated as a production function for Q:

$$\ln Q = 1.37 + \underset{(0.257)}{0.632} \ln K + \underset{(0.219)}{0.452} \ln L$$

$$R^2 = 0.98 \qquad \operatorname{cov}(b_k, b_l) = 0.055$$

where the standard errors are given in parentheses. Test the following hypotheses:

(*a*) The capital and labor elasticities of output are identical.

(*b*) There are constant returns to scale.

(University of Washington, 1980)

Note: The problem does not give the number of sample observations. Does this omission affect your conclusions?

3.12. Consider a multiple regression model for which all classical assumptions hold, but in which there is *no constant term.* Suppose you wish to test the null hypothesis that there is no relationship between y and X, that is,

$$H_0: \beta_2 = \cdots = \beta_k = 0$$

against the alternative that at least one of the β's is nonzero. Present the appropriate test statistic and state its distributions (including the number[s] of degrees of freedom).

(University of Michigan, 1978)

3.13. One aspect of the rational expectations hypothesis involves the claim that expectations are unbiased, that is, that the average prediction is equal to the observed realization of the variable under investigation. This claim can be tested by reference to announced predictions and to actual values of the rate of interest on three-month U.S. Treasury Bills published in *The Goldsmith-Nagan Bond and Money Market Letter.* The results of least-squares estimation (based on 30 quarterly observations) of the regression of the actual on the predicted interest rates were as follows:

$$r_t = \underset{(0.86)}{0.24} + \underset{(0.14)}{0.94} r_t^* + e_t \qquad \text{RSS} = 28.56$$

where r_t is the observed interest rate, and r_t^* is the average expectation of r_t held at the end of the preceding quarter. Figures in parentheses are estimated standard errors. The sample data on r^* give

$$\sum_t r_t^*/30 = 10 \qquad \sum_t (r_t^* - \bar{r}^*)^2 = 52$$

Carry out the test, assuming that all basic assumptions of the classical regression model are satisfied.

(University of Michigan, 1981)

3.14. Consider the following regression model in deviation form:

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + u_t$$

with sample data

$$n = 100 \qquad \sum y^2 = \frac{493}{3} \qquad \sum x_1^2 = 30 \qquad \sum x_2^2 = 3$$

$$\sum x_1 y = 30 \qquad \sum x_2 y = 20 \qquad \sum x_1 x_2 = 0$$

(*a*) Compute the LS estimates of β_1 and β_2, and also calculate R^2.
(*b*) Test the hypothesis H_0: $\beta_2 = 7$ against H_1: $\beta_2 \neq 7$.
(*c*) Test the hypothesis H_0: $\beta_1 = \beta_2 = 0$ against H_1: $\beta_1 \neq 0$ or $\beta_2 \neq 0$.
(*d*) Test the hypothesis H_0: $\beta_2 = 7\beta_1$ against H_1: $\beta_2 \neq 7\beta_1$.

(University of London, 1981)

3.15. Given the following least-squares estimates,

$$\begin{aligned} C_t &= \text{constant} + 0.92Y_t + e_{1t} \\ C_t &= \text{constant} + 0.84C_{t-1} + e_{2t} \\ C_{t-1} &= \text{constant} + 0.78Y_t + e_{3t} \\ Y_t &= \text{constant} + 0.55C_{t-1} + e_{4t} \end{aligned}$$

calculate the least-squares estimates of β_2 and β_3 in

$$C_t = \beta_1 + \beta_2 Y_t + \beta_3 C_{t-1} + u_t$$

(University of Michigan, 1989)

3.16. Prove that R^2 is the square of the simple correlation between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$, where $\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

3.17. Prove that if a regression is fitted *without* a constant term, the residuals will not necessarily sum to zero, and R^2, if calculated as $1 - \boldsymbol{e}'\boldsymbol{e}/(\boldsymbol{y}'\boldsymbol{y} - n\bar{Y}^2)$, may be negative.

3.18. Prove that $\bar{R}^2$ increases with the addition of an extra explanatory variable only if the $F\ (= t^2)$ statistic for testing the significance of that variable exceeds unity. If r_i denotes the partial correlation coefficient associated with X_i, prove that

$$r_i^2 = \frac{F}{F + \text{df}} = \frac{t^2}{t^2 + \text{df}}$$

where F and t are the values of the test statistics for X_i and df is the number of degrees of freedom in the regression.

3.19. An economist is studying the variation in fatalities from road traffic in different states. She hypothesizes that the fatality rate depends on the average speed and the standard deviation of speed in each state. No data are available on the standard deviation, but in a normal distribution the standard deviation can be approximated by the difference between the 85th percentile and the 50th percentile. Thus, the specified model is

$$Y = \beta_1 + \beta_2 X_2 + \beta_3(X_3 - X_2) + u$$

where Y = fatality rate
X_2 = average speed
X_3 = 85th percentile speed

Instead of regressing Y on X_2 and $(X_3 - X_2)$ as in the specified model, the research assistant fits

$$Y = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + u$$

with the following results:

$$Y = \text{constant} - 0.24X_2 + 0.20X_3 + e$$

with $R^2 = 0.62$. The t statistics for the two slope coefficients are 1.8 and 2.3, respectively, and the covariance of the regression slopes is 0.003. Use these regression results to calculate a point estimate of β_2 and test the hypothesis that average speed has no effect on fatalities.

3.20. Data on a three-variable problem yield the following results:

$$\boldsymbol{X'X} = \begin{bmatrix} 33 & 0 & 0 \\ 0 & 40 & 20 \\ 0 & 20 & 60 \end{bmatrix} \qquad \boldsymbol{X'y} = \begin{bmatrix} 132 \\ 24 \\ 92 \end{bmatrix} \qquad \sum(Y - \bar{Y})^2 = 150$$

(*a*) What is the sample size?

(*b*) Compute the regression equation.

(*c*) Estimate the standard error of b_2 and test the hypothesis that β_2 is zero.

(*d*) Test the same hypothesis by running the appropriate restricted regression and examining the difference in the residual sums of squares.

(*e*) Compute a conditional prediction for Y_f, given $x_{2f} = -4$ and $x_{3f} = 2$. Obtain also a 95 percent interval for this prediction. If the actual value of Y_f turned out to be 12, would you think it came from the relationship underlying the sample data?

CHAPTER 4

Some Tests of the *k*-Variable Linear Equation for Specification Error

The least-squares technique of Chapter 3 is the workhorse of econometrics and applied statistics, routinely used in the analysis of myriad data sets. It is often referred to as **Ordinary Least Squares** (OLS) because it is derived from the simplest set of assumptions about the equation. Given the assumptions of Chapter 3, the least-squares estimators have the desirable properties enumerated and can also be employed in an attractive array of exact inference procedures. However, there is a crucial question. *How do we know if the assumptions underlying OLS are valid for a given data set?* How do we know the properties of the unobservable disturbance term? How do we know which variables should be included in the X matrix and in what functional form? If any of the underpinning assumptions are invalid, what happens to the OLS estimators? Are they still useful in any sense, or are they seriously flawed and misleading? Are there alternative estimators and inference procedures that may be more appropriate under alternative assumptions? These questions will be pursued in this and subsequent chapters.

If any of the underlying assumptions are wrong, there is a **specification error.** Although some specification errors may have minor implications, others may be much more serious; and it is extremely important to be alert to the possibility of specification errors and to test for their presence. It will be seen in later chapters that more complex specifications than those underlying the OLS technique are often required and lead to the corresponding development of appropriate inference procedures.

4.1 SPECIFICATION ERROR

The specification of the linear model centers on the disturbance vector $\boldsymbol{u}$ and the X matrix. The assumptions made in Chapter 3 were as follows:

$$y = X\beta + u \tag{4.1}$$

$$u_i \text{ are iid } (0, \sigma^2) \qquad i = 1, \ldots, n \tag{4.2a}$$

or

$$u_i \text{ are iid } N(0, \sigma^2) \qquad i = 1, \ldots, n \tag{4.2b}$$

$$E(X_{it}u_s) = 0 \text{ for all } i = 1, \ldots, k \text{ and } t, s = 1, \ldots, n \tag{4.3}$$

$$X \text{ is nonstochastic with full column rank } k \tag{4.4}$$

Assumption (4.2*a*) postulates that the disturbances are **white noise** and (4.2*b*) that they are **Gaussian white noise**. Given the fixed regressor assumption, Eq. (4.3) follows trivially from the assumed zero mean for the disturbance term.

What might go wrong? We will indicate some of the major possibilities for departure from these assumptions. This outline, however, is only preliminary, and several important topics will be dealt with in later chapters.

4.1.1 Possible Problems with *u*

1. Assumption (4.2*a*) holds but (4.2*b*) does not. As already indicated in Chapter 2, this does not destroy the BLUE property of OLS, but the inference procedures are now only asymptotically valid.
2. $E(uu') = \text{diag}[\sigma_1^2 \cdots \sigma_n^2]$. The variance-covariance matrix for u is diagonal with different variances on the main diagonal and zeros everywhere else, so the assumption of **homoscedasticity** is violated. This is the simplest form of **heteroscedasticity**, frequently found in cross-section applications, although this and more complicated forms may also be found in time series applications. The detection of heteroscedasticity and the development of appropriate inference procedures will be taken up in Chapter 6.
3. $E(u_t u_{t-s}) \neq 0, (s \neq 0)$. Here the disturbances are assumed to be pairwise correlated. In time series applications there may be strong correlations between adjacent disturbances and, perhaps, smaller correlations between disturbances further apart. Similarly in cross-section data certain units may share common disturbances. Tests for **autocorrelated disturbances** and relevant inference procedures will also be discussed in Chapter 6.

4.1.2 Possible Problems with *X*

1. Exclusion of relevant variables. Economic theory teaches that income and prices jointly affect demand, so we would not expect to obtain a good estimate of a price elasticity if we left income out of a demand equation. However, in more complicated situations, which variables should be incorporated into a relation is often not clear, and this becomes an important specification problem.
2. Inclusion of irrelevant variables. This is the converse of Problem 1. Now the maintained hypothesis contains some variables that have no business there. There are some consequences for inference procedures, but they are generally less serious than those pertaining to the exclusion of relevant variables.

3. Incorrect functional form. We may have an appropriate list of variables but have embedded them in an incorrect functional form. Sometimes this can still be dealt with in the context of the linear model. For instance, a relation $Y = f(X_2, X_3)$ might be specified as

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

or perhaps

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_2 X_2^2 + \gamma_3 X_3^2 + \delta(X_2 X_3) + u$$

The second equation allows for both a quadratic response to the regressors and an interaction effect. The interaction effect is based on a new variable, which is the product of the two regressors. Thus, the expected effect of a unit change in X_2 is $\beta_2 + 2\gamma_2 X_2 + \delta X_3$, so it depends on β_2 and the current levels of both X_2 and X_3. The expected effect of a unit change in X_3, likewise, will depend on the level of X_2 as well as X_3. If the specification error consists of using the first equation rather than the second, it can easily be corrected by adding terms in X_2^2, X_3^2, and $(X_2 X_3)$. Other times an intrinsically nonlinear specification may be required, as in some of the examples in Chapter 2.

4. The X matrix has less than full column rank. This precludes estimation of a unique $\boldsymbol{b}$ vector. Often the regressors may be close to linear dependence, in which case the OLS vector can be computed, but the elements are likely to have large standard errors. This is known as the **collinearity** problem.

5. Nonzero correlations between the regressors and the disturbance. This is a breakdown of Assumption (4.3). It may occur in a variety of ways. As shown in Chapter 2 it can happen when a lagged value of Y appears as a regressor. Such a value will be uncorrelated with the current and future disturbances, but will be correlated with past disturbances. OLS estimates will now be biased in finite samples, but will be consistent and asymptotically normally distributed, since the Mann-Wald theorem, described in Chapter 2, is applicable. A more serious breakdown occurs if a regressor is correlated with the *current* disturbance. The OLS estimates are then biased and inconsistent. Such a condition occurs when there are measurement errors in the regressor(s) or when the equation under consideration is one of a system of simultaneous equations. These situations will be discussed in later chapters.

6. Nonstationary variables. As mentioned briefly in Chapter 2 most traditional inference procedures implicitly assume stationary variables. When this is not the case inference procedures are nonstandard, and we enter the realm of integrated variables, cointegration, error correction models, and the rest, all to be discussed at a later stage.

4.1.3 Possible Problems with $\boldsymbol{\beta}$

The assumption implicit in Eq. (4.1) is that the $\boldsymbol{\beta}$ vector is constant over all actual or possible sample observations. There may, however, be sudden structural breaks in coefficients, or slow evolution in response to changing social and environmental factors. It would not be reasonable to expect the elasticity of the demand for apples to be the same in the Garden of Eden as in Los Angeles toward the close of the twentieth century. That circumstance, however, would not preclude the development of a demand function that would have useful practical applications in the current situation.

4.2 MODEL EVALUATION AND DIAGNOSTIC TESTS

Standard econometric practice for a long time was to (*i*) formulate a model on the basis of theory or previous econometric findings, (*ii*) estimate the parameters of the model using what relevant sample data one could obtain, and (*iii*) inspect the resultant estimates and associated statistics to judge the adequacy of the specified model. That inspection typically focused on the overall fit, the agreement of the signs of the coefficients with a priori expectation, the statistical significance of the coefficients, and a test of autocorrelation in the disturbances. If the model were deemed "satisfactory" on these criteria, a new equation would be added to the literature and might well be used to make predictions for data points outside the time scale or empirical range of the sample. If the estimated model were deemed "unsatisfactory," the investigator would engage in a specification search, trying out different reformulations in an attempt to reach a "satisfactory" equation. That search process went largely unreported, for it smacked of **data mining,** which was held to be reprehensible, and also because it was practically impossible to determine correct P-values and confidence coefficients for the final statistics.[1]

In recent years there has been a substantial shift in opinion about good econometric practice, largely initiated by Denis Sargan of the London School of Economics, who wrote in 1975.

> Despite the problems associated with "data mining" I consider that a suggested specification should be tested in all possible ways, and only those specifications which survive and correspond to a reasonable economic model should be used.[2]

This approach has been actively developed, especially by David Hendry and associates.[3] The result is a battery of available diagnostic tests. Their use is not routine or automatic but requires judgment and economic intuition or good sense. Some tests may point to a particular specification error or errors. Others may indicate that a specification does not perform very well without locating a precise problem, and a specification may survive some tests and not others.

[1] See, for example, Michael C. Lovell, "Data Mining," *The Review of Economics and Statistics*, LXV, 1983, 1–12.

[2] J. D. Sargan, Discussion on Misspecification in *Modelling the Economy*, ed. G. A. Renton, Heinemann, 1975, quoted in Adrian R. Pagan, "Model Evaluation by Variable Addition," Chapter 5, *Econometrics and Quantitative Economics*, eds. David F. Hendry and Kenneth Wallis, Blackwell, 1984, p. 131.

[3] For details see any recent PC-GIVE manual (David F. Hendry et al, Institute of Economics and Statistics, University of Oxford, UK).

4.3 TESTS OF PARAMETER CONSTANCY

One of the most important criteria for an estimated equation is that it should have relevance for *data outside the sample data used in the estimation.* This criterion is embodied in the notion of **parameter constancy,** that is, that the $\boldsymbol{\beta}$ vector should apply both outside and within the sample data. Parameter constancy may be examined in various ways. One of the most useful is a test of predictive accuracy.

4.3.1 The Chow Forecast Test[4]

If the parameter vector is constant, out-of-sample predictions will have specified probabilities of lying within bounds calculated from the sample data. "Large" prediction errors therefore cast doubt on the constancy hypothesis, and the converse for "small" prediction errors. Instead of using all the sample observations for estimation, the suggested procedure is to divide the data set of n sample observations into n_1 observations to be used for estimation and $n_2 = n - n_1$ observations to be used for testing. With time series data one usually takes the first n_1 observations for estimation and the last n_2 for testing. In cross-section applications the data could be partitioned by the values of a *size* variable, such as household income or firm revenue, profits, employment, etc. There are no hard and fast rules for determining the relative sizes of n_1 and n_2. It is not uncommon to reserve 5, 10, or 15 percent of the observations for testing.

The test of predictive accuracy, widely referred to as the Chow test in honor of Chow's influential 1960 article, is as follows:

1. Estimate the OLS vector from the n_1 observations, obtaining
$$\boldsymbol{b}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}_1 \tag{4.5}$$
where $\boldsymbol{X}_i, \boldsymbol{y}_i$ $(i = 1, 2)$ indicate the partitioning of the data into n_1, n_2 observations.
2. Use $\boldsymbol{b}_1$ to obtain a prediction of the $\boldsymbol{y}_2$ vector, namely,
$$\hat{\boldsymbol{y}}_2 = \boldsymbol{X}_2\boldsymbol{b}_1 \tag{4.6}$$
3. Obtain the vector of prediction errors and analyze its sampling distribution under the null hypothesis of parameter constancy.

The vector of prediction errors is

$$\boldsymbol{d} = \boldsymbol{y}_2 - \hat{\boldsymbol{y}}_2 = \boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{b}_1 \tag{4.7}$$

If the equation $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, with $E(\boldsymbol{uu}') = \sigma^2\boldsymbol{I}$. holds for both data sets, the vector of prediction errors may be written

$$\boldsymbol{d} = \boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{b}_1 = \boldsymbol{u}_2 - \boldsymbol{X}_2(\boldsymbol{b}_1 - \boldsymbol{\beta})$$

[4] G. C. Chow, "Tests of Equality between Sets of Coefficients in Two Linear Regressions," *Econometrica*, **52,** 1960, 211–22.

Thus $E(\boldsymbol{d}) = \mathbf{0}$, and it may be shown[5] that the variance-covariance matrix for $\boldsymbol{d}$ is

$$\begin{aligned} \text{var}(\boldsymbol{d}) &= E(\boldsymbol{dd}') \\ &= \sigma^2 \boldsymbol{I}_{n_2} + X_2 \cdot \text{var}(\boldsymbol{b}_1) \cdot X_2' \\ &= \sigma^2 \left[\boldsymbol{I}_{n_2} + X_2(X_1'X_1)^{-1}X_2'\right] \end{aligned} \tag{4.8}$$

If we assume Gaussian disturbances,

$$\boldsymbol{d} \sim N[\mathbf{0}, \text{var}(\boldsymbol{d})]$$

and so $$\boldsymbol{d}'[\text{var}(\boldsymbol{d})]^{-1}\boldsymbol{d} \sim \chi^2(n_2)$$

Further, $$\boldsymbol{e}_1'\boldsymbol{e}_1/\sigma^2 \sim \chi^2(n_1 - k)$$

where $\boldsymbol{e}_1'\boldsymbol{e}_1$ is the residual sum of squares from the estimated regression. The two χ^2 statistics are distributed independently. Thus, under the hypothesis of parameter constancy,

$$F = \frac{\boldsymbol{d}'\left[\boldsymbol{I}_{n_2} + X_2(X_1'X_1)^{-1}X_2'\right]^{-1}\boldsymbol{d}/n_2}{\boldsymbol{e}_1'\boldsymbol{e}_1/(n_1 - k)} \sim F(n_2, n_1 - k) \tag{4.9}$$

Large values of this F statistic would reject the hypothesis that the same $\boldsymbol{\beta}$ vector applies within and outside the estimation data. The derivation of var($\boldsymbol{d}$) has assumed that σ^2 is the same in each subset of data. The F test in Eq. (4.9) is therefore conditional on that assumption. If the disturbances are heteroscedastic, the F statistic may overstate the true significance level.

There is an illuminating alternative way of deriving this test of predictive accuracy, due to Pagan and Nicholls, following on an earlier article by Salkever.[6] It also provides a simpler calculation. Suppose we allow for the possibility of a different coefficient vector in the forecast period. The complete model could then be written as

$$\begin{aligned} \boldsymbol{y}_1 &= X_1\boldsymbol{\beta} + \boldsymbol{u}_1 \\ \boldsymbol{y}_2 = X_2\boldsymbol{\alpha} + \boldsymbol{u}_2 &= X_2\boldsymbol{\beta} + X_2(\boldsymbol{\alpha} - \boldsymbol{\beta}) + \boldsymbol{u}_2 = X_2\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{u}_2 \end{aligned} \tag{4.10}$$

where $\boldsymbol{\gamma} = X_2(\boldsymbol{\alpha} - \boldsymbol{\beta})$. If $\boldsymbol{\gamma} = \mathbf{0}$, then $\boldsymbol{\alpha} = \boldsymbol{\beta}$, and the coefficient vector is constant over the estimation and forecast periods. The model is written compactly as

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} X_1 & \mathbf{0} \\ X_2 & \boldsymbol{I}_{n_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix} \tag{4.11}$$

If X denotes the augmented matrix in Eq. (4.11), then

$$X'X = \begin{bmatrix} X_1'X_1 + X_2'X_2 & X_2' \\ X_2 & I \end{bmatrix}$$

[5] See Appendix 4.1.

[6] Adrian Pagan and D. Nicholls, "Estimating Prediction Errors and Their Standard Deviations Using Constructed Variables," *Journal of Econometrics*, **24**, 1984, 293–310; and D. Salkever, "The Use of Dummy Variables to Compute Predictions, Prediction Errors, and Confidence Intervals," *Journal of Econometrics*, **4**, 1976, 393–397.

where the n_2 subscript on $\boldsymbol{I}$ has been omitted for simplicity. The inverse is

$$(X'X)^{-1} = \begin{bmatrix} (X_1'X_1)^{-1} & -(X_1'X_1)^{-1}X_2' \\ -X_2(X_1'X_1)^{-1} & I + X_2(X_1'X_1)^{-1}X_2' \end{bmatrix}$$

Estimating Eq. (4.11) by OLS gives

$$\begin{aligned} \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{c} \end{bmatrix} &= \begin{bmatrix} (X_1'X_1)^{-1} & -(X_1'X_1)^{-1}X_2' \\ -X_2(X_1'X_1)^{-1} & I + X_2(X_1'X_1)^{-1}X_2' \end{bmatrix} \begin{bmatrix} X_1'y_1 + X_2'y_2 \\ y_2 \end{bmatrix} \\ &= \begin{bmatrix} (X_1'X_1)^{-1}X_1'y_1 \\ y_2 - X_2(X_1'X_1)^{-1}X_1'y_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{d} \end{bmatrix} \end{aligned} \tag{4.12}$$

Thus the first k OLS coefficients replicate the $\boldsymbol{b}_1$ estimate of $\boldsymbol{\beta}$, obtained from the n_1 data points; and the n_2 remaining coefficients, which estimate the $\boldsymbol{\gamma}$ vector, are simply the prediction errors defined in Eq. (4.7). The OLS estimation of Eq. (4.11) may be written

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ X_2 & I \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{d} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \end{bmatrix} \tag{4.13}$$

However, the second equation in Eq. (4.13) is

$$y_2 = X_2\boldsymbol{b}_1 + \boldsymbol{d} + \boldsymbol{e}_2$$

Substituting for $\boldsymbol{d}$ gives $\boldsymbol{e}_2 = \boldsymbol{0}$. Thus the RSS from fitting Eq. (4.11) is simply $\boldsymbol{e}_1'\boldsymbol{e}_1$. It is clear from Eq. (4.10) that the hypothesis of constant $\boldsymbol{\beta}$ is equivalent to $H_0\colon \boldsymbol{\gamma} = \boldsymbol{0}$. This hypothesis is tested by examining the joint significance of the last n_2 variables in the augmented regression (4.13). This is a direct application of the general procedure for testing a set of linear restrictions developed in Eq. (3.38). Making the appropriate substitutions in Eq. (3.38) gives the test statistic as

$$F = \frac{\boldsymbol{d}'[\text{var}(\boldsymbol{d})]^{-1}\boldsymbol{d}/n_2}{\boldsymbol{e}_1'\boldsymbol{e}_1/(n_1 - k)} \sim F(n_2, n_1 - k) \tag{4.14}$$

The degrees of freedom in the denominator of this expression are obtained from $(n_1 + n_2)$ observations less $(k + n_2)$ estimated parameters. The variance matrix var $(\boldsymbol{d})$ is given by σ^2 times the submatrix in the bottom right-hand corner of $(X'X)^{-1}$ and is the same as the expression already obtained in Eq. (4.8). Substitution in Eq. (4.14) replicates the Chow statistic in Eq. (4.9).

Finally, an even simpler calculation is available in terms of a restricted and an unrestricted regression. As seen in Chapter 3, tests of linear restrictions can always be reformulated in this way. In the present case the unrestricted regression is, Eq. (4.13), with RSS $= \boldsymbol{e}_1'\boldsymbol{e}_1$. The restricted regression is obtained by setting $\boldsymbol{\gamma}$ to zero, that is, by estimating

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \boldsymbol{b}_* + \boldsymbol{e}_*$$

with RSS $= \boldsymbol{e}_*'\boldsymbol{e}_*$. The resultant test statistic for $\boldsymbol{\gamma} = \boldsymbol{0}$ is

$$F = \frac{(e_*'e_* - e_1'e_1)/n_2}{e_1'e_1/(n_1 - k)} \sim F(n_2, n_1 - k) \tag{4.15}$$

The Chow test may thus be implemented as follows:

1. Using the designated n_1 observations, regress y_1 on X_1 and obtain the RSS, $e_1'e_1$.
2. Fit the same regression to all $(n_1 + n_2)$ observations and obtain the restricted RSS, $e_*'e_*$.
3. Substitute in Eq. (4.15) and reject the hypothesis of parameter constancy if F exceeds a preselected critical value.

4.3.2 The Hansen Test[7]

A difficulty with the Chow test is the arbitrary nature of the partitioning of the data set. One such partitioning might reject the null hypothesis and another fail to reject. This difficulty does not apply to the Hansen test, which fits the linear equation to all n observations. It allows a somewhat more general specification of the model than that used in Chapter 3, although it does rule out nonstationary regressors. The technical derivation of the test is beyond the level of this book, but it is possible to give an outline of the procedure. Write the equation as

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + u_t \qquad t = 1, \ldots, n \tag{4.16}$$

The lowercase letters now denote the actual levels of the variables, not deviations from sample means; and the first variable, x_{1_t}, is typically equal to one. Denote the OLS residuals, as usual, by

$$e_t = y_t - b_1 x_{1t} - b_2 x_{2t} - \cdots - b_k x_{kt} \qquad t = 1, \ldots, n$$

The OLS fit gives the conditions

$$\sum_{t=1}^{n} x_{it} e_t = 0 \qquad i = 1, \ldots, k \tag{4.17}$$

and

$$\sum_{t=1}^{n} (e_t^2 - \hat{\sigma}^2) = 0 \tag{4.18}$$

This latter condition defines an estimate of the disturbance variance as $\hat{\sigma}^2 = \sum_{t=1}^{n} e_t^2/n$, which differs slightly from the unbiased estimator $s^2 = \sum_{t=1}^{n} e_t^2/(n-k)$. Define

$$f_{it} = \begin{cases} x_{it}e_t & i = 1, \ldots, k \\ e_t^2 - \hat{\sigma}^2 & i = k + 1 \end{cases}$$

Combining Eqs. (4.17) and (4.18) gives

[7] Bruce E. Hansen, "Testing for Parameter Instability in Linear Models," *Journal of Policy Modeling*, **14**, 1992, 517–533.

$$\sum_{t=1}^{n} f_{it} = 0 \qquad i = 1, \ldots, k+1 \tag{4.19}$$

The Hansen test statistics are based on *cumulative sums* of the f_{it}, namely,

$$S_{it} = \sum_{j=1}^{t} f_{ij} \tag{4.20}$$

He develops tests for the stability of each parameter individually and for the joint stability of all parameters. The individual test statistics are

$$L_i = \frac{1}{nV_i}\sum_{t=1}^{n} S_{it}^2 \qquad i = 1, \ldots, k+1 \tag{4.21}$$

where

$$V_i = \sum_{t=1}^{n} f_{it}^2 \tag{4.22}$$

For the joint stability test let

$$f_t = [f_{1t} \quad \cdots \quad f_{k+1,t}]'$$

and

$$s_t = \begin{bmatrix} S_{1t} & \cdots & S_{k+1,t} \end{bmatrix}'$$

The joint stability test statistic is then

$$L_c = \frac{1}{n}\sum_{t=1}^{n} s_t'V^{-1}s_t \tag{4.23}$$

where

$$V = \sum_{t=1}^{n} f_t f_t' \tag{4.24}$$

Under the null hypothesis the cumulative sums will tend to be distributed around zero. Thus "large" values of the test statistics suggest rejection of the null. The distribution theory is nonstandard, and only asymptotic critical values are available. These are tabulated in Table 7 in Appendix D. There is a line for each number of parameters from 1 to 20. The first line of the table thus gives critical values for the individual coefficient test. The 5 percent critical value is 0.470, leading to a rough rule of thumb that a test value in excess of one-half suggests an unstable parameter. For five parameters, including the variance, the 5 percent critical value is approximately 1.5. The Hansen test is already incorporated in some econometric packages.

4.3.3 Tests Based on Recursive Estimation

The model in Eq. (4.1) may be written

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + u_t \qquad t = 1, \ldots, n \tag{4.25}$$

where y_t is the tth observation on the dependent variable, and $\boldsymbol{x}_t' = [1 \; x_{2t} \; \cdots \; x_{kt}]$

is the row vector of regressors at the tth sample point, again using lowercase letters to denote the levels of variables. The complete sample matrix of regressors is

$$X = \begin{bmatrix} \cdots & x_1' & \cdots \\ \cdots & x_2' & \cdots \\ & \vdots & \\ \cdots & x_n' & \cdots \end{bmatrix}$$

The idea behind recursive estimation is very simple. Fit the model to the first k observations. The fit will be perfect since there are only k regression coefficients to be estimated. Next use the first $k + 1$ data points and compute the coefficient vector again. Proceed in this way, adding one sample point at a time, until the final coefficient vector is obtained, based on all n sample points. This process generates a *sequence* of vectors, $\boldsymbol{b}_k, \boldsymbol{b}_{k+1}, \ldots, \boldsymbol{b}_n$, where the subscript indicates the number of sample points used in the estimation. In general,

$$\boldsymbol{b}_t = (X_t'X_t)^{-1}X_t'\boldsymbol{y}_t \tag{4.26}$$

where X_t is the $t \times k$ matrix of regressors for the first t sample points, and $\boldsymbol{y}_t$ is the t-vector of the first t observations on the dependent variable. The standard errors of the various coefficients may be computed at each stage of the recursion, except at the first step, since the RSS is zero when $t = k$. Some software packages *initialize* the recursive calculations at some $m > k$, generating the sequence $\boldsymbol{b}_m, \boldsymbol{b}_{m+1}, \ldots, \boldsymbol{b}_n$. Graphs may be prepared showing the evolution of each coefficient, plus and minus two standard errors. Visual inspection of the graphs may suggest parameter constancy, or its reverse. As data are added, graphs sometimes display substantial vertical movement, to a level outside previously estimated confidence limits. This phenomenon is usually the result of the model trying to digest a structural change and leads one to suspect parameter inconstancy. Recursive estimation is an appealing procedure with time series data, since time gives a unique ordering of the data. However, the procedure is readily applicable to cross-section data, which can be ordered by a suitable "size" variable, if required.

4.3.4 One-Step Ahead Prediction Errors

By using all data up to and including period $t - 1$, the one-step ahead prediction of y_t is $\boldsymbol{x}_t'\boldsymbol{b}_{t-1}$. The one-step ahead prediction error is thus

$$v_t = y_t - \boldsymbol{x}_t'\boldsymbol{b}_{t-1} \tag{4.27}$$

From Eq. (4.8) the variance of the one-step ahead prediction error is

$$\text{var}(v_t) = \sigma^2\left[1 + \boldsymbol{x}_t'(X_{t-1}'X_{t-1})^{-1}\boldsymbol{x}_t\right] \tag{4.28}$$

The unknown σ^2 in Eq. (4.28) can be replaced by the residual variance estimated from the first $(t-1)$ observations, provided $t-1 > k$. Taking the square root gives the estimated standard error of regression (S.E.R.). Plus or minus twice these recursively estimated standard errors can be plotted around the zero line and the actual prediction errors (also referred to as recursive residuals) shown on the same graph. Residuals lying outside the standard error bands are suggestive of parameter inconstancy. At each point the probability of the observed error under the null hypothesis can be calculated from the appropriate t distribution, as in Eq. (3.48).

4.3.5 CUSUM and CUSUMSQ Tests

Scaled recursive residuals are defined as

$$w_t = \frac{v_t}{\sqrt{[1 + \boldsymbol{x}_t'(\boldsymbol{X}_{t-1}'\boldsymbol{X}_{t-1})^{-1}\boldsymbol{x}_t]}} \qquad t = k+1, \ldots, n \tag{4.29}$$

Under the assumptions in Eqs. (4.1) and (4.2b)

$$w_t \sim N(0, \sigma^2)$$

It can also be shown that the scaled recursive residuals are pairwise uncorrelated.[8] Thus,

$$\boldsymbol{w} = N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{n-k}) \tag{4.30}$$

Brown et al. suggest a pair of tests of parameter constancy, based on these scaled recursive residuals. The first test statistic is the **CUSUM** quantity,

$$W_t = \sum_{j=k+1}^{t} w_j/\hat{\sigma} \qquad t = k+1, \ldots, n \tag{4.31}$$

where $\hat{\sigma}^2 = \mathrm{RSS}_n/(n-k)$, with RSS_n being the residual sum of squares calculated from the full-sample regression. W_t is a *cumulative* sum, and it is plotted against t. With constant parameters, $E(W_t) = 0$, but with nonconstant parameters W_t will tend to diverge from the zero mean value line. The significance of the departure from the zero line may be assessed by reference to a pair of straight lines that pass through the points

$$(k, \pm a\sqrt{n-k}) \qquad \text{and} \qquad (n, \pm 3a\sqrt{n-k})$$

where a is a parameter depending on the significance level α chosen for the test. The correspondence for some conventional significance levels is

$$\alpha = 0.01 \qquad a = 1.143$$
$$\alpha = 0.05 \qquad a = 0.948$$
$$\alpha = 0.10 \qquad a = 0.850$$

The lines are shown in Fig. 4.1.

The second test statistic is based on cumulative sums of squared residuals, namely,

[8] R. L. Brown, J. Durbin, and J. M. Evans, "Techniques for Testing the Constancy of Regression Relationships over Time," *Journal of the Royal Statistical Society, Series B*, 35, 1975, 149–192.

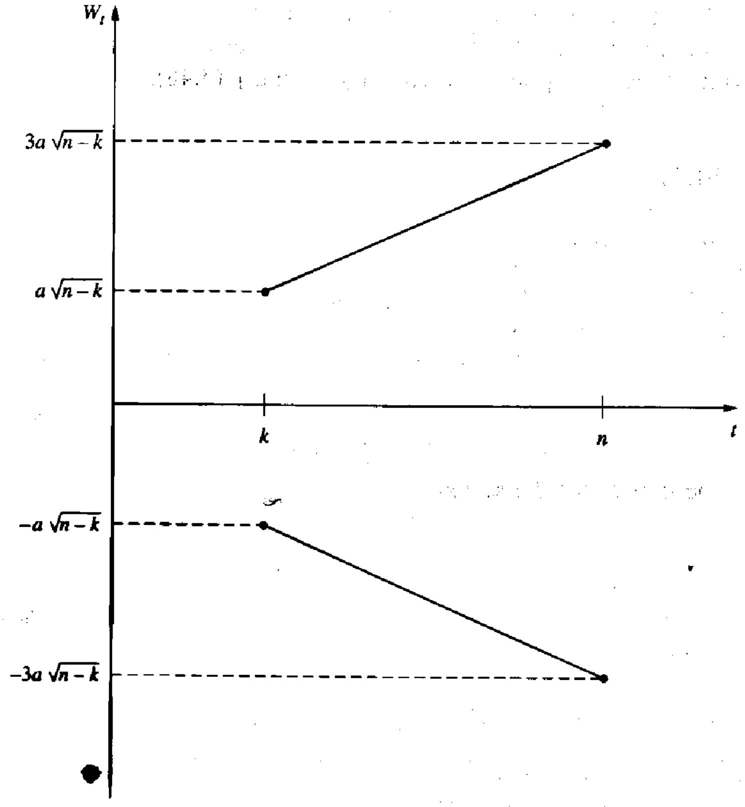


FIGURE 4.1
CUSUM plot.

$$S_t = \frac{\sum_{k+1}^{t} w_j^2}{\sum_{k+1}^{n} w_j^2} \qquad t = k+1, \ldots, n \tag{4.32}$$

Under the null hypothesis the squared w's are independent $\chi^2(1)$ variables. The numerator thus has an expected value of $t-k$, and the denominator an expected value of $n-k$. The mean value line, giving the *approximate* expected value of the test statistic under the null hypothesis, is

$$E(S_t) = \frac{t-k}{n-k}$$

which goes from zero at $t = k$ to unity at $t = n$. The significance of departures from the expected value line is assessed by reference to a pair of lines drawn parallel to the $E(S_t)$ line at a distance c_0 above and below. Values of c_0 from the Brown, Durbin, and Evans article for various sample sizes and significance levels are tabulated in Appendix D. Hansen (*op. cit.*) suggests that the CUSUM test is akin to his L_1 test (of the stability of the intercept), and that the CUSUMSQ test is akin to his L_{k+1} test (of the stability of the variance).

4.3.6 A More General Test of Specification Error: The Ramsey RESET Test

Ramsey has argued that various specification errors listed in Section 4.1 (omitted variables, incorrect functional form, correlation between X and u) give rise to a nonzero u vector.[9] Thus the null and alternative hypotheses are

$$H_0: u \sim N(0, \sigma^2 I)$$
$$H_1: u \sim N(\mu, \sigma^2 I) \qquad \mu \neq 0$$

The test of H_0 is based on an augmented regression

$$y = X\beta + Z\alpha + u$$

The test for specification error is then $\alpha = 0$. Ramsey's suggestion is that Z should contain powers of the *predicted values of the dependent variable*. Using the second, third, and fourth powers gives

$$Z = [\hat{y}^2 \quad \hat{y}^3 \quad \hat{y}^4]$$

where $\hat{y} = Xb$, and $\hat{y}^2 = [\hat{y}_1^2 \hat{y}_2^2 \ldots \hat{y}_n^2]'$ etc. The first power, $\hat{y}$, is not included since it is an exact linear combination of the columns of X. Its inclusion would make the regressor matrix $[X \; Z]$ have less than full rank.

4.4 A NUMERICAL ILLUSTRATION

This numerical example is not econometrically realistic. It is meant merely to illustrate the tests outlined in the previous section. The variables are those already introduced in Chapter 1, that is,

Y = log of per capita real expenditure on gasoline and oil

$X2$ = log of the real price index for gasoline and oil

$X3$ = log of per capita real disposable personal income

The first oil shock hit in 1973.4, so for this exercise we chose a sample period from 1959.1 to 1973.3, a period for which it might seem reasonable to postulate parameter constancy. As shown in Fig. 4.2, consumption and income trended fairly steadily upward during this period, and price trended downward with a greater rate of decrease in the second half of the sample period. The pairwise correlations are obviously fairly high, so it may be difficult to disentangle the relative contributions of the two explanatory variables. The first 51 observations were used for estimation and the remaining 8 reserved for the Chow forecast test. The simple specification

[9] J. B. Ramsey, "Tests for Specification Error in Classical Linear Least Squares Analysis," *Journal of the Royal Statistical Society, Series B*, **31**, 1969, 350–371. See also J. B. Ramsey and P. Schmidt, "Some Further Results on the Use of OLS and BLUS Residuals in Specification Error Tests," *Journal of the American Statistical Association*, **71**, 1976, 389–390.

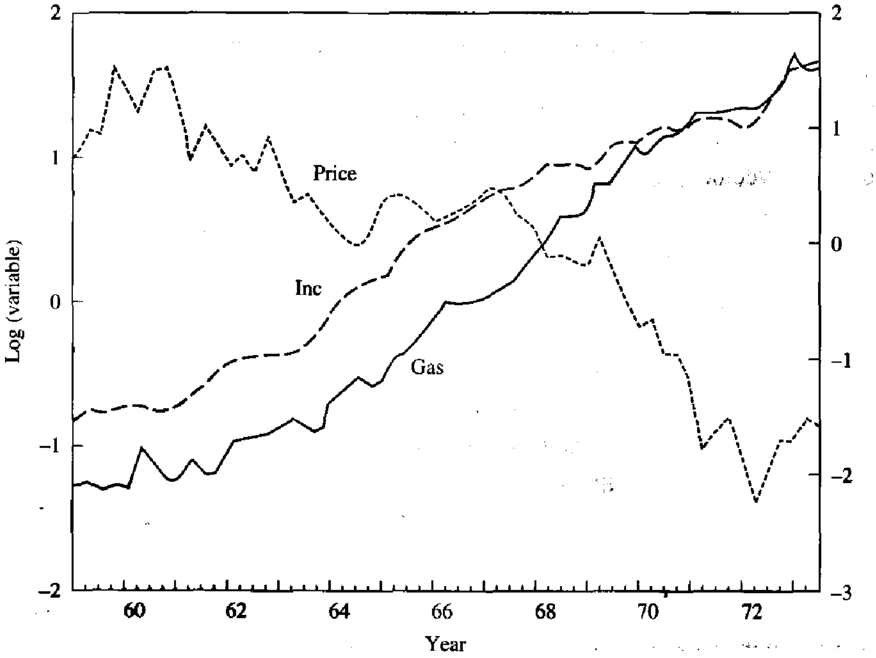


FIGURE 4.2
Gasoline consumption (Y), price (X2), and income (X3).

$$Y = \beta_1 + \beta_2(X2) + \beta_3(X3) + u$$

was employed. The results are shown in Table 4.1. They are a mixture of *good news* and *bad news*. Looking at the *good news* first, we see the specification appears economically sensible. The price elasticity is negative (−0.66), and the income elasticity is positive (0.85). Both coefficients are well determined by conventional standards, and the overall F statistic overwhelmingly rejects the null hypothesis that price and income have nothing to do with consumption. Furthermore, the specification passes the Chow test with flying colors, the F statistic being only 0.18. The one-step forecasts in Table 4.1 should not be confused with the one-step ahead prediction errors discussed in Section 4.2. The latter come from *recursive* estimation. The forecasts in Table 4.1 take the form $\hat{y}_t = x_t'b$, where b is the coefficient vector estimated from the 51 observations and x_t is the vector of explanatory variables in the forecast period. The forecast SE is given by $s\sqrt{1 + x_t'(X'X)^{-1}x_t}$, where s is the estimated standard error of the regression and X is the matrix of regressors for the first 51 sample points. This section of the table tests each forecast point individually, and we see that the two standard error range about the forecast includes each actual value of the dependent variable.

Now for the *bad news*. Two items of *bad news* are already contained in Table 4.1. The column headed Instab contains the Hansen statistics for testing the stability of individual coefficients. The hypothesis of stability is rejected for all three coefficients, and, not surprisingly, the joint stability test decisively rejects the null

TABLE 4.1
OLS Regression of Y on X2 and X3

Present sample: 1959.1 to 1973.3 less 8 forecasts
Forecast period: 1971.4 to 1973.3

Variable	Coefficient	Std Error	t-value	Instab
Constant	−1.1041	0.47043	−2.347	0.92**
X2	−0.66234	0.14546	−4.553	0.92**
X3	0.84791	0.060387	14.041	0.91**

(**: significant at 1% level)
$R^2 = 0.962856$ $F(2, 48) = 622.13$ [0.0000] $\sigma = 0.0232118$ $DW = 0.316$

Variance instability test: 0.12107; joint instability test: 4.0695**
Analysis of one-step forecasts

Date	Actual	Forecast	$Y - \hat{Y}$	Forecast SE	t-Value
1971.4	−7.65916	−7.67531	0.0161520	0.0249789	0.646624
1972.1	−7.65555	−7.66462	0.00907682	0.0264099	0.343690
1972.2	−7.65785	−7.64868	−0.00917203	0.0276412	−0.331825
1972.3	−7.65144	−7.64589	−0.00555639	0.0263309	−0.211021
1972.4	−7.63462	−7.63089	−0.00373734	0.0251797	−0.148427
1973.1	−7.60615	−7.62421	0.0180611	0.0251543	0.718011
1973.2	−7.62518	−7.63150	0.00631639	0.0247320	0.255394
1973.3	−7.62061	−7.62581	0.00519761	0.0247990	0.209590

Tests of parameter constancy over: 1971.4 to 1973.3
Chow $F(8, 48) = 0.18421$ [0.9920]

of parameter constancy.[10] A second disturbing piece of information is the very low value of the Durbin-Watson (DW) statistic. As will be explained in Chapter 6, this indicates substantial autocorrelation in the disturbance term, so a lot of action is not being explained by the included regressors. This also vitiates the estimated t and F values.

The presence of major specification error in this too simple equation is more firmly demonstrated by the recursive tests. Figure 4.3 shows the recursive residuals along with two standard error bands. The calculations underlying this figure are defined in Eqs. (4.27) and (4.28). A point on the graph lying outside the standard error bands is equivalent to a t statistic $[v_t/\text{s.e}(v_t)]$ being numerically greater than two and thus suggestive of parameter inconstancy. There is one such point in 1966, and a number of similar points from 1968 through 1970.

Figure 4.4, generated by PC-GIVE, is an alternative way of showing the same information as that given by Fig. 4.3, which is generated by EViews. The test implicit in Fig. 4.3 is a t test, whereas that implicit in Fig. 4.4 is an F test. The F statistic is the square of the corresponding t statistic. However, by following Eq. (4.15), the one-step Chow test for parameter constancy through the first j observations is based on

[10] The Hansen test may be inappropriate here, since Fig. 4.2 seems to suggest nonstationary variables. However, a more extended data sample will show price and consumption both reversing course, though income tends to move slowly upward.

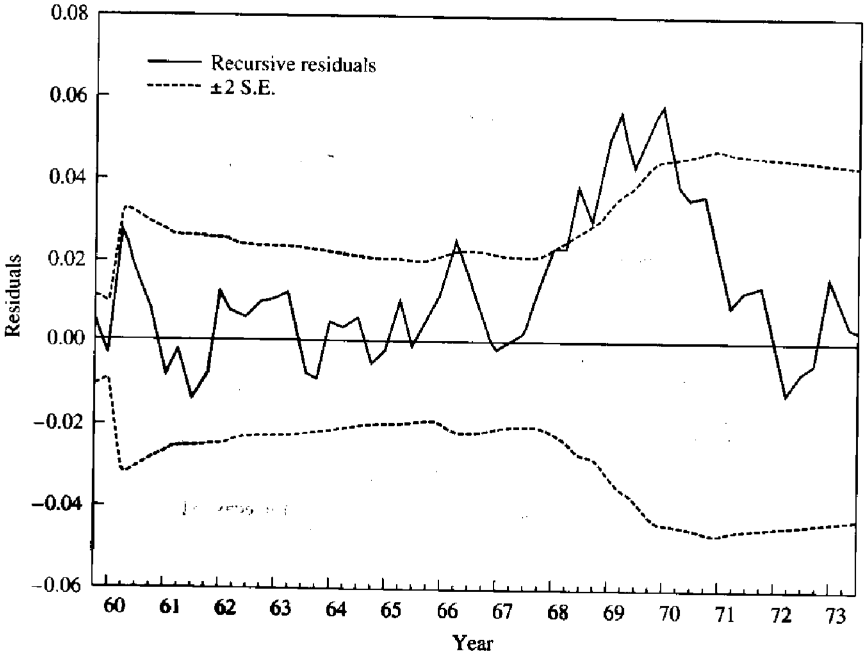


FIGURE 4.3
Recursive residuals and standard error bands for the gasoline equation of Section 4.4.

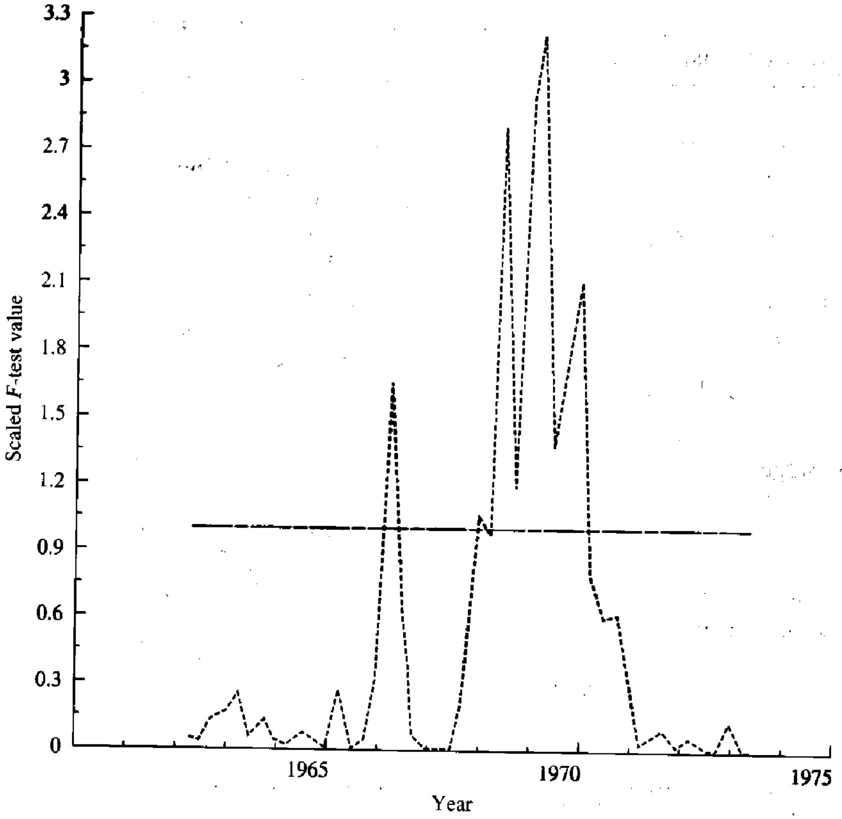
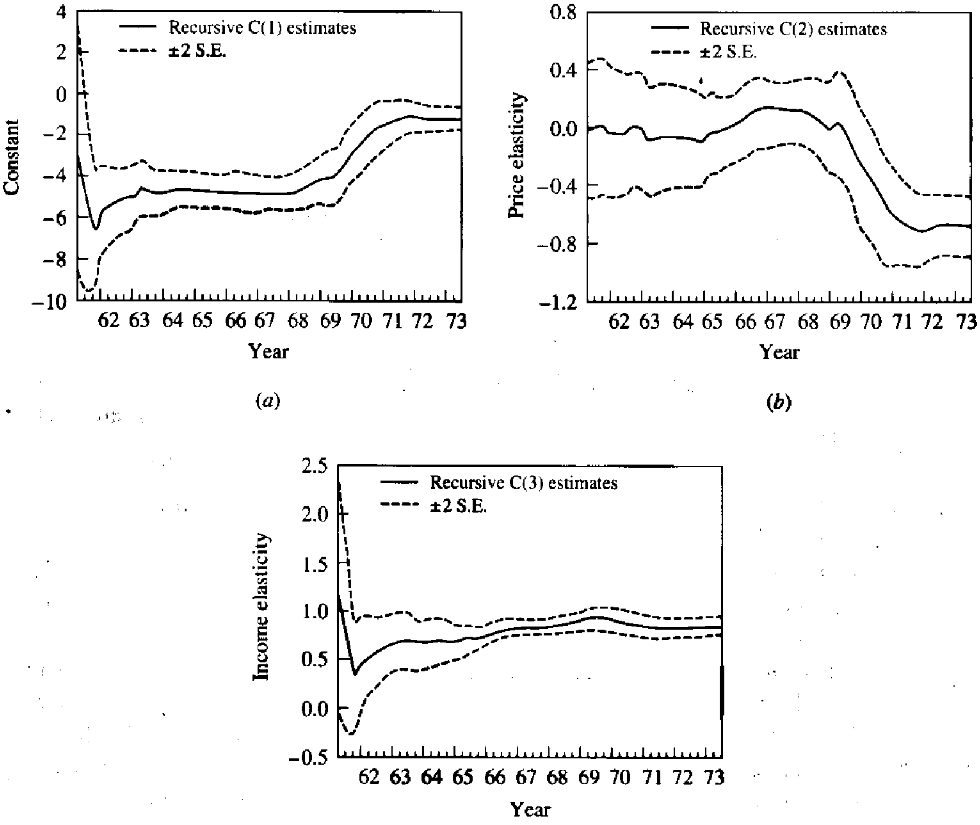


FIGURE 4.4
One-step forecast errors (scaled by 5 percent critical values).

$$F = \frac{\text{RSS}_j - \text{RSS}_{j-1}}{\text{RSS}_{j-1}/(j-k-1)} \qquad j = m+1,\ldots,n \tag{4.33}$$

where m is the number of observations used in the initial recursion. Under the null this statistic follows $F(1, j-k-1)$. Dividing the F statistic in Eq. (4.33) by the 5 percent critical value from $F(1, j-k-1)$ gives the series plotted in Fig. 4.4. Any point lying above the horizontal line at 1 implies rejection of parameter constancy, whereas points below do not lead to rejection. As in Fig. 4.3, there is one rejection in 1966 and a group of rejections in 1968 through 1970.

The three panels in Fig. 4.5 show the recursively estimated coefficients, with two standard error bands. As might be anticipated from Figs. 4.3 and 4.4, there are dramatic changes in the late 1960s, especially in the constant, C(1), and the price elasticity, C(2). In the first half of the sample, the price elasticity is not significantly different from zero, and the point estimate of price elasticity is positive. Only when data for the 1970s are included does the price elasticity turn negative, and significantly so. The income elasticity, C(3), is positive and reasonably stable.



(*a*)

(*b*)

(*c*)

FIGURE 4.5
Recursively estimated coefficients: (*a*) Constant, C(1); (*b*) price elasticity, C(2); (*c*) income elasticity, C(3).

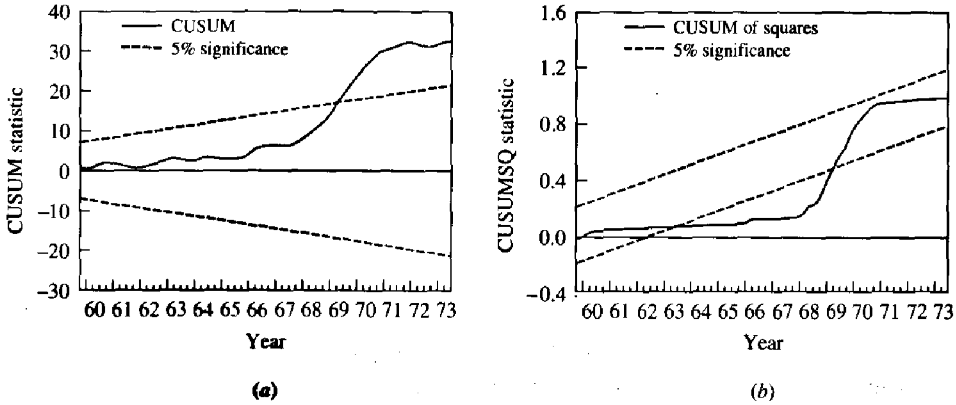


FIGURE 4.6
CUSUM tests of the gasoline equation in Section 4.4.

The CUSUM tests reported in Fig. 4.6 confirm the message of the previous figures. Finally, the Ramsey RESET test, using just $\hat{y}^2$, gives $F = 47.2$, which is a very strong indicator of specification error.

The process is somewhat akin to the medical examination of a patient by a doctor who is (one hopes) skillful. Some vital signs may be in the acceptable range, even as others give disturbing readings. The doctor must assess the patient's overall state of health and what, if anything, she is fit for. More importantly, can the doctor suggest appropriate remedial measures for the serious problems? In this case it seems clear that we have a very sick patient indeed. It remains to be seen in future chapters what remedies may be available.

4.5 TESTS OF STRUCTURAL CHANGE

The Chow forecast test leads naturally to more general tests of structural change. A structural change or structural break occurs if the parameters underlying a relationship differ from one subset of the data to another. There may, of course, be several relevant subsets of the data, with the possibility of several structural breaks. For the moment we will consider just two subsets of n_1 and n_2 observations making up the total sample of $n = n_1 + n_2$ observations. Suppose, for example, that one wishes to investigate whether the aggregate consumption in a country differs between peacetime and wartime and that we have observations on the relevant variables for n_1 peacetime years and n_2 wartime years. A Chow test could be performed by using the estimated peacetime function to forecast consumption in the wartime years. However, provided $n_2 > k$, one might alternatively use the estimated wartime function to forecast consumption in the peacetime years. It is not clear which choice should be made, and the two procedures might well yield different answers. If the subsets are large enough it is better to estimate both functions and test for common parameters.

4.5.1 Test of One Structural Change

The test of structural change may be carried out in three different, but equivalent, ways. Let y_i, X_i $(i = 1, 2)$ indicate the appropriate partitioning of the data. The unrestricted model may be written

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & \mathbf{0} \\ \mathbf{0} & X_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \boldsymbol{u} \qquad \boldsymbol{u} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}) \tag{4.34}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the k-vectors of peacetime and wartime coefficients respectively. The null hypothesis of no structural break is

$$H_0 \colon \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 \tag{4.35}$$

The first approach is a straightforward application of the test for linear restrictions, defined in Eqs. (3.28) and (3.38). The null hypothesis defines $\boldsymbol{R} = [\boldsymbol{I}_k \ -\boldsymbol{I}_k]$ and $\boldsymbol{r} = \mathbf{0}$. Substituting in Eq. (3.38) gives $\boldsymbol{Rb} - \boldsymbol{r} = \boldsymbol{b}_1 - \boldsymbol{b}_2$, where $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are the OLS estimates of the coefficient vectors in Eq. (4.34). Fitting Eq. (4.34) also provides the unrestricted RSS, $\boldsymbol{e}'\boldsymbol{e}$. The OLS coefficients may be written

$$\begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix} = \begin{bmatrix} X_1'X_1 & \mathbf{0} \\ \mathbf{0} & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1'y_1 \\ X_2'y_2 \end{bmatrix} = \begin{bmatrix} (X_1'X_1)^{-1}X_1'y_1 \\ (X_2'X_2)^{-1}X_2'y_2 \end{bmatrix}$$

Thus the unrestricted model may be estimated by setting up the data as in Eq. (4.34) and running OLS estimation once, or by fitting the equation separately to the peacetime data and to the wartime data. In the latter case the two RSSs must be summed to give the unrestricted RSS, that is, $\boldsymbol{e}'\boldsymbol{e} = \boldsymbol{e}_1'\boldsymbol{e}_1 + \boldsymbol{e}_2'\boldsymbol{e}_2$. Substitution in Eq. (3.38) tests the linear restrictions.

We have seen in Chapter 3 that a test of linear restrictions may also be formulated in terms of an unrestricted RSS and a restricted RSS. In this case the null hypothesis gives the restricted model as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \boldsymbol{\beta} + \boldsymbol{u} \tag{4.36}$$

Denoting the RSS from fitting Eq. (4.36) as $\boldsymbol{e}_*'\boldsymbol{e}_*$, the test of the null is given by

$$F = \frac{(\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e})/k}{\boldsymbol{e}'\boldsymbol{e}/(n-2k)} \sim F(k, n-2k)$$

For a third possibility, consider an alternative setup of the unrestricted model,

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & \mathbf{0} \\ X_2 & X_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1 \end{bmatrix} + \boldsymbol{u} \tag{4.37}$$

Now the test of H_0 is simply a test of the joint significance of the last k regressors. The choice of the most convenient procedure largely depends on the software package one is using.

4.5.2 Tests of Slope Coefficients

Frequently in economic research interest centers on the slope coefficients and one does not wish to impose restrictions on the intercept term. To explain such tests, partition the X matrices as

$$X_1 = \begin{bmatrix} i_1 & X_1^* \end{bmatrix} \qquad X_2 = \begin{bmatrix} i_2 & X_2^* \end{bmatrix}$$

where i_1, i_2 are n_1 and n_2 vectors of ones, and the X_i^* are matrices of the $k-1$ regressor variables. The conformable partitioning of the $\boldsymbol{\beta}$ vectors is

$$\boldsymbol{\beta}_1' = \begin{bmatrix} \alpha_1 & \boldsymbol{\beta}_1^{*\prime} \end{bmatrix} \qquad \boldsymbol{\beta}_2' = \begin{bmatrix} \alpha_2 & \boldsymbol{\beta}_2^{*\prime} \end{bmatrix}$$

The null hypothesis is now

$$\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_2^*$$

The unrestricted model is

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} i_1 & \mathbf{0} & X_1^* & \mathbf{0} \\ \mathbf{0} & i_2 & \mathbf{0} & X_2^* \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \boldsymbol{\beta}_1^* \\ \boldsymbol{\beta}_2^* \end{bmatrix} + u \tag{4.38}$$

The restricted model is

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} i_1 & \mathbf{0} & X_1^* \\ \mathbf{0} & i_2 & X_2^* \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \boldsymbol{\beta}^* \end{bmatrix} + u \tag{4.39}$$

The test of the null can be based on the RSS from these two regressions. Some regression packages automatically supply an intercept term (by inserting a column of ones in the regressors). This step must be suppressed in fitting Eqs. (4.38) and (4.39) to avoid generating linearly dependent regressors.

An alternative formulation of the unrestricted model is

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} i_1 & \mathbf{0} & X_1^* & \mathbf{0} \\ i_2 & i_2 & X_2^* & X_2^* \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 - \alpha_1 \\ \boldsymbol{\beta}_1^* \\ \boldsymbol{\beta}_2^* - \boldsymbol{\beta}_1^* \end{bmatrix} + u \tag{4.40}$$

The test of the joint significance of the last $k-1$ regressors in this setup is a test of the null hypothesis. The same caveat about the intercept term applies to the estimation of Eq. (4.40).

4.5.3 Tests of Intercepts

It might appear that a test of H_0: $\alpha_1 = \alpha_2$ is given by testing the significance of the second regressor in Eq. (4.40). However, such a test would normally make little sense. Since the estimation of Eq. (4.40) places no restrictions on the slope coefficients, the hypothesis of equal intercepts is then asking whether two *different* regression surfaces intersect the y axis at the same point. A test of differential intercepts makes more sense if it is made conditional on the assumption of *common* regression slopes. It now amounts to asking whether one regression plane is higher or lower than the other in the direction of the dependent variable. The unrestricted model for this test is that already specified in Eq. (4.39), where it appeared as the restricted model in the test of slope coefficients. The restricted model is

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} i_1 & X_1^* \\ i_2 & X_2^* \end{bmatrix} \begin{bmatrix} \alpha \\ \boldsymbol{\beta}^* \end{bmatrix} + \boldsymbol{u} \tag{4.41}$$

Contrasting RSS between Eqs. (4.39) and (4.41) then provides a test of equality of intercepts, given equal regression slopes.

The alternative setup of the unrestricted model is

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} i_1 & \mathbf{0} & X_1^* \\ i_2 & i_2 & X_2^* \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 - \alpha_1 \\ \boldsymbol{\beta}^* \end{bmatrix} + \boldsymbol{u} \tag{4.42}$$

Now a test of the significance of the second regressor tests the conditional hypothesis that the intercepts are equal.

4.5.4 Summary

There is a hierarchy of three models, namely,

$$\text{I:} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} i_1 & X_1^* \\ i_2 & X_2^* \end{bmatrix} \begin{bmatrix} \alpha \\ \boldsymbol{\beta}^* \end{bmatrix} + \boldsymbol{u}$$

Common parameters

$$\text{II:} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} i_1 & \mathbf{0} & X_1^* \\ \mathbf{0} & i_2 & X_2^* \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \boldsymbol{\beta}^* \end{bmatrix} + \boldsymbol{u}$$

Differential intercepts, common slope vectors

$$\text{III:} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} i_1 & \mathbf{0} & X_1^* & \mathbf{0} \\ \mathbf{0} & i_2 & \mathbf{0} & X_2^* \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \boldsymbol{\beta}_1^* \\ \boldsymbol{\beta}_2^* \end{bmatrix} + \boldsymbol{u}$$

Differential intercepts, differential slope vectors

Application of OLS to each model will yield a residual sum of squares, RSS, with associated degrees of freedom, respectively, of $n - k$, $n - k - 1$, and $n - 2k$. The test statistics for various hypotheses are then as follows:

H_0: $\alpha_1 = \alpha_2$ Test of differential intercepts

$$F = \frac{\text{RSS}_1 - \text{RSS}_2}{\text{RSS}_2/(n - k - 1)} \sim F(1, n - k - 1)$$

H_0: $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_2^*$ Test of differential slope vectors

$$F = \frac{(\text{RSS}_2 - \text{RSS}_3)/(k - 1)}{\text{RSS}_3/(n - 2k)} \sim F(k - 1, n - 2k)$$

H_0: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ Test of differential parameters (intercepts and slopes)

$$F = \frac{(\text{RSS}_1 - \text{RSS}_3)/k}{\text{RSS}_3/(n - 2k)} \sim F(k, n - 2k)$$

The degrees of freedom in the numerator of these expressions are the number of restrictions imposed in going from the relevant unrestricted model to the restricted model. This number is also equal to the difference in the degrees of freedom of the residual sums of squares in the numerator.

4.5.5 A Numerical Example

Suppose we have the following data:

$$n_1 = 5 \qquad n_2 = 10 \qquad k = 2$$

$$\boldsymbol{y}_1 = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 4 \\ 6 \end{bmatrix} \qquad \boldsymbol{x}_1 = \begin{bmatrix} 2 \\ 4 \\ 6 \\ 10 \\ 13 \end{bmatrix} \qquad \boldsymbol{y}_2 = \begin{bmatrix} 1 \\ 3 \\ 3 \\ 5 \\ 6 \\ 6 \\ 7 \\ 9 \\ 9 \\ 11 \end{bmatrix} \qquad \boldsymbol{x}_2 = \begin{bmatrix} 2 \\ 4 \\ 6 \\ 8 \\ 10 \\ 12 \\ 14 \\ 16 \\ 18 \\ 20 \end{bmatrix}$$

From these data we form the following n-vectors:

$$\boldsymbol{d}_1 = \begin{bmatrix} \boldsymbol{i}_1 \\ \boldsymbol{0} \end{bmatrix} \qquad \boldsymbol{d}_2 = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{i}_2 \end{bmatrix} \qquad \boldsymbol{i} = \begin{bmatrix} \boldsymbol{i}_1 \\ \boldsymbol{i}_2 \end{bmatrix}$$

$$\boldsymbol{z}_1 = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{0} \end{bmatrix} \qquad \boldsymbol{z}_2 = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{x}_2 \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}$$

The hierarchy of three models is then as follows:

I: Regression of $\boldsymbol{y}$ on $\boldsymbol{i}$ and $\boldsymbol{x}$
II: Regression of $\boldsymbol{y}$ on $\boldsymbol{d}_1$, $\boldsymbol{d}_2$, and $\boldsymbol{x}$
III: Regression of $\boldsymbol{y}$ on $\boldsymbol{d}_1$, $\boldsymbol{d}_2$, $\boldsymbol{z}_1$, and $\boldsymbol{z}_2$

Table 4.2 shows these three regressions. From the third part of the table the separate regressions are

$$\hat{y}_1 = -0.0625 + 0.4375x_1$$

$$\hat{y}_2 = 0.4000 + 0.5091x_2$$

In each regression the RSS is given by the entry labeled Sum of squared resid. The test of equality of the $\boldsymbol{\beta}$ vectors is thus

$$F = \frac{(6.5561 - 3.1602)/2}{3.1602/(15-4)} = 5.91$$

From the tables of the F distribution,

$$F_{0.95}(2,11) = 3.98 \qquad \text{and} \qquad F_{0.99}(2,11) = 7.21$$

TABLE 4.2

LS // Dependent Variable is Y

VARIABLE	COEFFICIENT	STD. ERROR	T-STAT.	2-TAIL SIG.
C	−0.0697842	0.3678736	−0.1896961	0.8525
X	0.5244604	0.0329917	15.896733	0.0000

R-squared	0.951074	Mean of dependent var	5.000000
Adjusted R-squared	0.947310	S.D. of dependent var	3.093773
S.E. of regression	0.710152	Sum of squared resid	6.556115
Log likelihood	−15.07669	F-statistic	252.7061
Durbin-Watson stat	1.343686	Prob(F-statistic)	0.000000

LS // Dependent Variable is Y

VARIABLE	COEFFICIENT	STD. ERROR	T-STAT.	2-TAIL SIG.
D1	−0.4658537	0.3048463	−1.5281589	0.1524
D2	0.5536585	0.3390025	1.6331992	0.1284
X	0.4951220	0.0266345	18.589463	0.0000

R-squared	0.973953	Mean of dependent var	5.000000
Adjusted R-squared	0.969612	S.D. of dependent var	3.093773
S.E. of regression	0.539309	Sum of squared resid	3.490244
Log likelihood	−10.34849	F-statistic	224.3564
Durbin-Watson stat	2.462522	Prob(F-statistic)	0.000000

LS // Dependent Variable is Y

VARIABLE	COEFFICIENT	STD. ERROR	T-STAT.	2-TAIL SIG.
D1	−0.0625000	0.4831417	−0.1293616	0.8994
D2	0.4000000	0.3661560	1.0924304	0.2980
Z1	0.4375000	0.0599263	7.3006283	0.0000
Z2	0.5090909	0.0295057	17.253988	0.0000

R-squared	0.976416	Mean of dependent var	5.000000
Adjusted R-squared	0.969984	S.D. of dependent var	3.093773
S.E. of regression	0.535998	Sum of squared resid	3.160227
Log likelihood	−9.603531	F-statistic	151.8074
Durbin-Watson stat	2.820099	Prob(F-statistic)	0.000000

Thus the hypothesis of no structural change would be rejected at the 5 percent level of significance, but not at the 1 percent level. The test of change in the regression slope is based on

$$F = \frac{3.4902 - 3.1602}{3.1602/11} = 1.15$$

with $F_{0.95}(1,11) = 4.84$. Thus the null of a common regression slope is not rejected. Given the assumption of a common regression slope, it is possible to test for common intercepts. The appropriate test statistic is

$$F = \frac{6.5561 - 3.4902}{3.4902/12} = 10.54$$

TABLE 4.3

LS // Dependent Variable is Y

VARIABLE	COEFFICIENT	STD. ERROR	T-STAT.	2-TAIL SIG.
C	−0.0625000	0.4831417	−0.1293616	0.8994
D2	0.4625000	0.6062146	0.7629312	0.4616
X	0.4375000	0.0599263	7.3006283	0.0000
Z2	0.0715909	0.0667964	1.0717786	0.3068

R-squared	0.976416	Mean of dependent var	5.000000
Adjusted R-squared	0.969984	S.D. of dependent var	3.093773
S.E. of regression	0.535998	Sum of squared resid	3.160227
Log likelihood	−9.603531	F-statistic	151.8074
Durbin-Watson stat	2.820099	Prob(F-statistic)	0.000000

LS // Dependent Variable is Y

VARIABLE	COEFFICIENT	STD. ERROR	T-STAT.	2-TAIL SIG.
C	−0.4658537	0.3048463	−1.5281589	0.1524
D2	1.0195122	0.3140167	3.2466815	0.0070
X	0.4951220	0.0266345	18.589463	0.0000

R-squared	0.973953	Mean of dependent var	5.000000
Adjusted R-squared	0.969612	S.D. of dependent var	3.093773
S.E. of regression	0.539309	Sum of squared resid	3.490244
Log likelihood	−10.34849	F-statistic	224.3564
Durbin-Watson stat	2.462522	Prob(F-statistic)	0.000000

with $F_{0.99}(1, 12) = 9.33$, so the difference in intercepts is significant at the 1 percent level.

Table 4.3 illustrates the alternative approach to the same three tests. The first part of the table gives the fit of the model in Eq. (4.40). Testing the *joint significance* of the second and fourth variables tests whether the $\boldsymbol{\beta}$ vectors (intercept and slope) are the same in the two subsets. MicroTSP (a pre-Windows program from Quantitative Micro Software) returns an F statistic for this test of 5.91, the same as that obtained in the preceding paragraph. The same table also provides a test of the equality of the regression slopes. This merely involves testing the significance of the fourth variable. The t statistic is 1.0718, which is clearly insignificant. Squaring the t statistic gives an F value of 1.15 as before.

The second part of Table 4.3 reports the results of fitting Eq. (4.42). The hypothesis of equal intercepts is tested by examining the significance of the second variable. The t statistic is 3.2467, and its square is 10.54, as before, so the null is rejected.

4.5.6 Extensions

There are two major ways in which these tests of structural change may be extended. The first is to split the total sample data into more than two subsets. One might examine the stability of a relation across several subperiods, (World War II, Cold War, Post–Cold War), or across countries, industries, social groups, or whatever. The classification of the subsets need not necessarily be time related, nor need the data within each subset be time series data. The same hierarchy of models applies, but now there are $p > 2$ subvectors in each column. The second extension is concerned with testing for common subsets of parameters. Testing for common intercepts and for common slope vectors are special cases, and we already know how to deal with them. More generally, we may wish to test any subset of the k parameters. The test procedure follows the general principles already established. Fit the restricted model with the subset of coefficients whose stability is under test taking the same value in each data subset. Leave all other coefficients free to vary across data subsets. Then fit the completely unrestricted model. The test is based on the contrast between the two residual sums of squares.

4.6 DUMMY VARIABLES

4.6.1 Introduction

We have already encountered dummy variables but have not so labeled them. The last n_2 variables in the augmented matrix in Eq. (4.11) take the form

$$\begin{bmatrix} \mathbf{0} \\ \boldsymbol{I}_{n_2} \end{bmatrix}$$

The $\mathbf{0}$ matrix is of order $n_1 \times n_2$, and $\boldsymbol{I}_{n_2}$ is the identity matrix of order n_2. Each n-vector column is a dummy variable, where a single element is one and the other $n - 1$ elements are all zero. As shown in Eq. (4.12) the effect of the dummies is to *exclude* the last n_2 observations from the estimation of the $\boldsymbol{\beta}$ vector. The coefficients of the dummy variables are the forecast errors for the last n_2 observations, and the regression residuals are zero at these points.

Sometimes a single dummy variable of this type is defined for an observation that is thought to be unusual. For example, 1973.4 was the quarter in which the Organization of Petroleum Exporting Countries (OPEC) oil embargo hit. In estimating an energy demand function, one might define a dummy variable with the value of one for this quarter and zero for all other quarters. The effect is shown in the first panel of Fig. 4.7, where a two-variable relation has been assumed for simplicity. The regression line is determined from the observations other than 1973.4, and for that point the regression line shifts to give the actual value of the dependent variable (Fig. 4.7a). Fig. 4.7b shows the case of three such dummy variables. The basic regression line is estimated from the $n - 3$ observations, and it makes three one-period shifts to pass through the three chosen values.

A second type of dummy variable takes the form

$$\boldsymbol{d}_2 = \begin{bmatrix} 0 \\ i_2 \end{bmatrix}$$

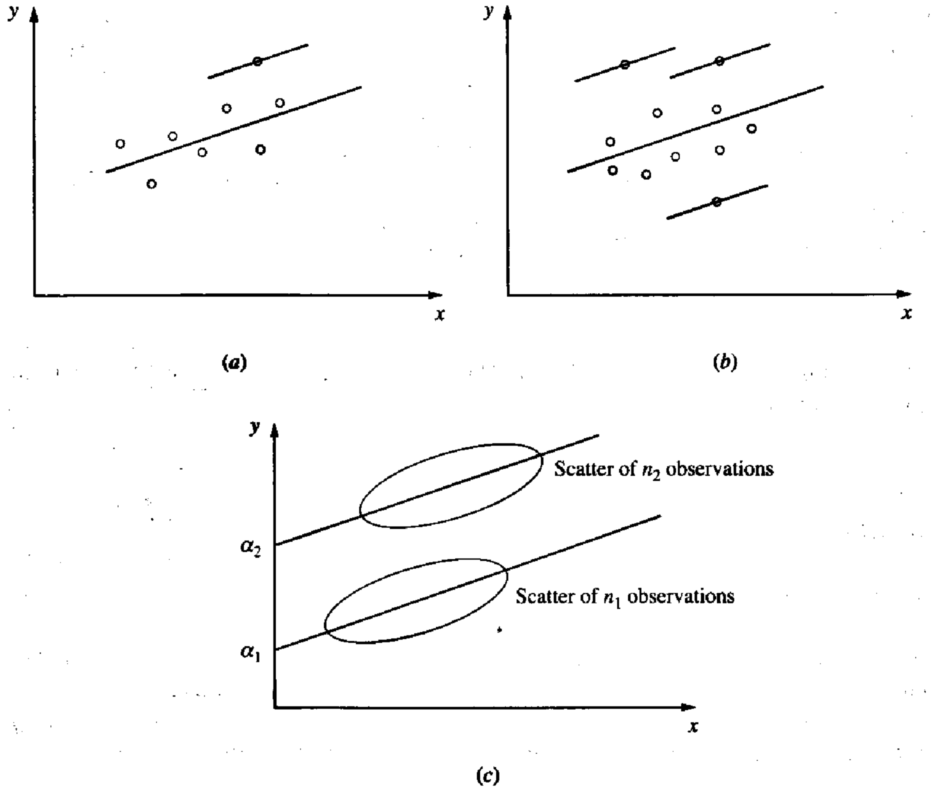


FIGURE 4.7
Regressions with dummy variables.

as in the models of structural change. The effect of fitting a model such as Eq. (4.42) is shown in Fig. 4.7*c*. There are two parallel regression lines, with all *n* observations being used to estimate the common slope.

4.6.2 Seasonal Dummies

In working with, say, quarterly data one may wish to allow for seasonal shifts in a relation. Vacation expenditure will depend on income but may have a positive or negative shift in certain quarters. This requires the specification of quarterly dummy variables, such as,

$$\begin{aligned} Q_{it} &= 1 \qquad \text{if observation is in quarter } i \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

for $i = 1, \ldots, 4$. For the four quarters of each year these dummies are

Q_1	Q_2	Q_3	Q_4
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

The relationship might then be written

$$Y_t = \alpha_1 Q_{1t} + \cdots + \alpha_4 Q_{4t} + \boldsymbol{x}_t' \boldsymbol{\beta} + \boldsymbol{u}_t \tag{4.43}$$

where x_t' contains observations on relevant regressors but must not contain an element of one, since a column of ones would be perfectly collinear with the four seasonal dummies, yielding a singular data matrix. The function has four intercepts, denoted by the α's. An alternative specification is

$$Y_t = \alpha_1 + \gamma_2 Q_{2t} + \gamma_3 Q_{3t} + \gamma_4 Q_{4t} + \boldsymbol{x}_t' \boldsymbol{\beta} + u_t \tag{4.44}$$

Comparing coefficients of the dummy variables in the two equations gives

$$\gamma_2 = \alpha_2 - \alpha_1 \qquad \gamma_3 = \alpha_3 - \alpha_1 \qquad \gamma_4 = \alpha_4 - \alpha_1$$

The γ's thus measure *differential* intercepts, by reference to α_1. The hypothesis of interest is usually

$$H_0: \quad \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$$

This could be tested by estimating Eq. (4.43) and testing the appropriate linear restrictions. Alternatively, the null hypothesis may be expressed as

$$H_0: \quad \gamma_2 = \gamma_3 = \gamma_4 = 0$$

This is easily tested by fitting Eq. (4.44) and testing the joint significance of the three quarterly dummies. The test statistic is invariant to the choice of which quarterly dummy to omit in moving from Eq. (4.43) to Eq. (4.44).

4.6.3 Qualitative Variables

Suppose a labor economist postulates an earnings function as

$$\text{Income} = f(\text{sex, race, educational level, age})$$

The first two explanatory variables are qualitative rather than quantitative. That is, they are not subject to cardinal measurement. They may, however, be represented by dummy variables. Sex is represented by two variables, namely,

$$\begin{aligned} S_1 &= 1 \text{ if male} \\ &= 0 \text{ otherwise} \end{aligned}$$

and

$$\begin{aligned} S_2 &= 1 \text{ if female} \\ &= 0 \text{ otherwise} \end{aligned}$$

The two categories are mutually exclusive and exhaustive. For each individual the sum of S_1 and S_2 is one. Suppose that race is defined by, say, three mutually exclusive and exhaustive categories (for example, Caucasian, Black, and other). Typical entries for the S and R dummy variables would look like

S_1	S_2	R_1	R_2	R_3
0	1	1	0	0
1	0	0	1	0

The first entry is for a female Caucasian, and the second is a male Black. The R columns sum to one. as do the S columns.

Educational level is a different type of variable. It could be represented in a cardinal fashion by years of education. Alternatively, it could be expressed in terms of dummy variables. One way is to classify educational level by the *highest* level of diploma awarded. say. high school diploma, bachelor diploma, or graduate diploma. Individuals with diplomas would have an entry of one for the relevant dummy variable and zero for the other two. Although the three categories are mutually exclusive by definition. they are not exhaustive, since there is no category for people without a diploma. Let E_1 be the dummy variable for dropouts, and E_2, E_3, E_4 be the dummy variables for the highest diploma awarded. If one modeled income just as a function of educational level, we would have

$$Y = \alpha_1 E_1 + \alpha_2 E_2 + \alpha_3 E_3 + \alpha_4 E_4 + u \tag{4.45}$$

The expected level of income, conditional on educational level, is

$$E(Y \mid E_i) = \alpha_i \qquad i = 1, \ldots, 4$$

If we suppress E_1. an alternative specification is

$$Y = \alpha_1 + \gamma_2 E_2 + \gamma_3 E_3 + \gamma_4 E_4 + u \tag{4.46}$$

The γ's measure the marginal increment in expected income for a diploma over the no-diploma level. The marginal increment for a bachelor diploma over a high school diploma is $\gamma_3 - \gamma_2$ and the marginal increment for a graduate diploma over a bachelor diploma is $\gamma_4 - \gamma_3$. The significance of each marginal increment may be tested by testing the relevant linear restriction. Alternatively, a reformulation of the dummy variables can provide direct estimates of the stepwise marginal increments. Let the dummy variable have a value of one if a person has the relevant diploma, *irrespective of whether he or she has one or more higher diplomas*. Also define E_1 as a pre–high school diploma dummy so that every individual has an entry of one for this dummy. The educational dummies for a person with only a high school diploma would then be [1 1 0 0], and a person with a graduate degree would show as [1 1 1 1]. The equation to be fitted is

$$Y = \alpha_1 + \delta_2 E_2 + \delta_3 E_3 + \delta_4 E_4 + u \tag{4.47}$$

The expected values are

$$\begin{aligned}
E(Y \mid \text{pre–HS diploma}) &= \alpha_1 \\
E(Y \mid \text{HS diploma}) &= \alpha_1 + \delta_2 \\
E(Y \mid \text{bachelor diploma}) &= \alpha_1 + \delta_2 + \delta_3 \\
E(Y \mid \text{graduate diploma}) &= \alpha_1 + \delta_2 + \delta_3 + \delta_4
\end{aligned}$$

Now the δ's provide direct estimates of the marginal increment from one level to the next higher level, and the corresponding standard errors provide a direct test of significance.

4.6.4 Two or More Sets of Dummy Variables

As we have seen, a mutually exclusive and exhaustive set of dummy variables sums to give the unit vector, i_n. Estimating a constant in a relationship is done by inserting a unit vector into the set of regressors. To avoid a singular data matrix the constant must be suppressed if a complete set of dummy variables is used; or, if the constant is retained, one of the dummy variables must be dropped. If there are *two* sets of dummy variables in a relationship and the constant is suppressed the estimation procedure still breaks down, because the included dummy variables are linearly dependent (the sum of the first set minus the sum of the second set gives the zero vector). If a constant is retained, *one dummy variable must be dropped from each set;* and this rule obviously extends to three or more sets.

4.6.5 A Numerical Example

Table 4.4 shows some hypothetical data on income (Y), conditional on sex (S), and education (E). For example, three persons in the first category for both sex and education have incomes of 8, 10, and 12. Inserting a constant and suppressing the first dummy variable in each set specifies the equation to be estimated as

$$Y = \mu + \alpha_2 E_2 + \alpha_3 E_3 + \beta_2 S_2 + u \tag{4.48}$$

TABLE 4.4

	E_1	E_2	E_3
S_1	8, 10, 12	12, 14	20, 22
S_2	5, 6	10, 12	20, 24

The relevant variables appear in column form as

Y	E_2	E_3	S_2
8	0	0	0
10	0	0	0
12	0	0	0
12	1	0	0
14	1	0	0
20	0	1	0
22	0	1	0
5	0	0	1
6	0	0	1
10	1	0	1
12	1	0	1
20	0	1	1
24	0	1	1

The estimated relationship is

$$\hat{Y} = 9 + 4E_2 + 13.5E_3 - 2S_2$$

TABLE 4.5

	E_1	E_2	E_3
S_1	9	13	22.5
S_2	7	11	20.5

TABLE 4.6

	E_1	E_2	E_3
S_1	10	13	21
S_2	5.5	11	22

The resultant estimated mean incomes for various combinations of sex and education are shown in Table 4.5. Every specification forces the data to conform to the straitjacket implicit in the specification. In this case a possibly undesirable feature of the specification is that the marginal increments for education are the same for each sex, and conversely, the sex difference is the same at all educational levels. It is preferable to test for this possibility rather than to impose it. The appropriate test is a test for **interaction effects.** It is carried out by adding two new dummy variables to the specification. These are the products (E_2S_2) and (E_3S_2). The revised specification is

$$Y = \mu + \alpha_2E_2 + \alpha_3E_3 + \beta_2S_2 + \gamma_2(E_2S_2) + \gamma_3(E_3S_2) + u \qquad (4.49)$$

Expected values are now

$$\begin{aligned}
E(Y \mid S_1, E_1) &= \mu \\
E(Y \mid S_1, E_2) &= \mu + \alpha_2 \\
E(Y \mid S_1, E_3) &= \mu + \alpha_3 \\
E(Y \mid S_2, E_1) &= \mu + \beta_2 \\
E(Y \mid S_2, E_2) &= \mu + \alpha_2 + \beta_2 + \gamma_2 \\
E(Y \mid S_2, E_3) &= \mu + \alpha_3 + \beta_2 + \gamma_3
\end{aligned}$$

Fitting Eq. (4.49) gives

$$\hat{Y} = 10 + 3E_2 + 11E_3 - 4.5S_2 + 2.5(E_2S_2) + 5.5(E_3S_2)$$

Table 4.6 shows the resultant estimated mean incomes. The entries in the cells are seen to be the arithmetic means of the raw data in Table 4.4. This special result is due to two factors. First, the only regressors are dummy variables; and allowing for interaction effects gives as many parameters to be estimated as there are cells in the original table. In practice, of course, most specifications contain cardinal regressors in addition to any dummy variables, so this effect would not be observed in general.

APPENDIX

APPENDIX 4.1
To show var$(\boldsymbol{d}) = \sigma^2\left[I_{n_2} + X_2(X_1'X_1)^{-1}X_2'\right]$

As shown in the text

$$\boldsymbol{d} = \boldsymbol{u}_2 - X_2(\boldsymbol{b}_1 - \boldsymbol{\beta})$$

From Eq. (3.23) $\quad \boldsymbol{b}_1 - \boldsymbol{\beta} = (X_1'X_1)^{-1}X_1'\boldsymbol{u}_1$

Thus $\quad \boldsymbol{d} = \boldsymbol{u}_2 - X_2(X_1'X_1)^{-1}X_1'\boldsymbol{u}_1$

Then $\quad E(\boldsymbol{d}) = \boldsymbol{0}$

and

$$\begin{aligned}\text{var}(\boldsymbol{d}) &= E(\boldsymbol{dd}') \\ &= E\left[\boldsymbol{u}_2 - X_2(X_1'X_1)^{-1}X_1'\boldsymbol{u}_1\right]\left[\boldsymbol{u}_2 - X_2(X_1'X_1)^{-1}X_1'\boldsymbol{u}_1\right]' \\ &= E(\boldsymbol{u}_2\boldsymbol{u}_2') + X_2(X_1'X_1)^{-1}X_1' \cdot E(\boldsymbol{u}_1\boldsymbol{u}_1') \cdot X_1(X_1'X_1)^{-1}X_2'\end{aligned}$$

The cross-product terms vanish since $E(\boldsymbol{u}_1\boldsymbol{u}_2') = \boldsymbol{0}$ by assumption. Substituting $E(\boldsymbol{u}_i\boldsymbol{u}_i') = \sigma^2 \boldsymbol{I}_{n_i}$ then gives the desired result. The variance of a single prediction error in Chapter 3, shown just before Eq. (3.48), is a special case of this result with $n_2 = 1$.

PROBLEMS

4.1. A bivariate regression is fitted to 20 sample observations on Y and X, where the data are expressed as

$$X'X = \begin{bmatrix} 20 & 10 \\ 10 & 30 \end{bmatrix} \qquad X'y = \begin{bmatrix} 30 \\ 40 \end{bmatrix} \qquad y'y = 75$$

A new observation is obtained, showing $X = 2$ and $Y = 4$. Calculate a Chow test for parameter constancy with these data.

4.2. A four-variable regression using quarterly data from 1958 to 1976 inclusive gave an estimated equation

$$\hat{Y} = 2.20 + 0.104X_2 - 3.48X_3 + 0.34X_4$$

The explained sum of squares was 109.6, and the residual sum of squares, 18.48. When the equation was re-estimated with three seasonal dummies added to the specification, the explained sum of squares rose to 114.8. Test for the presence of seasonality.

Two further regressions based on the original specification were run for the subperiods 1958.1 to 1968.4 and 1969.1 to 1976.4, yielding residual sums of squares of 9.32 and 7.46, respectively. Test for the constancy of the relationship over the two subperiods.

4.3. Gasoline sales in a regional market were modeled by the following regression equation, estimated with quarterly data:

$$\hat{Q} = 70 - 0.01P + 0.2Y - 1.5S_1 + 3.6S_2 + 4.7S_3$$

where Q is sales, P is price, Y is disposable income, and the S_i are quarterly dummy variables. The expected paths of P and Y for the next year are as follows:

Quarter	1	2	3	4
P	110	116	122	114
Y	100	102	104	103

Calculate the sales of gasoline to be expected in each quarter of the year.

Suppose another researcher proposes to use the same data to estimate an equation of the same form, except that she wishes to employ dummy variables S_2, S_3, and S_4. Write down the equation that will come from her calculations.

Yet another investigator proposes to suppress the intercept and use all four seasonal dummies. Write down the results of his estimation.

4.4. The model

$$Y = \alpha_1 + \gamma_2 E_2 + \gamma_3 E_3 + u$$

is estimated by OLS, where E_2 and E_3 are dummy variables indicating membership of the second and third educational classes. Show that the OLS estimates are

$$\begin{bmatrix} a_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 - \bar{Y}_1 \\ \bar{Y}_3 - \bar{Y}_1 \end{bmatrix}$$

where $\bar{Y}_i$ denotes the mean value of Y in the ith educational class.

4.5. Using the data in Table 4.4 estimate the specification

$$Y = \alpha_1 E_1 + \alpha_2 E_2 + \alpha_3 E_3 + \beta_2 S_2 + u$$

and estimate the resultant mean values for Table 4.5. Compare your results with the values given in the text and comment.

Repeat the exercise for the specification

$$Y = \mu + \alpha_1 E_1 + \alpha_2 E_2 + \beta_2 S_2 + u$$

4.6. Using the data of Table 4.4 estimate a specification of your own choosing without a constant term but with appropriate dummy variables to allow for interaction effects. Calculate the resultant version of Table 4.6 and compare with the results in the text.

4.7. Survey records for a sample of 12 families show the following weekly consumption expenditures (Y) and weekly incomes (X):

Y	70	76	91	100	105	113	122	120	146	135	147	155
X	80	95	105	115	125	135	145	155	165	175	185	200
		*			*	*			*	*	*	*

Families with an asterisk (*) reported that their income is higher than in the previous year. Using a linear consumption function, test whether the consumption behavior of families experiencing an increase in income is different from that of families who did not experience an increase.

4.8. Annual data for 1929–1967 (39 observations) are obtained for

Q = an index of U.S. GDP in constant dollars

L = an index of labor input

K = an index of capital input

A production function is estimated for the whole period as

$$\log Q = -3.8766 + 1.4106 \log L + 0.4162 \log K$$

with $R^2 = 0.9937$ and $s = 0.03755$.

Regressions for two subperiods yield

1929–1948

$$\log Q = -4.0576 + 1.6167 \log L + 0.2197 \log K$$

with $R^2 = 0.9759$ and $s = 0.04573$.

1949–1967

$$\log Q = -1.9564 + 0.8336 \log L + 0.6631 \log K$$

with $R^2 = 0.9904$ and $s = 0.02185$.
Test for the stability of the production function over the two subperiods.

4.9. The usual two-variable linear model is postulated, and a sample of 20 observations is drawn from an urban area and another sample of 10 observations from a rural area. The sample information in raw form is summarized as follows:
Urban

$$X'X = \begin{bmatrix} 20 & 20 \\ 20 & 25 \end{bmatrix} \qquad X'y = \begin{bmatrix} 10 \\ 20 \end{bmatrix} \qquad y'y = 30$$

Rural

$$X'X = \begin{bmatrix} 10 & 10 \\ 10 & 20 \end{bmatrix} \qquad X'y = \begin{bmatrix} 8 \\ 20 \end{bmatrix} \qquad y'y = 24$$

Test the hypothesis that the same relationship holds in both urban and rural areas.

4.10. A study of vacation expenditures in relation to income was based on data for 256 households, which were grouped into three separate income classes. Log linear regressions (with an intercept term) were computed for each income group and for all households with the following results:

Household income	Regression slope	Residual variance	Number of households
Low income	0.02	0.26	102
Middle income	0.09	0.42	102
High income	0.14	0.30	52
All households	0.07	0.38	256

Test whether the expenditure function is the same for all income groups.

What additional information would you need to test whether the expenditure elasticity is the same across income groups?

Given that the variance of the log of income in the complete sample is 24, test the hypothesis that the expenditure elasticity for all households is 0.10.

CHAPTER 5

Maximum Likelihood (ML), Generalized Least Squares (GLS), and Instrumental Variable (IV) Estimators

The maximum likelihood principle was introduced in Chapter 2. Now is the time to give a more comprehensive treatment.

5.1 MAXIMUM LIKELIHOOD ESTIMATORS

In recent years there has been a rapid development of new econometric tests, variously based on the Wald and Lagrange multiplier approaches. This has also led to a resurgence of interest in the maximum likelihood approach.

Let $\boldsymbol{y}' = [y_1 \quad y_2 \quad \cdots \quad y_n]$ be an n-vector of sample values, dependent on some k-vector of unknown parameters, $\boldsymbol{\theta}' = [\theta_1 \quad \theta_2 \quad \cdots \quad \theta_k]$. Let the joint density be written $f(\boldsymbol{y};\boldsymbol{\theta})$, which indicates the dependence on $\boldsymbol{\theta}$. This density may be interpreted in two different ways. For a given $\boldsymbol{\theta}$ it indicates the probability of a set of sample outcomes. Alternatively, it may be interpreted as a function of $\boldsymbol{\theta}$, conditional on a set of sample outcomes. In the latter interpretation it is referred to as a *likelihood* function. The formal definition is

$$\text{Likelihood function} = L(\boldsymbol{\theta};\boldsymbol{y}) = f(\boldsymbol{y};\boldsymbol{\theta}) \tag{5.1}$$

It is customary to reverse the order of the symbols in writing the likelihood function to emphasize the new focus of interest. Maximizing the likelihood function with respect to $\boldsymbol{\theta}$ amounts to finding a specific value. say $\hat{\boldsymbol{\theta}}$, that maximizes the probability of obtaining the sample values that have actually been observed. Then $\hat{\boldsymbol{\theta}}$ is said to be the maximum likelihood estimator (MLE) of the unknown parameter vector $\boldsymbol{\theta}$.

In most applications it is simpler to maximize the log of the likelihood function. We will denote the log-likelihood by

$$l = \ln L$$

Then
$$\frac{\partial l}{\partial \theta} = \frac{1}{L}\frac{\partial L}{\partial \theta}$$

and the $\hat{\theta}$ that maximizes l will also maximize L. The derivative of l with respect to $\boldsymbol{\theta}$ is known as the **score,** $s(\boldsymbol{\theta};\boldsymbol{y})$. The MLE, $\hat{\boldsymbol{\theta}}$, is obtained by setting the score to zero, that is, by finding the value of $\boldsymbol{\theta}$ that solves

$$s(\boldsymbol{\theta};\boldsymbol{y}) = \frac{\partial l}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{5.2}$$

The widespread use of maximum likelihood estimators is largely due to a range of desirable properties, which are summarized in the next section.

5.1.1 Properties of Maximum Likelihood Estimators

The major properties of MLEs are *large-sample,* or *asymptotic,* ones. They hold under fairly general conditions.

1. **Consistency**

$$\text{plim}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$$

2. **Asymptotic normality**

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} N(\boldsymbol{\theta}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$$

This states that the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is normal with mean $\boldsymbol{\theta}$ and variance given by the inverse of $\boldsymbol{I}(\boldsymbol{\theta})$. $\boldsymbol{I}(\boldsymbol{\theta})$ is the **information matrix** and is defined in two equivalent ways by

$$\boldsymbol{I}(\boldsymbol{\theta}) = E\left[\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)'\right] = -E\left[\frac{\partial^2 l}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right] \tag{5.3}$$

In practice it is usually much easier to evaluate the second expression. When $\boldsymbol{\theta}$ is a k-vector, $\partial l/\partial \boldsymbol{\theta}$ denotes a column vector of k partial derivatives, that is,

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \partial l/\partial \theta_1 \\ \partial l/\partial \theta_2 \\ \vdots \\ \partial l/\partial \theta_k \end{bmatrix}$$

Each element in this score (or gradient) vector is itself a function of $\boldsymbol{\theta}$, and so may be differentiated partially with respect to each element in $\boldsymbol{\theta}$. For example,

$$\frac{\partial[\partial l/\partial \boldsymbol{\theta}]}{\partial \theta_1} = \begin{bmatrix} \frac{\partial^2 l}{\partial \theta_1^2} & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_k} \end{bmatrix}$$

where the second-order derivatives have been written as a row vector. Proceeding in this way yields a square, symmetric matrix of second-order derivatives, known as the Hessian matrix,

$$\frac{\partial^2 l}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} = \begin{bmatrix} \frac{\partial^2 l}{\partial \theta_1^2} & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 l}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 l}{\partial \theta_2^2} & \cdots & \frac{\partial^2 l}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 l}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 l}{\partial \theta_k^2} \end{bmatrix}$$

Notice that this is *not* the same matrix as $(\partial l/\partial \boldsymbol{\theta})(\partial l/\partial \boldsymbol{\theta})'$. The latter is also a square, symmetric $k \times k$ matrix, but its i,jth element is the product $(\partial l/\partial \theta_i)$ $(\partial l/\partial \theta_j)$.

3. **Asymptotic efficiency.** If $\hat{\theta}$ is the maximum likelihood estimator of a *single parameter* $\boldsymbol{\theta}$**, the** previous property means that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2)$$

for some finite constant σ^2. If $\tilde{\theta}$ denotes any other consistent, asymptotically normal estimator of θ, then $\sqrt{n}\tilde{\theta}$ has a normal limiting distribution whose variance is greater than or equal to σ^2. The MLE has minimum variance in the class of consistent, asymptotically normal estimators. The term **asymptotic variance** refers to the variance of a limiting distribution. Thus the asymptotic variance of $\sqrt{n}\hat{\theta}$ is σ^2. However, the term is also used to describe the variance of the asymptotic approximation to the unknown finite sample distribution. Thus an equivalent statement is that the asymptotic variance of $\hat{\theta}$ is σ^2/n. When $\boldsymbol{\theta}$ is a vector of parameters and $\hat{\boldsymbol{\theta}}$ is the MLE,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{V})$$

for some positive definite matrix V. If $\tilde{V}$ denotes the variance matrix of any other consistent, asymptotically normal estimator, then $\tilde{V} - V$ is a positive semidefinite matrix.

4. **Invariance.** If $\hat{\theta}$ is the MLE of θ and $g(\theta)$ is a continuous function of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.
5. **The score has zero mean and variance $I(\theta)$.** To demonstrate the zero mean we note that integrating the joint density over all possible values of $\boldsymbol{y}$ gives a value of one, that is,

$$\int \cdots \int f(y_1, y_2, \ldots, y_n; \theta)\, dy_1 \cdots dy_n = \int \cdots \int L\, dy = 1$$

Differentiating both sides with respect to $\boldsymbol{\theta}$ yields

$$\int \cdots \int \frac{\partial L}{\partial \theta} dy = 0$$

But
$$E(s) = \int \cdots \int \frac{\partial l}{\partial \theta} L\,dy$$
$$= \int \cdots \int \frac{\partial L}{\partial \theta} dy$$
$$= 0$$

It then follows that the variance of s is

$$\text{var}(s) = E(ss') = E\left[\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)'\right] = I(\boldsymbol{\theta})$$

5.2
ML ESTIMATION OF THE LINEAR MODEL

This section covers the maximum likelihood estimation of the linear model, which comprises many of the econometric applications. The equation is

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} \qquad \text{with} \qquad \boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

The multivariate normal density for $\boldsymbol{u}$ is

$$f(\boldsymbol{u}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)(\boldsymbol{u}'\boldsymbol{u})}$$

The multivariate density for $\boldsymbol{y}$ conditional on $\boldsymbol{X}$ is then

$$f(\boldsymbol{y} \mid \boldsymbol{X}) = f(\boldsymbol{u}) \left| \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{y}} \right|$$

where $|(\partial \boldsymbol{u}/\partial \boldsymbol{y})|$ is the absolute value of the determinant formed from the $n \times n$ matrix of partial derivatives of the elements of $\boldsymbol{u}$ with respect to the elements of $\boldsymbol{y}$.[1] Here this matrix is simply the identity matrix. Thus the log-likelihood function is

$$l = \ln f(\boldsymbol{y} \mid \boldsymbol{X}) = \ln f(\boldsymbol{u}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\boldsymbol{u}'\boldsymbol{u}$$
$$= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) \qquad (5.4)$$

The vector of unknown parameters, $\boldsymbol{\theta}$, has $k + 1$ elements, namely,

$$\boldsymbol{\theta}' = [\boldsymbol{\beta}', \sigma^2]$$

Taking partial derivatives gives

[1] See Appendix 5.1 on the change of variables in density functions.

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = -\frac{1}{\sigma^2}(-X'y + X'X\boldsymbol{\beta})$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\boldsymbol{\beta})'(y - X\boldsymbol{\beta})$$

Setting these partial derivatives to zero gives the MLEs as

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'y \tag{5.5}$$

and

$$\hat{\sigma}^2 = (y - X\hat{\boldsymbol{\beta}})'(y - X\hat{\boldsymbol{\beta}})/n \tag{5.6}$$

The MLE, $\hat{\boldsymbol{\beta}}$, is seen to be the OLS estimator $\boldsymbol{b}$, and $\hat{\sigma}^2$ is $\boldsymbol{e}'\boldsymbol{e}/n$ where $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb}$ is the vector of OLS residuals.[2] We know from least-squares theory that $E(e'e/(n-k)) = \sigma^2$. Thus $E(\hat{\sigma}^2) = \sigma^2(n-k)/n$, so that $\hat{\sigma}^2$ is biased for σ^2, though $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$. The second-order derivatives are

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = -\frac{X'X}{\sigma^2} \quad \text{with} \quad -E\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right) = \frac{X'X}{\sigma^2}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \sigma^2} = -\frac{X'\boldsymbol{u}}{\sigma^4} \quad \text{with} \quad -E\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \sigma^2}\right) = \boldsymbol{0}$$

$$\frac{\partial^2 l}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{\boldsymbol{u}'\boldsymbol{u}}{\sigma^6} \quad \text{with} \quad -E\left(\frac{\partial^2 l}{\partial (\sigma^2)^2}\right) = \frac{n}{2\sigma^4}$$

since $E(\boldsymbol{u}'\boldsymbol{u}) = n\sigma^2$.

The information matrix is

$$\boldsymbol{I}(\boldsymbol{\theta}) = \boldsymbol{I}\begin{pmatrix}\boldsymbol{\beta}\\ \sigma^2\end{pmatrix} = \begin{bmatrix}\frac{1}{\sigma^2}(X'X) & \boldsymbol{0}\\ \boldsymbol{0} & \frac{n}{2\sigma^4}\end{bmatrix}$$

and its inverse is

$$\boldsymbol{I}^{-1}\begin{pmatrix}\boldsymbol{\beta}\\ \sigma^2\end{pmatrix} = \begin{bmatrix}\sigma^2(X'X)^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \frac{2\sigma^4}{n}\end{bmatrix}$$

The zero off-diagonal terms indicate that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are distributed independently of one another.

Substituting the MLE values from Eqs. (5.5) and (5.6) in the log-likelihood function, Eq. (5.4), and exponentiating gives the *maximum* of the likelihood function as

$$\begin{aligned} L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) &= (2\pi e)^{-n/2}(\hat{\sigma}^2)^{-n/2} \\ &= \left(\frac{2\pi e}{n}\right)^{-n/2}(\boldsymbol{e}'\boldsymbol{e})^{-n/2} \qquad (5.7) \\ &= \text{constant} \cdot (\boldsymbol{e}'\boldsymbol{e})^{-n/2} \end{aligned}$$

where the constant does not depend on any of the parameters of the model.

[2] Do not confuse $\boldsymbol{e}$ or its elements with the mathematical constant e = 2.71828.

5.3 LIKELIHOOD RATIO, WALD, AND LAGRANGE MULTIPLIER TESTS

We will illustrate these tests in the context of linear hypotheses about $\boldsymbol{\beta}$. A general linear hypothesis takes the form

$$H_0\text{: } \boldsymbol{R\beta} = \boldsymbol{r} \tag{5.8}$$

where $\boldsymbol{R}$ is a $q \times k$ $(q < k)$ matrix of known constants and $\boldsymbol{r}$ a $q \times 1$ known vector. The main tests are LR, W, and LM tests.

5.3.1 Likelihood Ratio (LR) Tests

The MLEs in Eqs. (5.5) and (5.6) maximize the likelihood function without imposing any restrictions. The resultant value of $L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ in Eq. (5.7) is the *unrestricted* maximum likelihood and is expressible as a function of the *unrestricted residual sum of squares, e'e*. The model may also be estimated in *restricted* form by maximizing the likelihood subject to the restrictions, $\boldsymbol{R\beta} = \boldsymbol{r}$. Let the resultant estimators be denoted by $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}^2$. The relevant maximum of the likelihood is then obtained by substituting these values in the likelihood function to get $L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2)$. The restricted maximum cannot exceed the unrestricted maximum, but if the restrictions are valid one would expect the restricted maximum to be "close" to the unrestricted maximum. The *likelihood ratio* is defined as

$$\lambda = \frac{L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2)}{L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)}$$

and intuitively we expect to reject the null hypothesis if λ is "small." In some cases exact finite-sample tests of the "smallness" of λ can be derived for some special transformations of λ. However, a *large-sample test* of general applicability is available in that

$$\text{LR} = -2\ln\lambda = 2[\ln L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) - \ln L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2)] \overset{a}{\sim} \chi^2(q) \tag{5.9}$$

The restricted MLEs are derived by maximizing

$$l^* = l - \boldsymbol{\mu}'(\boldsymbol{R\beta} - \boldsymbol{r}) \tag{5.10}$$

where $\boldsymbol{\mu}$ is a $q \times 1$ vector of Lagrange multipliers, and l is the log-likelihood specified in Eq. (5.4). It can be shown that $\tilde{\boldsymbol{\beta}}$ is simply the restricted $\boldsymbol{b}_*$ vector already derived in the standard least squares analysis [see Eq. (3.43)]. This vector satisfies the constraints $\boldsymbol{Rb}_* = \boldsymbol{r}$. If we denote the corresponding residuals by

$$\boldsymbol{e}_* = \boldsymbol{y} - \boldsymbol{Xb}_*$$

the restricted MLE of σ^2 is $\tilde{\sigma}^2_* = \boldsymbol{e}_*'\boldsymbol{e}_*/n$, and so

$$L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) = \text{constant} \cdot (\boldsymbol{e}_*'\boldsymbol{e}_*)^{-n/2} \tag{5.11}$$

Substitution of Eqs. (5.7) and (5.11) into Eq. (5.9) gives the LR test statistic as LR $= n(\ln e_*'e_* - \ln e'e)$. For future reference we note some alternative forms of the LR statistic, namely,

$$\begin{aligned} \text{LR} &= n(\ln e_*'e_* - \ln e'e) \\ &= n\ln\left(1 + \frac{e_*'e_* - e'e}{e'e}\right) \\ &= n\ln\left(\frac{1}{1 - (e_*'e_* - e'e)/e_*'e_*}\right) \end{aligned} \tag{5.12}$$

The calculation of the LR statistic thus requires the fitting of both the restricted and the unrestricted model.

5.3.2 The Wald (W) Test

In the Wald procedure only the unrestricted $\hat{\boldsymbol{\beta}}$ is calculated. The vector $(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$ then indicates the extent to which the unrestricted ML estimates fit the null hypothesis. This vector being "close" to zero would tend to support the null hypothesis; "large" values would tend to contradict it. Since $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed with mean vector $\boldsymbol{\beta}$ and variance-covariance matrix $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$ it follows that, under H_0, $(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$ is asymptotically distributed as multivariate normal with zero mean vector and variance-covariance matrix $\boldsymbol{R}\boldsymbol{I}^{-1}(\boldsymbol{\beta})\boldsymbol{R}'$, where $\boldsymbol{I}^{-1}(\boldsymbol{\beta}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. As shown earlier, the information matrix for the linear regression model is block diagonal, so we can concentrate on the submatrix relating to $\boldsymbol{\beta}$. It then follows that[3]

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'[\boldsymbol{R}\boldsymbol{I}^{-1}(\boldsymbol{\beta})\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \overset{a}{\sim} \chi^2(q) \tag{5.13}$$

where q is the number of restrictions in $\boldsymbol{R}$. The asymptotic distribution still holds when the unknown σ^2 in $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$ is replaced by a consistent estimator $\hat{\sigma}^2 = e'e/n$. The result is the Wald statistic,

$$\text{W} = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})}{\hat{\sigma}^2} \overset{a}{\sim} \chi^2(q) \tag{5.14}$$

It was shown earlier [see Eq. (3.44)] that the numerator in Eq. (5.14) can also be expressed as $(e_*'e_* - e'e)$, so an alternative form of the Wald statistic for testing (5.8) is

$$\text{W} = \frac{n(e_*'e_* - e'e)}{e'e} \overset{a}{\sim} \chi^2(q) \tag{5.15}$$

[3] As seen in Chapter 3, this statistic also has an exact, finite sample $\chi^2(q)$ distribution when the disturbances are normally distributed.

5.3.3 Lagrange Multiplier (LM) Test

The LM test, also known as the score test, is based on the score (or gradient) vector,

$$s(\boldsymbol{\theta}) = \frac{\partial \ln L}{\partial \boldsymbol{\theta}} = \frac{\partial l}{\partial \boldsymbol{\theta}}$$

The unrestricted estimator, $\hat{\boldsymbol{\theta}}$, is found by solving $s(\hat{\boldsymbol{\theta}}) = \mathbf{0}$, where the notation $s(\hat{\boldsymbol{\theta}})$ indicates the score vector *evaluated at* $\hat{\boldsymbol{\theta}}$. When the score vector is evaluated at $\tilde{\boldsymbol{\theta}}$, the restricted estimator, it will in general not be zero. However, if the restrictions are valid, the restricted maximum, $l(\tilde{\boldsymbol{\theta}})$, should be close to the unrestricted maximum, $l(\hat{\boldsymbol{\theta}})$, and so the gradient at the former should be close to zero. As shown earlier, the score vector has zero mean and variance-covariance matrix given by the information matrix, $\boldsymbol{I}(\boldsymbol{\theta})$. The quadratic form, $s'(\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta})s(\boldsymbol{\theta})$, will then have a χ^2 distribution. Evaluating this quadratic form at $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ provides a test of the null hypothesis. The basic result is that, under the null hypothesis,

$$\text{LM} = s'(\tilde{\boldsymbol{\theta}})\boldsymbol{I}^{-1}(\tilde{\boldsymbol{\theta}})s(\tilde{\boldsymbol{\theta}}) \overset{a}{\sim} \chi^2(q) \tag{5.16}$$

Notice that each element in Eq. (5.16) is evaluated at $\tilde{\boldsymbol{\theta}}$. In contrast to the Wald test, we now need calculate only the restricted estimator. The popularity of LM tests is due to the fact that in many cases it is much easier to calculate the restricted estimator than the unrestricted estimator.

From the development leading up to Eqs. (5.5) and (5.6) the score vector is

$$s(\boldsymbol{\theta}) = \begin{bmatrix} \dfrac{\partial l}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial l}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{u} \\ -\dfrac{n}{2\sigma^2} + \dfrac{\boldsymbol{u}'\boldsymbol{u}}{2\sigma^4} \end{bmatrix}$$

n 6^2

To evaluate the score vector at the restricted estimator $\tilde{\boldsymbol{\theta}}$, we replace $\boldsymbol{u}$ by $\boldsymbol{e}_* = \boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}$ and σ^2 by $\tilde{\sigma}^2 = \boldsymbol{e}_*'\boldsymbol{e}_*/n$. The $\tilde{\boldsymbol{\beta}}$ vector satisfies $\boldsymbol{R}\tilde{\boldsymbol{\beta}} = \boldsymbol{r}$. Thus,

$$s(\tilde{\boldsymbol{\theta}}) = \begin{bmatrix} \dfrac{1}{\tilde{\sigma}^2}\boldsymbol{X}'\boldsymbol{e}_* \\ 0 \end{bmatrix}$$

The inverse of the information matrix was given before. Evaluating this at $\tilde{\boldsymbol{\theta}}$ gives

$$\boldsymbol{I}^{-1}(\tilde{\boldsymbol{\theta}}) = \begin{bmatrix} \tilde{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \dfrac{2\tilde{\sigma}^4}{n} \end{bmatrix}$$

Substitution in Eq. (5.16) then gives

$$\begin{aligned} \text{LM} &= \begin{bmatrix} \dfrac{1}{\tilde{\sigma}^2}\boldsymbol{e}_*'\boldsymbol{X} & 0 \end{bmatrix} \begin{bmatrix} \tilde{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \dfrac{2\tilde{\sigma}^4}{n} \end{bmatrix} \begin{bmatrix} \dfrac{1}{\tilde{\sigma}^2}\boldsymbol{X}'\boldsymbol{e}_* \\ 0 \end{bmatrix} \\ &= \frac{\boldsymbol{e}_*'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}_*}{\tilde{\sigma}^2} \qquad (5.17) \\ &= \frac{n\boldsymbol{e}_*'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}_*}{\boldsymbol{e}_*'\boldsymbol{e}_*} \end{aligned}$$

By recalling the expressions for the explained and total sum of squares from a multiple regression in Chapter 3, the LM statistic in Eq. (5.17) is

$$\mathrm{LM} = nR^2$$

where R^2 is the squared multiple correlation coefficient from the regression of $\boldsymbol{e}_*$ on X. If $\boldsymbol{e}_*$ does not have zero mean, R^2 is the uncentered R^2. However, if the restrictions involve only the slope coefficients, $\beta_2, \beta_3, \ldots, \beta_k$, as is usually the case, then $\boldsymbol{e}_*$ will have zero mean and the R^2 in this expression is the centered statistic from the conventional regression packages.[4] The LM test can therefore be implemented in two steps. First compute the restricted estimator $\tilde{\boldsymbol{\beta}}$ and obtain the resultant residual vector $\boldsymbol{e}_*$. Then regress $\boldsymbol{e}_*$ on X and refer nR^2 from this regression to $\chi^2(q)$. This two-step procedure occurs frequently in cases where maximization of the likelihood is equivalent to minimization of a sum of squares. It may be shown[5] that Eq. (5.17) may be rewritten as

$$\mathrm{LM} = \frac{n(\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e})}{\boldsymbol{e}_*'\boldsymbol{e}_*} \tag{5.18}$$

We can now illustrate the famous inequality for these three test statistics in the linear model, namely, W $\geq$ LR $\geq$ LM. Applying the first two terms of the logarithmic expansion $\ln(1 + z) = z - \frac{1}{2}z^2 + \cdots$ to the second expression for LR in Eq. (5.12) gives

$$\mathrm{LR} = \frac{n(\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e})}{\boldsymbol{e}'\boldsymbol{e}} - \frac{n}{2}\left(\frac{\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{e}'\boldsymbol{e}}\right)^2$$

which yields LR $\leq$ W. Similarly, using the third expression in Eq. (5.12) gives

$$\begin{aligned}\mathrm{LR} &= -n\ln\left(1 - \frac{\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{e}_*'\boldsymbol{e}_*}\right)\\ &= \frac{n(\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e})}{\boldsymbol{e}_*'\boldsymbol{e}_*} + \frac{n}{2}\left(\frac{\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{e}_*'\boldsymbol{e}_*}\right)^2\end{aligned}$$

so that LR $\geq$ LM and finally W $\geq$ LR $\geq$ LM. The tests are asymptotically equivalent but in general will give different numerical results in finite samples.

EXAMPLE 5.1. We return to the data of Example 3.3 and test H_0: $\beta_3 = 0$ by these asymptotic tests. From Table 3.2 we see that the unrestricted regression of Y on X_2 and X_3 gives $\boldsymbol{e}'\boldsymbol{e} = 1.5$; and the restricted regression, when X_3 is excluded, gives $\boldsymbol{e}_*'\boldsymbol{e}_* = 2.4$. Substitution in Eqs. (5.15), (5.12), and (5.18) in that order gives

$$\mathrm{W} = 5(2.4 - 1.5)/1.5 = 3.00$$

$$\mathrm{LR} = 5\ln(2.4/1.5) = 2.35$$

$$\mathrm{LM} = 5(2.4 - 1.5)/2.4 = 1.875$$

[4] See Appendix 5.2 for an explanation of centered and uncentered R^2.

[5] See Appendix 5.3.

The 5 percent point of $\chi^2(1)$ is 3.841, so the null hypothesis is not rejected by any of the tests, although not much can be expected from the use of asymptotic tests on such a small sample. When the residual series from the restricted regression of Y on X_2 is used as the dependent variable in a regression on X_2 and X_3, the resultant R^2 is 0.375, giving $nR^2 = 5(0.375) = 1.875$, in agreement with the value of the preceding LM statistic.

5.4 ML ESTIMATION OF THE LINEAR MODEL WITH NONSPHERICAL DISTURBANCES

The postulated model is now

$$y = X\boldsymbol{\beta} + u \qquad \text{with} \qquad u \sim N(\mathbf{0}, \sigma^2\boldsymbol{\Omega}) \tag{5.19}$$

where $\boldsymbol{\Omega}$ is a positive definite matrix of order n. This model is referred to as the case of nonspherical disturbances, compared with $\text{var}(u) = \sigma^2 I$, which is the case of spherical disturbances. For the present the elements of $\boldsymbol{\Omega}$ will be assumed to be known. For example, if the disturbance variance at each sample point is proportional to the square of one of the regressors, say, X_2, we have

$$\text{var}(u_i) = \sigma_i^2 = \sigma^2 X_{2i}^2 \qquad i = 1, 2, \ldots, n$$

where σ^2 is a scaling factor. The variance-covariance matrix of the disturbance is then

$$\text{var}(u) = \sigma^2 \begin{bmatrix} X_{21}^2 & 0 & \cdots & 0 \\ 0 & X_{22}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_{2n}^2 \end{bmatrix} = \sigma^2 \text{diag}\{X_{21}^2 \quad X_{22}^2 \quad \cdots \quad X_{2n}^2\}$$

From Eq. (5.19), the multivariate normal density for u is

$$f(u) = (2\pi)^{-n/2}|\sigma^2\boldsymbol{\Omega}|^{-1/2}\exp\left[-\tfrac{1}{2}u'(\sigma^2\boldsymbol{\Omega})^{-1}u\right]$$

Noting that $|\sigma^2\boldsymbol{\Omega}| = \sigma^{2n}|\boldsymbol{\Omega}|$, we may rewrite the density as

$$f(u) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}|\boldsymbol{\Omega}|^{-1/2}\exp[(-1/2\sigma^2)u'\boldsymbol{\Omega}^{-1}u]$$

The log-likelihood is then

$$l = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\ln|\boldsymbol{\Omega}| - \frac{1}{2\sigma^2}(y - X\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(y - X\boldsymbol{\beta}) \tag{5.20}$$

Differentiating with respect to $\boldsymbol{\beta}$ and σ^2 gives

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}(X'\boldsymbol{\Omega}^{-1}y - X'\boldsymbol{\Omega}^{-1}X\boldsymbol{\beta})$$

and
$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(y - X\boldsymbol{\beta})$$

Setting the partial derivatives to zero gives the ML estimators

$$\hat{\boldsymbol{\beta}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \tag{5.21}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\boldsymbol{\beta}})'\Omega^{-1}(y - X\hat{\boldsymbol{\beta}}) \tag{5.22}$$

These estimators are, of course, only operational if the $\boldsymbol{\Omega}$ matrix is known.

5.4.1 Generalized Least Squares

Since $\boldsymbol{\Omega}$ is positive definite, its inverse is positive definite. Thus it is possible to find a nonsingular matrix $\boldsymbol{P}$ such that

$$\boldsymbol{\Omega}^{-1} = \boldsymbol{P'P} \tag{5.23}$$

Substitution in Eq. (5.21) gives

$$\hat{\boldsymbol{\beta}} = (X'P'PX)^{-1}X'P'Py = [(PX)'(PX)]^{-1}(PX)'(Py)$$

This is exactly the vector of estimated coefficients that would be obtained from the OLS regression of the vector $\boldsymbol{Py}$ on the matrix $\boldsymbol{PX}$. These are transformations of the original y, X data. This provides an alternative way of looking at the maximum likelihood estimator of the nonspherical model.

Premultiply the linear model, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, by a nonsingular matrix, $\boldsymbol{P}$, satisfying Eq. (5.23), to obtain

$$y_* = X_*\boldsymbol{\beta} + u_* \tag{5.24}$$

where $\boldsymbol{y}_* = \boldsymbol{Py}$, $\boldsymbol{X}_* = \boldsymbol{PX}$, and $\boldsymbol{u}_* = \boldsymbol{Pu}$. It follows from Eq. (5.23) that $\boldsymbol{\Omega} = \boldsymbol{P}^{-1}(\boldsymbol{P}')^{-1}$. Then

$$\begin{aligned} \mathrm{var}(\boldsymbol{u}_*) &= E(\boldsymbol{Puu'P'}) \\ &= \sigma^2\boldsymbol{P\Omega P'} \\ &= \sigma^2\boldsymbol{PP}^{-1}(\boldsymbol{P}')^{-1}\boldsymbol{P}' \\ &= \sigma^2\boldsymbol{I} \end{aligned}$$

Thus the transformed variables in Eq. (5.24) satisfy the conditions under which OLS is BLUE. The coefficient vector from the OLS regression of y_* on X_* is the **generalized least squares** (GLS) estimator.

$$\begin{aligned} \boldsymbol{b}_{\mathrm{GLS}} &= (X_*'X_*)^{-1}X_*'y_* \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \end{aligned} \tag{5.25}$$

This is seen to be the ML estimator already defined in Eq. (5.21). From OLS theory it follows directly that

$$\begin{aligned} \mathrm{var}(\boldsymbol{b}_{\mathrm{GLS}}) &= \sigma^2(X_*'X_*)^{-1} \\ &= \sigma^2(X'\Omega^{-1}X)^{-1} \end{aligned} \tag{5.26}$$

This is also the asymptotic variance matrix that would be yielded by the ML approach. An unbiased estimate of the unknown σ^2 in Eq. (5.26) is readily obtained from the application of OLS to the transformed model. It is

$$\begin{aligned} s^2 &= (y_* - X_* b_{GLS})'(y_* - X_* b_{GLS})/(n-k) \\ &= [P(y - Xb_{\text{GLS}})]'\,[P(y - Xb_{\text{GLS}})]/(n-k) \\ &= (y - Xb_{\text{GLS}})'\Omega^{-1}(y - Xb_{\text{GLS}})/(n-k) \end{aligned} \tag{5.27}$$

This differs from the biased ML estimator in Eq. (5.22) by the factor $n/(n-k)$.

Finally, since Eq. (5.24) satisfies the conditions for the application of OLS, an **exact, finite sample test of** the linear restrictions

$$H_0\colon R\beta = r$$

can be based on

$$F = \frac{(r - Rb_{\text{GLS}})'[R(X'\Omega^{-1}X)^{-1}R']^{-1}(r - Rb_{\text{GLS}})/q}{s^2} \tag{5.28}$$

having the $F(q, n-k)$ distribution under the null, and s^2 is defined in Eq. (5.27). There are many important practical applications of GLS, particularly in the areas of heteroscedasticity and autocorrelation, which are the subjects of the next chapter. However, we note that the procedures outlined so far imply knowledge of Ω. In practice this condition is rarely satisfied, and it is important to develop **feasible generalized least squares** (FGLS) estimators, where unknown parameters are replaced by consistent estimates. Examples of this important technique will be given in the next chapter.

Rather than writing $u \sim N(0, \sigma^2\Omega)$ as in Eq. (5.19), it is sometimes more convenient to use the specification $u \sim N(0, V)$, where V is a positive definite, variance-covariance matrix. This no longer separates out the scale factor σ^2. With the alternative specification, it follows directly that

$$\begin{aligned} b_{\text{GLS}} &= (X'V^{-1}X)^{-1}X'V^{-1}y \\ \text{var}(b_{\text{GLS}}) &= (X'V^{-1}X)^{-1} \end{aligned} \tag{5.29}$$

5.5 INSTRUMENTAL VARIABLE (IV) ESTIMATORS

Under the classical assumptions OLS estimators are best linear unbiased. One of the major underpinning assumptions is the independence of regressors from the disturbance term. If this condition does not hold, OLS estimators are *biased* and *inconsistent*. This statement may be illustrated by a simple **errors in variables** example.

Consider the relation

$$y = \beta x + u \tag{5.30}$$

where, for simplicity, the constant term has been dropped. So far we have implicitly assumed that variables are measured without error. Suppose, however, that the observed value x can be represented as the sum of the true value $\tilde{x}$ and a random measurement error v, that is,

$$x = \tilde{x} + v$$

In this case the appropriate relation may be

$$y = \beta\tilde{x} + u \tag{5.31}$$

In practice one has to think carefully about whether Eq. (5.30) or Eq. (5.31) is the relevant specification. If, for example, traders in the bond market respond to last quarter's *reported* GDP, then the analysis and forecasting of traders' behavior requires the use of Eq. (5.30), and the problems to be discussed in this section do not arise. For many economic variables the *true* value is an elusive and unattainable concept. Most values only become definitive when the statisticians cease to revise them.

Measurement error in y need not be modeled separately since, if present, it can be merged with the disturbance (equation error) u. If we assume that Eq. (5.31) is the maintained specification but that observations are only available on x and not on $\tilde{x}$, what happens if we use OLS to estimate β? The OLS slope is

$$\begin{aligned} b &= \frac{\sum yx}{\sum x^2} \\ &= \frac{\sum x(\beta\tilde{x} + u)}{\sum x^2} \\ &= \beta\frac{\sum x\tilde{x}}{\sum x^2} + \frac{\sum xu}{\sum x^2} \end{aligned} \tag{5.32}$$

It is assumed that u, $\tilde{x}$, and v are mutually independent, and that appropriate second-order moments and their probability limits exist. It then follows that

$$\operatorname{plim}\left(\frac{1}{n}\sum x^2\right) = \sigma_{\tilde{x}}^2 + \sigma_v^2$$

$$\operatorname{plim}\left(\frac{1}{n}\sum x\tilde{x}\right) = \sigma_{\tilde{x}}^2$$

$$\operatorname{plim}\left(\frac{1}{n}\sum xu\right) = 0$$

Substitution in (5.32) gives

$$\operatorname{plim} b = \beta\left(\frac{\sigma_{\tilde{x}}^2}{\sigma_{\tilde{x}}^2 + \sigma_v^2}\right) \tag{5.33}$$

Thus OLS is biased and inconsistent, with a probability limit numerically less than β. Therefore, whether β is positive or negative, the probability limit of the OLS slope is closer to zero than the true slope, which is called **attenuation bias.** This is an example of a specification error. By assumption the relevant model is Eq. (5.31) but for data reasons we have had to use Eq. (5.30). The result is a flawed estimation procedure.

It is enlightening to derive this result in an alternative fashion. Equation (5.31) may be rewritten as

$$y = \beta x + (u - \beta v)$$

which shows that if we model y as a function of x, the transformed disturbance contains the measurement error in x. We may then write

$$b = \beta + \frac{\sum x(u - \beta v)}{\sum x^2} \tag{5.34}$$

Then $\operatorname{plim}\frac{1}{n}\sum x(u - \beta v) = \operatorname{plim}\frac{1}{n}\sum xu - \beta \operatorname{plim}\frac{1}{n}\sum xv = -\beta\sigma_v^2$

The regressor and the transformed disturbance are correlated. Substitution in Eq. (5.34) gives

$$\operatorname{plim} b = \beta - \frac{\beta\sigma_v^2}{\sigma_{\tilde{x}}^2 + \sigma_v^2} = \beta\left(\frac{\sigma_{\tilde{x}}^2}{\sigma_{\tilde{x}}^2 + \sigma_v^2}\right)$$

as before.

Returning to the general linear model, $y = X\boldsymbol{\beta} + \boldsymbol{u}$, the OLS estimator is

$$\boldsymbol{b} = \boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{u}$$

giving $$\operatorname{plim}\boldsymbol{b} = \boldsymbol{\beta} + \operatorname{plim}\left(\frac{1}{n}X'X\right)^{-1} \cdot \operatorname{plim}\left(\frac{1}{n}X'\boldsymbol{u}\right)$$

If we assume that $\operatorname{plim}(X'X/n) = \boldsymbol{\Sigma}_{XX}$, a positive definite matrix of full rank, and $\operatorname{plim}(X'\boldsymbol{u}/n) = \boldsymbol{\Sigma}_{X\boldsymbol{u}} \neq \mathbf{0}$, then

$$\operatorname{plim}\boldsymbol{b} = \boldsymbol{\beta} + \boldsymbol{\Sigma}_{XX}^{-1} \cdot \boldsymbol{\Sigma}_{X\boldsymbol{u}} \tag{5.35}$$

so that correlation of the disturbance term with one or more of the regressors renders the OLS estimates inconsistent. As we have seen, such correlations can be caused by measurement error in one or more regressors. In Chapter 6 it will be shown that the combination of an autocorrelated disturbance and one or more lagged values of the dependent variable among the regressors will also produce such correlations. In Chapter 9 it will be seen that structural simultaneous models give rise to the same condition. It is therefore important to seek consistent estimators.

A consistent estimator may be obtained by the use of **instrumental variables,** which are also commonly referred to as **instruments.** We still postulate the model $y = X\boldsymbol{\beta} + \boldsymbol{u}$, with $\operatorname{var}(\boldsymbol{u}) = \sigma^2\boldsymbol{I}$, but we now assume $\operatorname{plim}(X'\boldsymbol{u}/n) \neq \mathbf{0}$. Suppose that it is possible to find a data matrix Z of order $n \times l$ ($l \geq k$), which possesses two vital properties:

1. The variables in Z are correlated with those in X, and in the limit $\operatorname{plim}(Z'X/n) = \boldsymbol{\Sigma}_{ZX}$, a finite matrix of full rank.
2. The variables in Z are in the limit uncorrelated with the disturbance term $\boldsymbol{u}$, that is, $\operatorname{plim}(Z'\boldsymbol{u}/n) = \mathbf{0}$.

Premultiplying the general relation by Z' gives

$$Z'y = Z'X\beta + Z'u \qquad \text{with} \qquad \operatorname{var}(Z'u) = \sigma^2(Z'Z) \tag{5.36}$$

This suggests the use of GLS. The resultant estimator is

$$\begin{aligned} b_{\text{GLS}} = b_{\text{IV}} &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= (X'P_ZX)^{-1}X'P_Zy \end{aligned} \tag{5.37}$$

where $P_Z = Z(Z'Z)^{-1}Z'$. The variance-covariance matrix is

$$\operatorname{var}(b_{\text{IV}}) = \sigma^2(X'P_ZX)^{-1} \tag{5.38}$$

and the disturbance variance may be estimated consistently from

$$\hat{\sigma}^2 = (y - Xb_{\text{IV}})'(y - Xb_{\text{IV}})/n \tag{5.39}$$

The use of n or $n-k$ or $n-l$ in the divisor here does not matter asymptotically. The consistency of the IV estimator may be seen as follows. From Eq. (5.37)

$$b_{\text{IV}} = \beta + \left(\frac{1}{n}X'P_ZX\right)^{-1}\left(\frac{1}{n}X'P_Zu\right)$$

Now
$$\frac{1}{n}X'P_ZX = \left(\frac{1}{n}X'Z\right)\left(\frac{1}{n}Z'Z\right)^{-1}\left(\frac{1}{n}Z'X\right) \tag{5.40}$$

If we assume the middle term to have probability limit Σ_{ZZ}^{-1} it follows that

$$\operatorname{plim}\left(\frac{1}{n}X'P_ZX\right) = \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$$

which will be a finite. nonsingular matrix. Similarly,

$$\operatorname{plim}\left(\frac{1}{n}XP_Zu\right) = \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{Zu} = \mathbf{0}$$

since the instruments are assumed to be uncorrelated in the limit with the disturbance. The IV estimator is thus consistent.

5.5.1 Special Case

When $l = k$, that is, when Z contains the same number of columns as X, we have a special case of the foregoing results. Now $X'Z$ is $k \times k$ and nonsingular. The estimator in Eq. (5.37) simplifies to

$$b_{\text{IV}} = (Z'X)^{-1}Z'y \tag{5.41}$$

with
$$\operatorname{var}(b_{\text{IV}}) = \sigma^2(Z'X)^{-1}(Z'Z)(X'Z)^{-1} \tag{5.42}$$

Note, however, that we must have at least as many instruments as there are columns in X. If that condition were not satisfied, the matrix in Eq. (5.40) would have rank $l < k$ and so would be singular.

5.5.2 Two-Stage Least Squares (2SLS)

The IV estimator may also be seen as the result of a double application of least squares.

Stage (*i*): Regress each of the variables in the X matrix on Z to obtain a matrix of fitted values $\hat{X}$,

$$\hat{X} = Z(Z'Z)^{-1}Z'X = P_Z X \tag{5.43}$$

Stage (*ii*): Regress y on $\hat{X}$ to obtain the estimated $\boldsymbol{\beta}$ vector

$$\begin{aligned} b_{2\text{SLS}} &= (\hat{X}'\hat{X})^{-1}(\hat{X}'y) \\ &= (X'P_Z X)^{-1}(X'P_Z y) \\ &= b_{\text{IV}} \end{aligned} \tag{5.44}$$

Thus the IV estimator can be obtained by a two-stage least-squares procedure. The variance matrix and the estimated disturbance variance are given in Eqs. (5.38) and (5.39).

5.5.3 Choice of Instruments

The crucial question is, where do we find these useful instruments? Some of them are often variables from the X matrix itself. Any variables that are thought to be exogenous and independent of the disturbance are retained to serve in the Z matrix. In dynamic analyses, as will be seen in Chapter 8, *lagged* variables can be used as instruments for *current* values. When some of the X variables are used as instruments, we may partition X and Z as

$$X = [X_1 \quad X_2] \qquad Z = [X_1 \quad Z_1]$$

where X_1 is of order $n \times r$ $(r < k)$, X_2 is $n \times (k-r)$, and Z_1 is $n \times (l-r)$. It can be shown[6] that $\hat{X}$, the matrix of regressors in the second-stage regression, is then

$$\hat{X} = [X_1 \quad \hat{X}_2] \qquad \text{where } \hat{X}_2 = Z(Z'Z)^{-1}Z'X_2 \tag{5.45}$$

The variables in X_1 serve as instruments for themselves, and the remaining second-stage regressors are the fitted values of X_2, obtained from the regression of X_2 on the *full* set of instruments. There still remains the question of how many instruments to use. The *minimum* number is k, including any variables that serve as their own instruments. The asymptotic efficiency increases with the number of instruments. However, the finite sample bias also increases with the number of instruments. If, in fact, we select n instruments, it is simple to show that $P_Z = I$, in which case the IV estimator is simply OLS, which is biased and inconsistent. If, on the other hand, we use the minimum, or close to the minimum, number of instruments, the results may also be poor. It has been shown[7] that the mth moment of the 2SLS estimator exists if and only if $m < l-k+1$. Thus, if there are just as many instruments as explanatory

[6] See Problem 5.9.

[7] T. W. Kinal, "The Existence of Moments of k-Class Estimators," *Econometrica*, **48**, 1980, 241–249.

variables, the 2SLS estimator will not have a mean. With one more instrument there will be a mean but no variance, and so forth.[8]

5.5.4 Tests of Linear Restrictions

We frequently need to test the usual kinds of linear restrictions on an equation that has been estimated by the IV (2SLS) method. This can be done by familiar-looking methods, but there are two important qualifications. First, the test procedures only have asymptotic validity; and, second, one has to be very careful about the definition and calculation of the residual sums of squares, which appear in the test statistics. We consider the usual linear model, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ with H_0: $\boldsymbol{R\beta} = \boldsymbol{r}$. We will assume that the first-stage regression of $\boldsymbol{X}$ on $\boldsymbol{Z}$ has been completed, giving the matrix of fitted values $\hat{\boldsymbol{X}} = \boldsymbol{P}_Z X$. The test procedure is as follows:

1. Regress $\boldsymbol{y}$ on $\hat{\boldsymbol{X}}$, *imposing the restrictions*. Denote the resultant coefficient vector by $\boldsymbol{b}_{\text{res}}$, and the vector of residuals by

$$\boldsymbol{e}_{\text{r}} = \boldsymbol{y} - \hat{\boldsymbol{X}}\boldsymbol{b}_{\text{res}} \tag{5.46}$$

2. Regress $\boldsymbol{y}$ on $\hat{\boldsymbol{X}}$, without restrictions. Denote the resultant coefficient vector by $\boldsymbol{b}_{\text{unres}}$ and the vector of residuals by

$$\boldsymbol{e}_{\text{ur}} = \boldsymbol{y} - \hat{\boldsymbol{X}}\boldsymbol{b}_{\text{unres}} \tag{5.47}$$

3. Using $\boldsymbol{b}_{\text{unres}}$ compute the vector

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_{\text{unres}} \tag{5.48}$$

where the actual X values have now been used, rather than the fitted values.

The relevant test statistic for H_0 is then

$$F = \frac{(\boldsymbol{e}_{\text{r}}'\boldsymbol{e}_{\text{r}} - \boldsymbol{e}_{\text{ur}}'\boldsymbol{e}_{\text{ur}})/q}{\boldsymbol{e}'\boldsymbol{e}/(n-k)} \overset{a}{\sim} F(q, n-k) \tag{5.49}$$

A detailed derivation of these results is available in Davidson and MacKinnon.[9]

APPENDIX

APPENDIX 5.1
Change of variables in density functions

The univariate case has already been dealt with in Appendix 2.1. In the multivariate case $\boldsymbol{u}$ and $\boldsymbol{y}$ now indicate vectors of, say, n variables each. The multivariate

[8] For a horror story of poor IV performance when $l = k = 1$ and there is low correlation between the instrument and the single explanatory variable, see the two articles by C. R. Nelson and R. Startz, "The Distribution of the Instrumental Variable Estimator and Its t-Ratio When the Instrument Is a Poor One," *Journal of Business*, **63**, 1990, S125–S140; and "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator," *Econometrica*, **58**, 1990, 967–976.

[9] Russell Davidson and James G. MacKinnon, *op. cit.*, 215–232.

extension of the previous result is

$$f(y) = f(u)\left|\frac{\partial u}{\partial y}\right|$$

where $|\partial u/\partial y|$ indicates the absolute value of the determinant formed from the matrix of partial derivatives,

$$\begin{bmatrix} \frac{\partial u_1}{\partial y_1} & \frac{\partial u_1}{\partial y_2} & \cdots & \frac{\partial u_1}{\partial y_n} \\ \frac{\partial u_2}{\partial y_1} & \frac{\partial u_2}{\partial y_2} & \cdots & \frac{\partial u_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial u_n}{\partial y_1} & \frac{\partial u_n}{\partial y_2} & \cdots & \frac{\partial u_n}{\partial y_n} \end{bmatrix}$$

The absolute value of this determinant is known as the **Jacobian of the transformation from *u* to *y*.**

APPENDIX 5.2
Centered and uncentered R^2

From Chapter 3 the OLS regression may be written as

$$y = Xb + e = \hat{y} + e$$

Squaring gives

$$y'y = \hat{y}'\hat{y} + e'e$$

since $\hat{y}'e = b'X'e = 0$. Substituting for b gives

$$y'y = y'X(X'X)^{-1}X'y + e'e \tag{A 5.1}$$

The *uncentered* R^2 is defined as

$$\text{uncentered } R^2 = \frac{y'X(X'X)^{-1}X'y}{y'y} \tag{A 5.2}$$

This is called uncentered because the denominator is $y'y = \sum_{t=1}^{n} Y_t^2$, which is the total variation in the dependent variable, measured about the origin. The *uncentered* R^2 thus indicates the proportion of this total variation "explained" by the regression.

In many economic and other applications, interest is focused on the variation of the dependent variable *about its mean level,* rather than its variation about zero. Then the quantity to be explained is $\sum_{t=1}^{n}(Y_t - \bar{Y})^2 = \sum_{t=1}^{n} Y_t^2 - (\sum_{t=1}^{n} Y)^2/n = y'y - (i'y)^2/n$. Subtracting the correction for the mean, $(i'y)^2/n$, from both sides of (A 5.1) gives

$$y'y - (i'y)^2/n = [y'X(X'X)^{-1}X'y - (i'y)^2/n] + e'e$$

This expression gives the decomposition of the total sum of squares (TSS) about the mean, into the sum of the explained sum of squares (ESS), the term in brackets, and the residual, or unexplained, sum of squares (RSS). The *centered* R^2 is then

$$\text{centered } R^2 = \frac{y'X(X'X)^{-1}X'y - (i'y)^2/n}{y'y - (i'y)^2/n} \tag{A 5.3}$$

This R^2 statistic is already defined in Eqs. (3.9) and (3.10), although the correspondence is not immediately obvious. To show the correspondence we need to show that the TSS and ESS defined in Eq. (3.9) are given by the denominator and numerator of Eq. (A 5.3). The denominator in Eq. (3.10) is

$$y_*'y_* = y'Ay = y'y - y'\left(\frac{1}{n}ii'\right)y = y'y - (i'y)^2/n$$

which is the denominator in Eq. (A 5.3). The ESS in Eq. (3.9) is $b_*'X_*'X_*b_*$. To see the connection with Eq. (A 5.3) return to the OLS regression in the raw data,

$$y = Xb + e$$

Premultiply by the deviation-producing matrix A, defined in Eq. (3.7). The result is

$$Ay = AXb + Ae = AXb + e$$

Squaring each side gives

$$y'Ay = b'X'AXb + e'e$$

Thus the ESS may now be expressed as

$$\begin{aligned}
\text{ESS} &= b'X'AXb \\
&= y'X(X'X)^{-1}X'AX(X'X)^{-1}X'y \\
&= y'X(X'X)^{-1}X'y - y'X(X'X)^{-1}X'\left(\frac{1}{n}ii'\right)X(X'X)^{-1}X'y \\
&= y'X(X'X)^{-1}X'y - (i'y)^2/n
\end{aligned}$$

which is the numerator in Eq. (A 5.3). The last step in this proof is based on the fact that $X(X'X)^{-1}X'i = i$. Since i is the first column in X, the product $(X'X)^{-1}X'i$ gives the first column in I_k. Premultiplying by X gives the first column of the X matrix, which is i.

APPENDIX 5.3
To show that $e_*'X(X'X)^{-1}X'e_* = e_*'e_* - e'e$

Consider $e_*'e_* - e_*'X(X'X)^{-1}X'e_* = e_*'Me_*$. Thus we need to show that $e_*'Me_* = e'e$, which will be true if $e = Me_*$. We have

$$e_* = y - Xb_*$$

where b_* is the restricted estimator satisfying $Rb_* = r$. Thus

$$Me_* = My = e$$

since $MX = \mathbf{0}$, which completes the proof.

PROBLEMS

5.1. Consider the binomial variable y, which takes on the values zero or one according to the probability density function (pdf)

$$f(y) = \theta^{y}(1-\theta)^{(1-y)} \qquad 0 \le \theta \le 1 \qquad y = 0, 1$$

Thus the probability of a "success" ($y = 1$) is given by $f(1) = \theta$, and the probability of a "failure" ($y = 0$) is given by $f(0) = 1 - \theta$. Verify that $E(y) = \theta$, and $\text{var}(y) = \theta(1-\theta)$. If a random sample of n observations is drawn from this distribution, find the MLE of θ and the variance of its sampling distribution. Find the asymptotic variance of the MLE estimator, using each of the two expressions for the information matrix in Eq. (5.3).

5.2. The pdf of the **uniform distribution** is given by

$$f(x \mid \alpha) = \frac{1}{\alpha} \qquad 0 < x < \alpha$$

Find the MLE of α.

5.3. When s successes occur in n independent trials, where the probability of a success is θ, the sample proportion, $p = s/n$, has been shown in Problem 5.1 to be the MLE of θ. Consider an alternative estimator,

$$p^* = \frac{s+1}{n+2}$$

Find the mean and variance of p^* in terms of the mean and variance of p. Is p^* a consistent estimator of θ?

5.4. A sample of three values $x = 1, 2, 3$ is drawn from the exponential distribution with the following pdf:

$$f(x) = \frac{1}{\theta}e^{-x/\theta}$$

Derive the ML *estimator* of θ, and compute the ML *estimate* for these sample values.

5.5. Prove that maximizing the log-likelihood in Eq. (5.10) yields the restricted LS estimator defined in Eq. (3.43).

5.6. Show that when the restricted least-squares equation is fitted, the residuals e_* do not necessarily have zero sample mean, but that this result will hold if the restrictions involve only the slope coefficients and do not involve the intercept term.

If the general proof at first eludes you, try the two-variable equation $Y = \alpha + \beta X + u$, imposing a restriction on the slope coefficient but none on the intercept, and, alternatively, imposing a restriction on the intercept but none on the slope.

5.7. Using the data of Example 3.3, compute the LR, W, and LM test statistics for H_0: $\beta_3 = -1$. Verify the LM statistic by computing the residual from the restricted regression and regressing it on X_2 and X_3.

5.8. Repeat Problem 5.7 for H_0: $\beta_2 + \beta_3 = 0$.

5.9. Prove the assertion in Eq. (5.45).

CHAPTER 6

Heteroscedasticity and Autocorrelation

Two of the major applications of the maximum likelihood and generalized least-squares procedures outlined in Chapter 5 occur in the estimation and testing of relations with heteroscedastic and/or autocorrelated disturbances. We will deal with the problems in that order.

When heteroscedasticity alone occurs, the variance matrix for the disturbance vector is

$$\operatorname{var}(\boldsymbol{u}) = E(\boldsymbol{uu}') = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} = \boldsymbol{V} \tag{6.1}$$

There are now $n + k$ unknown parameters; n unknown variances; and k elements in the $\boldsymbol{\beta}$ vector. Without some additional assumptions, estimation from n sample points is clearly impossible. Additional assumptions are usually made about the disturbance process. Heteroscedasticity is most commonly expected in cross-section data. For example, in a cross-section study of the relation between household vacation expenditure and household income, one would expect the *average* expenditure for a given income level to increase with income, but one might also expect the *variation* about average expenditure to increase as income increases. Suppose, for instance, that we adopt the formal hypothesis

$$\sigma_i^2 = \sigma^2 x_{2i} \qquad i = 1, 2, \dots, n$$

where σ^2 is a scale factor and x_2 denotes the income variable. Income is thus assumed to be both an explanatory variable in the expenditure equation and the causal factor in the heteroscedasticity process. The variance matrix for the disturbance vector is now

$$\text{var}(\boldsymbol{u}) = E(\boldsymbol{uu}') = \sigma^2 \begin{bmatrix} x_{21} & 0 & \cdots & 0 \\ 0 & x_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_{2n} \end{bmatrix} = \sigma^2 \boldsymbol{\Omega} \tag{6.2}$$

Specification (6.2) has only one unknown parameter, compared with n unknown parameters in Eq. (6.1). However, Eq. (6.2) is a very strong assumption and the heteroscedasticity process may not be that simple. It is therefore important to test for the possible presence of heteroscedasticity and, if found, to explore its structure in order to derive feasible GLS estimators of the equation of interest. These topics will be dealt with in Sections 6.2 and 6.3, but first we will examine the properties of OLS estimators, if they are applied to an equation with nonspherical disturbances.

6.1 PROPERTIES OF OLS ESTIMATORS

The specified equation is

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} \qquad \text{with } E(\boldsymbol{u}) = \boldsymbol{0} \qquad \text{and} \qquad E(\boldsymbol{uu}') = \sigma^2\boldsymbol{\Omega}$$

For nonstochastic X the following results hold.

1. OLS estimator is unbiased and consistent.

   From Eq. (3.23)

   $$\boldsymbol{b} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \qquad b = \beta + (X'X)^{-1}X'\epsilon$$

   It follows directly that $E(\boldsymbol{b}) = \boldsymbol{\beta}$, so the unbiased property holds. Mean square consistency follows provided the variance matrix, var($\boldsymbol{b}$), has a zero probability limit, as will be seen in Point 3.
2. OLS estimator is inefficient.

   Equation (5.25) showed that the GLS estimator, which consists of the regression of a *transformed* y vector on a *transformed* X matrix, gives a best linear unbiased estimator. Thus, OLS, which regresses *untransformed* variables, produces linear unbiased but not minimum variance estimators.
3. Conventional OLS coefficient standard errors are incorrect, and the conventional test statistics based on them are invalid.

   The correct variance matrix for the OLS coefficient vector is

   $$\begin{aligned} \text{var}(\boldsymbol{b}) &= E[(\boldsymbol{b}-\boldsymbol{\beta})(\boldsymbol{b}-\boldsymbol{\beta})'] \\ &= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{uu}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}] \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \quad \neq \sigma^2(X'X)^{-1} \end{aligned} \tag{6.3}$$

   The conventional formula calculates $\sigma^2(X'X)^{-1}$, which is only part of the correct expression in Eq. (6.3). Thus the conventional test statistics are invalidated. The variance matrix may also be expressed as

   $$\text{var}(\boldsymbol{b}) = \frac{\sigma^2}{n}\left[\frac{1}{n}(\boldsymbol{X}'\boldsymbol{X})\right]^{-1}\left[\frac{1}{n}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})\right]\left[\frac{1}{n}(\boldsymbol{X}'\boldsymbol{X})\right]^{-1} \tag{6.4}$$

The probability limit of the first term in Eq. (6.4) is zero. With stationary regressors the probability limit of the second term is a finite matrix. Consistency thus requires that the probability limit of $X'\Omega X/n$ also be a finite matrix, which in general will be true if the elements of Ω are finite. If the X matrix contains one or more lags of the dependent variable, the OLS estimator will have a finite sample bias; but it will still be consistent as long as V is diagonal, as in Eq. (6.1). That autocorrelated disturbances cause off-diagonal terms in V to be nonzero will be discussed in Section 6.5. When autocorrelation is combined with the fact of one or more regressors being lags of the dependent variable, consistency no longer holds.

Even if one suspects heteroscedasticity, one may wish to proceed with OLS in spite of the inefficiency. Valid inferences, however, would then require implementation of Eq. (6.3), with $\sigma^2\Omega = \text{diag}\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\}$. The problem of estimating $\sigma^2\Omega$ seems impossible because it contains n parameters and we only have n observations. However, White has shown in a very influential article that this way of looking at the problem is misleading. What matters is obtaining a satisfactory estimate of $X'\sigma^2\Omega X$, which is a square matrix of order k; and k (the number of regressors) is a constant, independent of the sample size n.[1] The nature of the White estimator is most easily seen by rewriting $X'\sigma^2\Omega X$ in an alternative form. Let y_t denote the tth observation on the dependent variable, and $x_t' = [1 \; x_{2t} \; \cdots \; x_{kt}]$ denote the tth row of the X matrix. Then

$$
\begin{aligned}
X'\sigma^2\Omega X &= \begin{bmatrix} \vdots & \vdots & & \vdots \\ x_1 & x_2 & \cdots & x_n \\ \vdots & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \begin{bmatrix} \cdots & x_1' & \cdots \\ \cdots & x_2' & \cdots \\ & \vdots & \\ \cdots & x_n' & \cdots \end{bmatrix} \\
&= \sum_{t=1}^{n} \sigma_t^2 x_t x_t'
\end{aligned}
\tag{6.5}
$$

The White estimator replaces the unknown $\sigma_t^2 (t = 1, 2, \ldots, n)$ by e_t^2, where the e_t denote the OLS residuals, $y_t - x_t'b \; (t = 1, 2, \ldots, n)$. This provides a consistent estimator of the variance matrix for the OLS coefficient vector and is particularly useful because it does not require any specific assumptions about the form of the heteroscedasticity. The empirical implementation of Eq. (6.3) is then

$$
\begin{aligned}
\text{est. var}(b) &= (X'X)^{-1}X'\sigma^2\hat{\Omega}X(X'X)^{-1} \\
\sigma^2\hat{\Omega} &= \text{diag}\{e_1^2, e_2^2, \ldots, e_n^2\}
\end{aligned}
\tag{6.6}
$$

The square roots of the elements on the principal diagonal of est. var(b) are the estimated standard errors of the OLS coefficients. They are often referred to as heteroscedasticity-consistent standard errors (HCSEs). The usual t and F tests are

[1] Halbert White, "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity," *Econometrica*, **48**, 1980, 817–838.

now only valid asymptotically. General linear hypotheses may be tested by the Wald statistic,

$$W = (\boldsymbol{Rb} - \boldsymbol{r})' \left\{ \boldsymbol{R}[\text{est. var}(\boldsymbol{b})]\boldsymbol{R}' \right\}^{-1} (\boldsymbol{Rb} - \boldsymbol{r}) \overset{a}{\sim} \chi^2(q) \tag{6.7}$$

There are two crucial questions:

1. What is the difference between the conventional and the correct estimates of the standard errors of the OLS coefficients?
2. What is the difference between the correct OLS standard errors and the GLS standard errors?

Davidson and MacKinnon provide some Monte Carlo evidence on these questions. Their model is

$$y_t = 1 + x_t + u_t \qquad u_t \sim N(0,\ x_t^{\alpha})$$

with $n = 100$, x_t uniformly distributed between 0 and 1, and α a parameter that takes on various values. For each specified value of α they drew 20,000 samples of 100 observations and calculated the OLS and GLS estimates of the intercept and slope. The standard deviations of these estimates give the (correct) OLS and GLS standard errors. The incorrect OLS standard errors are calculated from the conventional formula. A selection of their results is shown in Table 6.1.[2] For the intercept, the incorrect OLS standard errors are greater than the correct values. There is little difference between the correct and incorrect slope standard errors, except for the largest value of α. The inefficiency of OLS is shown by the contrast between the correct OLS standard errors and the GLS standard errors. The inefficiency increases substantially with α. These results, of course, are only illustrative and depend on the specification of the experiment.

The White procedure has large-sample validity. It may not work very well in finite samples. There is some evidence that corrections to e_t^2 can improve finite sample performance.[3] One correction is to replace e_t^2 by $ne_t^2/(n-k)$. A better correction

TABLE 6.1
Correct and incorrect standard errors

	OLS intercept			OLS slope		
α	Incorrect	Correct	GLS intercept	Incorrect	Correct	GLS slope
0.5	0.164	0.134	0.110	0.285	0.277	0.243
1.0	0.142	0.101	0.048	0.246	0.247	0.173
2.0	0.116	0.074	0.0073	0.200	0.220	0.109
3.0	0.100	0.064	0.0013	0.173	0.206	0.056

[2] Reprinted by permission from Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics,* Oxford University Press, 1993, 550.

[3] Davidson and MacKinnon, ibid., 554.

is to use $e_t^2/(1 - h_t)$, where $h_t = \boldsymbol{x}_t'(X'X)^{-1}\boldsymbol{x}_t$. The rationale of this correction may be shown as follows. From the development leading up to Eq. (3.17) it is seen that

$$\boldsymbol{e} = \boldsymbol{My}$$

where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, which is a symmetric, idempotent matrix with the properties $\boldsymbol{MX} = \boldsymbol{0}$ and $\boldsymbol{Me} = \boldsymbol{e}$. By assuming homoscedasticity, the variance matrix of the OLS residual vector is

$$E(\boldsymbol{ee}') = E(\boldsymbol{Muu}'\boldsymbol{M}) = \sigma^2\boldsymbol{M} \tag{6.8}$$

The tth element on the principal diagonal of the matrices in Eq. (6.8) gives

$$E(e_t^2) = \sigma^2(1 - \boldsymbol{x}_t'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_t)$$

The mean squared residual thus underestimates σ^2, which suggests the second correction given in this paragraph. The term h_t is the tth diagonal element in $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. This matrix is referred to as the **hat matrix**, because it is the matrix that premultiplies $\boldsymbol{y}$ to give the predicted values $\hat{\boldsymbol{y}} = \boldsymbol{Xb} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

6.2 TESTS FOR HETEROSCEDASTICITY

If the inefficiency of OLS is thought to be a serious drawback, testing for the presence of heteroscedasticity is then desirable. This section reviews four major tests, namely. the White test, the Breusch-Pagan/Godfrey test, the Goldfeld-Quandt test, and a likelihood ratio test for grouped data.

6.2.1 The White Test[4]

This asymptotic test does not require one to specify the variables thought to determine the heteroscedasticity. One simply computes an auxiliary regression of the squared OLS residuals on a constant and all nonredundant variables in the set consisting of the regressors, their squares, and their cross products. Suppose, for example, that

$$\boldsymbol{x}_t' = [1\ x_{2t}\ x_{3t}]$$

In principle there are nine possible variables, but the square of 1 is 1 and the cross product of 1 with each x merely replicates the x variable. Thus the set of nonredundant variables comprising regressors, squares, and cross products is then

$$[1\ x_{2t}\ x_{3t}\ x_{2t}^2\ x_{3t}^2\ x_{2t}x_{3t}]$$

This set already contains a constant, so the auxiliary regression is e_t^2 on these six regressors. On the hypothesis of homoscedasticity, nR^2 is asymptotically distributed as

[4] Halbert White, *op. cit.*

$\chi^2(5)$. The degrees of freedom are the number of variables in the auxiliary regression (excluding the constant). In general, under the null of homoscedasticity,

$$nR^2 \overset{a}{\sim} \chi^2(q)$$

where q is the number of variables in the auxiliary regression less one. If homoscedasticity is rejected, there is no indication of the form of the heteroscedasticity and, thus, no guide to an appropriate GLS estimator. Computing the White standard errors would, however, be wise if one is proceeding with OLS. A final problem with the White test is that the degrees of freedom in the χ^2 test may become rather large, which tends to reduce the power of the test. For instance, if there are k regressors, including a constant, in the original relation, the value of q will in general be $[k(k+1)/2] - 1$. With $k = 10$, $q = 54$. If the regressors include dummy variables, the degrees of freedom will be somewhat smaller. Sometimes ad hoc reductions in q are made by including the squares of the regressors but excluding the cross products.

6.2.2 The Breusch-Pagan/Godfrey Test[5]

This test is an example of an LM test, and the technical details are given in Appendix 6.1. The usual linear relation,

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + u_t \qquad t = 1, 2, \ldots, n \tag{6.9}$$

is postulated, where $\boldsymbol{x}_t' = [1 \; x_{2t} \; x_{3t} \; \cdots \; x_{kt}]$. It is assumed that heteroscedasticity takes the form

$$Eu_t = 0 \qquad \text{for all } t$$
$$\sigma_t^2 = Eu_t^2 = h(\boldsymbol{z}_t'\boldsymbol{\alpha}) \tag{6.10}$$

where $\boldsymbol{z}_t' = [1 \; z_{2t} \; \cdots \; z_{pt}]$ is a vector of known variables, $\boldsymbol{\alpha} = [\alpha_1 \; \alpha_2 \; \cdots \; \alpha_p]$ is a vector of unknown coefficients, and $h(\cdot)$ is some unspecified function that must take on only positive values. The null hypothesis of homoscedasticity is then

$$H_0\colon \alpha_2 = \alpha_3 = \cdots = \alpha_p = 0$$

for this gives $\sigma_t^2 = h(\alpha_1) =$ constant. The *restricted* model under the null is then simply estimated by applying OLS to Eq. (6.9), on the assumption of normally distributed disturbances. Simplicity is what makes the LM test very attractive. The test procedure is as follows:

1. Estimate the original relation, Eq. (6.9), by OLS; obtain the OLS residuals, $e_t = y_t - \boldsymbol{x}_t'\boldsymbol{b}$, and an estimated disturbance variance, $\tilde{\sigma}^2 = \sum e_t^2/n$.
2. Regress $e_t^2/\tilde{\sigma}^2$ on z_t by OLS and compute the explained sum of squares (ESS).
3. Under H_0,

$$\tfrac{1}{2}\text{ESS} \overset{a}{\sim} \chi^2(p-1) \tag{6.11}$$

[5] T. S. Breusch and A. R. Pagan, "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, **47**, 1979, 1287–1294; and L. Godfrey, "Testing for Multiplicative Heteroscedasticity," *Journal of Econometrics*, **8**, 1978, 227–236.

Thus, homoscedasticity is rejected if ESS/2 exceeds the preselected critical value from the χ^2 distribution.

4. As shown in Appendix 6.1, a simpler but asymptotically equivalent procedure is to regress e_t^2 on z_t. Then nR^2 from this regression is asymptotically distributed as $\chi^2(p-1)$ under the null.

This test requires one to know the z variables causing the heteroscedasticity, though not the functional form of the heteroscedasticity. Such knowledge may not be readily available. In practice the candidate variables may be one or more of the regressors already appearing in the x_t vector. In this case the test is essentially the same as an ad hoc version of the White test.

6.2.3 The Goldfeld-Quandt Test[6]

This very simple, finite-sample test is applicable if there is a single variable (typically one of the regressors) that is thought to be an indicator of the heteroscedasticity. Suppose, for instance, that one suspects that σ_t^2 is positively related to the ith regressor, X_i. The test procedure is as follows:

1. Reorder the observations by the value of X_i.
2. Omit c central observations.
3. Fit separate regressions by OLS to the first and last $(n-c)/2$ observations, provided, of course, that $(n-c)/2$ exceeds the number of parameters in the relation.
4. Let RSS_1 and RSS_2 denote the residual sums of squares from the two regressions, with subscript 1 indicating that from the smaller X_i values and 2 that from the larger X_i values. Then

$$R = \frac{\text{RSS}_2}{\text{RSS}_1}$$

will, on the assumption of homoscedasticity, have the F distribution with $[(n-c-2k)/2, (n-c-2k)/2]$ degrees of freedom. Under the alternative hypothesis F will tend to be large. Thus, if $R > F_{.95}$, one would reject the assumption of homoscedasticity at the 5 percent level.

The power of the test will depend, among other things, on the number of central observations excluded. The power will be low if c is too large, so that RSS_1 and RSS_2 have very few degrees of freedom. However, if c is too small, the power will also be low, since any contrast between RSS_1 and RSS_2 is reduced. A rough guide is to set c at approximately $n/3$.

6.2.4 Extensions of the Goldfeld-Quandt Test

If the sample size is sufficiently large, it is possible to group the data into four, five, or more groups on the basis of an indicator variable such as X_i as before, and then

[6] S. M. Goldfeld and R. E. Quandt, "Some Tests for Homoscedasticity," *Journal of the American Statistical Association*, **60**, 1965, 539–547; or S. M. Goldfeld and R. E. Quandt, *Nonlinear Methods in Econometrics*, North-Holland, Amsterdam, 1972, Chapter 3, for a more general discussion.

to derive a likelihood ratio test of the constancy of the disturbance variance across groups. Suppose there are g groups, with n_i observations in the ith group, and $n = \sum_{i=1}^{g} n_i$ is the total sample size. The model is

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_g \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_g \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_g \end{bmatrix} \tag{6.12}$$

or, more compactly,

$$y = X\boldsymbol{\beta} + u \tag{6.13}$$

The assumption of a common β for all groups is maintained. The only question is the nature of the disturbance vector. The null hypothesis of homoscedasticity is

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_g^2 = \sigma^2$$

or,

$$E(uu') = V = \sigma^2 I_n \tag{6.14}$$

The alternative hypothesis of heteroscedasticity is

$$E(uu') = V = \begin{bmatrix} \sigma_1^2 I_{n_1} & 0 & \cdots & 0 \\ 0 & \sigma_2^2 I_{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_g^2 I_{n_g} \end{bmatrix} \tag{6.15}$$

The restricted likelihood is based on the maximum likelihood estimates of the parameters in Eqs. (6.13) and (6.14), namely, $\boldsymbol{\beta}$ and σ^2. The unrestricted likelihood is based on the maximum likelihood estimates of the parameters in Eqs. (6.13) and (6.15), namely, $\boldsymbol{\beta}$ and $\sigma_1^2, \sigma_2^2, \ldots, \sigma_g^2$. The detailed derivation is given in Appendix 6.2. Here we will give a brief outline of the procedure.

The log likelihood is

$$l = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|V| - \frac{1}{2}u'V^{-1}u \tag{6.16}$$

Maximizing Eq. (6.16), with V as specified in Eq. (6.14), is the standard linear model already analyzed in Chapter 5. The restricted MLEs are given in Eqs. (5.5) and (5.6), namely,

$$b = (X'X)^{-1}X'y \qquad \text{and} \qquad \hat{\sigma}^2 = (y - Xb)'(y - Xb)/n$$

Substitution of these MLEs in Eq. (6.16) would give the restricted log likelihood.

To obtain the unrestricted log likelihood we maximize Eq. (6.16) with V specified in Eq. (6.15). As shown in Eq. (5.29) the ML (also GLS) estimate of $\boldsymbol{\beta}$ under Eq. (6.15) is $(X'V^{-1}X)^{-1}X'V^{-1}y$. The problem now is that V contains the g unknown variances. Maximum likelihood estimates of these variances would yield a $\hat{V}$ matrix; and a maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ could then be obtained from

$$\hat{\boldsymbol{\beta}} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y \tag{6.17}$$

To obtain the maximum likelihood estimates of the disturbance variances we note that, under Eq. (6.15), the log likelihood is

$$l = -\frac{n}{2}\ln(2\pi) - \frac{n_1}{2}\ln\sigma_1^2 - \cdots - \frac{n_g}{2}\ln\sigma_g^2 - \frac{1}{2}\sum_{i=1}^{g}\frac{1}{\sigma_i^2}(y_i - X_i\beta)'(y_i - X_i\boldsymbol{\beta}) \tag{6.18}$$

The disturbance MLEs obtained from this are

$$\hat{\sigma}_i^2 = (y_i - X_i\hat{\boldsymbol{\beta}})'(y_i - X_i\hat{\boldsymbol{\beta}})/n_i \tag{6.19}$$

giving

$$\hat{V} = \begin{bmatrix} \hat{\sigma}_1^2 I_{n_1} & \cdots & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\sigma}_g^2 I_{n_g} \end{bmatrix} \tag{6.20}$$

We see that $\hat{\boldsymbol{\beta}}$ in Eq. (6.17) depends on $\hat{V}$, which in turn depends on $\hat{\boldsymbol{\beta}}$. The MLEs for the restricted likelihood may thus be obtained by iteration between Eqs. (6.17) and (6.20). The iteration could be started by estimating the $\boldsymbol{\beta}$ vector separately for each group, obtaining $\hat{\boldsymbol{\beta}}_i = (X_i'X_i)^{-1}X_i'y_i$ for $i = 1, 2, \ldots, g$. Substitution in Eq. (6.19) gives a $\hat{V}$ in Eq. (6.20) that on substitution in Eq. (6.17) gives a single $\hat{\boldsymbol{\beta}}$ vector, which may be substituted in Eqs. (6.19) and (6.20) to produce a new $\hat{V}$. The process is continued until a satisfactory degree of convergence is reached. If computational resources are limited, the process could be terminated at the first step outlined above, which is what is done in the simple application of the Goldfeld-Quandt test. Substituting the ML estimates in Eq. (6.18) gives the unrestricted log likelihood. When the restricted and unrestricted log likelihoods are substituted in the likelihood ratio, the relevant test statistic is

$$\mathrm{LR} = n\ln\hat{\sigma}^2 - \sum_{i=1}^{g} n_i \ln\hat{\sigma}_i^2 \overset{a}{\sim} \chi^2(g-1) \tag{6.21}$$

Large values of the test statistic lead to rejection of the hypothesis of homoscedasticity.

6.3 ESTIMATION UNDER HETEROSCEDASTICITY

If one or more of the tests in the previous section reject homoscedasticity, there are two possible ways to proceed in the estimation of the $\boldsymbol{\beta}$ vector. The first is to estimate $\boldsymbol{\beta}$ by OLS but compute the White covariance matrix in Eq. (6.6). This choice provides consistent estimates of the OLS standard errors and also permits Wald tests of linear restrictions as in Eq. (6.7). This procedure is attractive because of its simplicity; but the estimator, although unbiased and consistent, is inefficient. The second procedure is to compute some feasible GLS estimator in an attempt to capture the efficiency of GLS. However, this requires knowledge of the structural form of the heteroscedasticity, which may not always be available. Even when it is available, one cannot be sure how much of the potential gain in efficiency is captured because of the inherent inaccuracy of the estimation process.

6.3.1 Estimation with Grouped Data

The simplest case of feasible GLS estimation occurs with the kind of grouped data considered in the discussion of the likelihood ratio test. The test procedure has already yielded consistent estimates of the disturbances in Eq. (6.19) and of the $\boldsymbol{\beta}$ vector in Eq. (6.17). The relevant variance matrix for inference purposes is then

$$\operatorname{var}(\hat{\boldsymbol{\beta}}) = (X'\hat{V}^{-1}X)^{-1} \tag{6.22}$$

6.3.2 Estimation of the Heteroscedasticity Relation

Apart from the case of grouped data, the test procedures of the previous section shed no light on the functional form of the heteroscedasticity. Suppose, however, that we hypothesize

$$\sigma_t^2 = \alpha_0 + \alpha_1 z_t^{\alpha_2} \qquad t = 1, 2, \ldots, n \tag{6.23}$$

where z is a single variable, possibly one of the regressors, thought to determine the heteroscedasticity. If $\alpha_1 = 0$ this specification gives homoscedasticity, with $\sigma_t^2 = \alpha_0 (> 0)$. If $\alpha_0 = 0$ and $\alpha_2 = 1$, the disturbance variance is simply proportional to z, as in Eq. (6.2). If $\alpha_0 = 0$ and $\alpha_2 = 2$, the disturbance variance is proportional to the square of the determining variable. One or the other of these two special cases is often assumed in practice, and GLS then reduces to a simple application of **weighted least squares.** For example, suppose the assumption is made that $\sigma_t^2 = \alpha_1 z_t$. It then follows that

$$V = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} = \alpha_1 \begin{bmatrix} z_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & z_n \end{bmatrix} = \alpha_1 \boldsymbol{\Omega}$$

Looking at the constituents of the GLS estimator, we see

$$X'\boldsymbol{\Omega}^{-1}y = \begin{bmatrix} \vdots & & \vdots \\ x_1 & \cdots & x_n \\ \vdots & & \vdots \end{bmatrix} \begin{bmatrix} \frac{1}{z_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{z_n} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \sum_{t=1}^{n} \left(\frac{1}{z_t}\right) x_t y_t$$

In like fashion, it may be seen that

$$X'\boldsymbol{\Omega}^{-1}X = \sum_{t=1}^{n} \left(\frac{1}{z_t}\right) x_t x_t'$$

and so

$$b_{\text{GLS}} = \left[\sum_{t=1}^{n} x_t x_t'/z_t\right]^{-1} \left[\sum_{t=1}^{n} x_t y_t/z_t\right]$$

If y_t and each element in x_t are all multiplied by the square root of the reciprocal of z_t, the application of OLS to these transformed variables will give the b_{GLS} estimator.

If the assumption $\sigma_t^2 = \alpha_1 z_t^2$ were made, the appropriate weighting factor would be the reciprocal of z_t.

Rather than assume special cases of Eq. (6.23), one might estimate it directly. Because the OLS residuals $e_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}$ are consistent estimates of u_t, one might estimate Eq. (6.23) by the nonlinear regression

$$e_t^2 = \hat{\alpha}_0 + \hat{\alpha}_1 z_t^{\hat{\alpha}_2} + \hat{v}_t$$

Estimates of the disturbance variances are then

$$\hat{\sigma}_t^2 = \hat{\alpha}_0 + \hat{\alpha}_1 z_t^{\hat{\alpha}_2} \qquad t = 1, 2, \ldots, n \tag{6.24}$$

assuming all three parameters to be significant. These estimates give the $\hat{V}$ matrix and a feasible GLS procedure. Specifications with more variables than Eq. (6.23) might be treated in a similar fashion.

EXAMPLE 6.1. TESTS FOR HETEROSCEDASTICITY. The CPS88 data file on the diskette contains a random sample of 1000 observations from the *Current Population Survey*, 1988. The first 100 observations from the file were taken, and a conventional earnings equation was estimated. The results are shown in Table 6.2. The dependent variable is the log of wage (LNWAGE). Years of education are indicated by GRADE. Years of experience and its square are given by POTEXP and EXP2, and UNION is a zero/one dummy variable for membership in a union. The results conform with expectations. Education has a significant positive effect, experience has a quadratic effect, and the union dummy variable has a positive but not very significant coefficient.

To apply the White test for heteroscedasticity to this relation, we need first of all to square the regression residuals. The resultant series is denoted by RESSQ. Next we need to regress RESSQ on the original regressors and their squares and cross products. Taking account of the nature of the specific regressors, there are eight new regressors, noting that the square of the union dummy replicates the original dummy. The new variables are these:

TABLE 6.2
A conventional earnings equation

LS // Dependent Variable is LNWAGE
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	T-Statistic	Prob.
C	0.595106	0.283485	2.099248	0.0384
GRADE	0.083543	0.020093	4.157841	0.0001
POTEXP	0.050274	0.014137	3.556214	0.0006
EXP2	−0.000562	0.000288	−1.951412	0.0540
UNION	0.165929	0.124454	1.333248	0.1856

R-squared	0.371796	Mean dependent var	2.359213
Adjusted R-squared	0.34534	S.D. dependent var	0.580941
S.E. of regression	0.470043	Akaike info criterion	−1.461153
Sum squared resi	20.9893	Schwartz criterion	−1.330895
Log likelihood	−63.83618	F-statistic	14.05620
Durbin-Watson stat	2.161735	Prob(F-statistic)	0.000000

$$\begin{aligned} \text{GRADE2} &= \text{GRADE\^{}2} & \text{EXP4} &= \text{EXP2\^{}2} \\ \text{EXP3} &= \text{POTEXP} * \text{EXP2} & \text{GX} &= \text{GRADE} * \text{POTEXP} \\ \text{GX2} &= \text{GRADE} * \text{EXP2} & \text{GU} &= \text{GRADE} * \text{UNION} \\ \text{XU} &= \text{POTEXP} * \text{UNION} & \text{XU2} &= \text{EXP2} * \text{UNION} \end{aligned}$$

The White regression is shown in Table 6.3. The test statistic is $nR^2 = 10.79$ and $\chi^2_{.05}(12) = 21.03$. So the hypothesis of homoscedasticity is not rejected.

To apply the Breusch-Pagan/Godfrey test one must specify the variable or variables that one thinks influence the heteroscedasticity. Selecting GRADE, POTEXP, and UNION as possible candidates gives the regression shown in Table 6.4. From the table, and correcting for the scale factor, $\hat{\sigma}^2 = 0.2099$, from Table 6.2

$$\frac{1}{2}\text{ESS} = \frac{1}{2}\frac{R^2}{1 - R^2}\text{RSS} = \frac{(0.0428)(10.5480)}{2(0.9572)(0.2099)^2} = 5.35$$

The relevant critical value is $\chi^2_{.05}(3) = 7.815$, so homoscedasticity is not rejected. The alternative test statistic is $nR^2 = 4.28$, which is likewise insignificant.

Finally we illustrate the Goldfeld-Quandt test on these data. We sort the data by POTEXP and take the first and last 35 observations. The ratio of the second RSS to the first RSS is R = 7.5069/7.2517 = 1.06, which is insignificant, since $F_{.05}(30, 30) = 1.84$.

TABLE 6.3
White auxiliary regression

LS // Dependent Variable is RESSQ
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	T-Statistic	Prob.
C	−0.077672	0.985804	−0.078790	0.9374
GRADE	−0.012200	0.125021	−0.097586	0.9225
POTEXP	0.077838	0.071880	1.082882	0.2819
EXP2	−0.003990	0.004095	−0.974433	0.3325
UNION	0.648787	0.861596	0.753006	0.4535
GRADE2	0.002196	0.004247	0.516939	0.6065
EXP4	−3.34E-07	1.51E-06	−0.220995	0.8256
EXP3	6.17E-05	0.000142	0.434796	0.6648
GX	−0.003752	0.004942	−0.759234	0.4498
GX2	0.000117	0.000111	1.052392	0.2955
GU	−0.051374	0.044304	−1.159596	0.2494
XU	0.001933	0.060614	0.031885	0.9746
XU2	−0.000222	0.001259	−0.176223	0.8605

R-squared	0.107881	Mean dependent var	0.209894
Adjusted R-squared	−0.015170	S.D. dependent var	0.333630
S.E. of regression	0.336151	Akaike info criterion	−2.059652
Sum squared resid	9.830776	Schwartz criterion	−1.720980
Log likelihood	−25.91123	F-statistic	0.876722
Durbin-Watson stat	1.807900	Prob(F-statistic)	0.573082

TABLE 6.4
The Breusch-Pagan/Godfrey test

LS // Dependent Variable is RESSQ
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	T-Statistic	Prob.
C	−0.068446	0.199232	−0.343551	0.7319
GRADE	0.020768	0.013507	1.537566	0.1274
POTEXP	0.002089	0.002770	0.754211	0.4526
UNION	−0.122248	0.083188	−1.469547	0.1450

R-squared	0.042797	Mean dependent var	0.209894
Adjusted R-squared	0.012884	S.D. dependent var	0.333630
S.E. of regression	0.331474	Akaike info criterion	−2.169236
Sum squared resid	10.54798	Schwartz criterion	−2.065029
Log likelihood	−29.43206	F-statistic	1.430730
Durbin-Watson stat	1.791593	Prob(F-statistic)	0.238598

6.4 AUTOCORRELATED DISTURBANCES

Heteroscedasticity affects the elements on the principal diagonal of var($\boldsymbol{u}$), but the disturbances are still assumed to have zero pairwise covariances, that is, $E(u_t u_{t+s}) = 0$ for all t and $s \neq 0$. When the disturbances are autocorrelated (correlated with themselves), this assumption no longer holds. The pairwise **autocovariances** are defined by

$$\gamma_s = E(u_t u_{t+s}) \qquad s = 0, \pm 1, \pm 2, \ldots \tag{6.25}$$

When $s = 0$, Eq. (6.25) gives

$$\gamma_0 = E(u_t^2) = \sigma_u^2 \tag{6.26}$$

Thus the assumption is made that the disturbances are homoscedastic. For $s \neq 0$, Eq. (6.25) shows that the autocovariances are symmetric in the lag length s and independent of time t. It is a simple step from autocovariances to **autocorrelations.** The autocorrelation coefficient at lag s is

$$\rho_s = \frac{\text{cov}(u_t u_{t+s})}{\sqrt{\text{var}(u_t)\text{var}(u_{t+s})}}$$

Given homoscedasticity, this expression reduces to

$$\rho_s = \frac{\gamma_s}{\gamma_0} \qquad s = 0, \pm 1, \pm 2, \ldots \tag{6.27}$$

With n sample points there are $n - 1$ autocovariances and autocorrelations. Thus,

$$\operatorname{var}(\boldsymbol{u}) = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{bmatrix} = \sigma_u^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix} \tag{6.28}$$

Without further information the estimation problem is intractable because there are $n+k$ unknown parameters and only n observation points. As with heteroscedasticity, progress requires the assumption of some structure for the autocorrelation of the disturbance term.

6.4.1 Forms of Autocorrelation: Autoregressive and Moving Average Schemes

By far the most common specification in the literature is that of a first-order, autoregressive AR(1) process, or

$$u_t = \varphi u_{t-1} + \epsilon_t \tag{6.29}$$

where $\{\epsilon_t\}$ is a white noise process. The AR(1) process has already been studied in Section 2.5. The necessary and sufficient condition for a stationary disturbance process is

$$|\varphi| < 1 \tag{6.30}$$

The constant expectation for $\{u_t\}$ is

$$E(u_t) = 0 \qquad \text{for all } t \tag{6.31}$$

and, given Eq. (6.30), the constant variance is

$$\operatorname{var}(u_t) = \sigma_u^2 = \frac{\sigma_\epsilon^2}{1-\varphi^2} \tag{6.32}$$

and the autocorrelation coefficients are

$$\rho_s = \varphi^s \qquad s = 0, 1, 2, \ldots \tag{6.33}$$

The autocorrelation coefficients start at $\rho_0 = 1$ and then decline exponentially, but they never quite disappear. The current disturbance u_t is a weighted sum of the current shock and all previous shocks, or innovations, $\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, \ldots$, but the more distant shocks receive ever-declining weights, as is seen by rewriting Eq. (6.29) in the equivalent form of

$$u_t = \epsilon_t + \varphi\epsilon_{t-1} + \varphi^2\epsilon_{t-2} + \cdots \tag{6.34}$$

Given Eq. (6.29), the variance matrix of the disturbance vector is

$$\operatorname{var}(\boldsymbol{u}) = \sigma_u^2 \begin{bmatrix} 1 & \varphi & \cdots & \varphi^{n-1} \\ \varphi & 1 & \cdots & \varphi^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi^{n-1} & \varphi^{n-2} & \cdots & 1 \end{bmatrix} \tag{6.35}$$

Now there are only $k + 2$ parameters to be estimated, and feasible GLS procedures exist, as will be seen in Section 6.7.

A first-order, moving average MA(1) process is defined as

$$u_t = \epsilon_t + \theta\epsilon_{t-1} \qquad (6.36)$$

where $\{\epsilon_t\}$ is white noise. The crucial parameters of this process are

$$\begin{aligned} \sigma_u^2 &= \sigma_\epsilon^2(1 + \theta^2) \\ \rho_1 &= \theta/(1 + \theta^2) \\ \rho_i &= 0 \qquad i = 2, 3, \ldots \end{aligned} \qquad (6.37)$$

Unlike the autoregressive process, the MA process has a short, finite memory, being affected only by the current and prior values of ϵ.

Higher-order autoregressive and moving average processes are easily defined, and combinations of AR and MA processes of quite low orders can describe complicated time series behavior. These issues will be examined at length in Chapter 7. In the present context we are concerned with the behavior of the unknown disturbance term. Precise a priori knowledge of such behavior is not readily available, and the conventional practice is to specify simple forms of autocorrelated disturbances.

6.4.2 Reasons for Autocorrelated Disturbances

In the specification $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, one hopes to include all relevant variables in the $\boldsymbol{X}$ matrix. In such a case the disturbances would be expected to be serially uncorrelated. Significantly autocorrelated disturbances would thus be an indication of an inadequate specification. Suppose, for example, that the relationship really is

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + u_t$$

and u_t is white noise. The researcher specifies

$$y_t = \beta_1 + \beta_2 x_t + v_t$$

The pseudodisturbance is then $v_t = \beta_3 y_{t-1} + u_t$, which is autocorrelated because the "correct" specification makes y autocorrelated. To avoid such a situation it is better to err on the side of generosity rather than parsimony in the specification of the original relation. However, we can never know for sure all the variables that may play a role in determining y, so some variables are necessarily excluded. If these variables are autocorrelated, as most economic time series are, then the disturbance in the specified relation is likely to be autocorrelated. It is therefore important to test for autocorrelation and seek feasible estimation procedures, but first we will look at what is likely to be the result of using OLS.

6.5 OLS AND AUTOCORRELATED DISTURBANCES

The consequences of applying OLS to a relationship with nonstochastic X and autocorrelated disturbances are the same as those derived in Section 6.1 for the heteroscedasticity case, namely unbiased consistent, but inefficient estimation and invalid inference procedures. If, however, any lags of the dependent variable appear in X the results are radically different, as may be illustrated by a simple example. Suppose the postulated relationship is

$$\begin{aligned} y_t &= \beta y_{t-1} + u_t \qquad |\beta| < 1 \\ u_t &= \varphi u_{t-1} + \epsilon_t \qquad |\varphi| < 1 \end{aligned} \tag{6.38}$$

where $\qquad E(\boldsymbol{\epsilon}) = \mathbf{0} \quad$ and $\quad E(\boldsymbol{\epsilon\epsilon}') = \sigma_\epsilon^2 \boldsymbol{I}$

Estimating β by OLS gives

$$b = \frac{\sum y_t y_{t-1}}{\sum y_{t-1}^2} = \beta + \frac{\sum y_{t-1} u_t}{\sum y_{t-1}^2}$$

Thus,

$$\operatorname{plim} b = \beta + \frac{\operatorname{plim}\left(\frac{1}{n}\sum y_{t-1}u_t\right)}{\operatorname{plim}\left(\frac{1}{n}\sum y_{t-1}^2\right)}$$

The consistency of b then depends on $\operatorname{plim}(\sum y_{t-1}u_t/n)$. From Eq. (6.38)

$$y_{t-1} = u_{t-1} + \beta u_{t-2} + \beta^2 u_{t-3} + \cdots$$

The process of multiplying both sides by u_t, assuming that the autocovariances are consistently estimated by the sample moments, and using Eq. (6.33) gives

$$\operatorname{plim}\left(\frac{1}{n}\sum y_{t-1}u_t\right) = \varphi\sigma_u^2 + \beta\varphi^2\sigma_u^2 + \beta^2\varphi^3\sigma_u^2 + \cdots = \frac{\varphi\sigma_u^2}{1 - \beta\varphi}$$

Thus, the combination of a lagged dependent variable and an autocorrelated disturbance renders OLS inconsistent. OLS should not be used in such a case, and alternative estimators will be examined in Section 6.7.

If OLS is used with nonstochastic regressors, the same two questions arise as in the heteroscedasticity case. What is the bias in the conventional (incorrect) OLS standard errors, and how inefficient is OLS compared with feasible GLS estimators? These issues have been examined for a very simple model, namely,

$$\begin{aligned} y_t &= \beta x_t + u_t \\ u_t &= \varphi u_{t-1} + \epsilon_t \qquad |\varphi| < 1 \end{aligned} \tag{6.39}$$

with $\{\epsilon_t\}$ being a white noise series. The crucial difference between Eqs. (6.38) and (6.39) is that the stochastic y_{t-1} in the former is replaced by the nonstochastic x_t. The OLS estimate of β in Eq. (6.39) is $b = \sum_{t=1}^n y_t x_t / \sum_{t=1}^n x_t^2$. The correct sampling variance of this coefficient is obtained from

$$\operatorname{var}(b) = \sigma_u^2 (\boldsymbol{X'X})^{-1}\boldsymbol{X'\Omega X}(\boldsymbol{X'X})^{-1}$$

In this case $\boldsymbol{X}' = [x_1\ x_2\ \cdots\ x_n]$ and $\boldsymbol{\Omega}$, for an AR(1) process, is given in Eq. (6.35). Substitution gives

$$\operatorname{var}(b) = \frac{\sigma_u^2}{\sum_{t=1}^n x_t^2} \times \left(1 + 2\varphi\frac{\sum_{t=2}^n x_t x_{t-1}}{\sum_{t=1}^n x_t^2} + 2\varphi^2\frac{\sum_{t=3}^n x_t x_{t-2}}{\sum_{t=1}^n x_t^2} + \cdots + 2\varphi^{n-1}\frac{x_1 x_n}{\sum_{t=1}^n x_t^2}\right) \tag{6.40}$$

The first term on the right side of Eq. (6.40) is the conventional, but here incorrect, expression for var(b). The term in brackets involves powers of φ and sample autocorrelations of the regressor. If the regressor variable is not autocorrelated, the term in brackets will be negligible and the conventional variance not seriously biased. However, if the regressor and disturbance are both positively autocorrelated, the conventional standard error is likely to be a serious underestimate of the true standard error. Let r denote the sample, first-order autocorrelation coefficient of the regressor. The sum of just the first two terms inside the brackets is then $1 + 2\varphi r$. If $\varphi r = 0.5$, the conventional variance will be approximately one-half the correct value. In general, positively autocorrelated regressors combined with a positively autocorrelated disturbance term are likely to lead to serious biases in the standard errors calculated on the assumption of a white noise disturbance term.

OLS may also lead to substantial inefficiencies. Making the appropriate substitutions in

$$\text{var}(b_{\text{GLS}}) = \sigma_u^2(X'\Omega^{-1}X)^{-1}$$

gives

$$\text{var}(b_{\text{GLS}}) = \frac{\sigma_u^2}{\sum_{t=1}^n x_t^2} \times \left(\frac{1-\varphi^2}{1+\varphi^2 - 2\varphi \sum_{t=2}^n x_t x_{t-1} / \sum_{t=1}^n x_t^2 - \varphi^2(x_1^2 + x_n^2)/\sum_{t=1}^n x_t^2} \right) \quad (6.41)$$

Dropping negligible terms in Eqs. (6.40) and (6.41) gives an approximate expression for the efficiency of OLS as

$$\frac{\text{var}(b_{\text{GLS}})}{\text{var}(b)} \simeq \frac{1-\varphi^2}{(1+\varphi^2-2\varphi r)(1+2\varphi r)}$$

If, for example, $\varphi^2 = 0.5 = \varphi r$, this ratio is one-half. That is, the least-squares coefficient has a sampling variance twice that of GLS. However, not all of this efficiency may be captured in a feasible GLS estimation, since the "true" disturbance structure is unknown and has to be estimated. Although these specific results merely illustrate a very simple model, we suspect that autocorrelated disturbances pose a potentially serious problem, so testing for autocorrelation and devising feasible GLS procedures are important topics.

6.6 TESTING FOR AUTOCORRELATED DISTURBANCES

Suppose that in the model $y = X\beta + u$ one suspects that the disturbance follows an AR(1) scheme, namely,

$$u_t = \varphi u_{t-1} + \epsilon_t$$

The null hypothesis of zero autocorrelation is then

$$H_0\colon \varphi = 0$$

and the alternative hypothesis is

$$H_1: \varphi \neq 0$$

The hypothesis is about the u's, which are unobservable. One therefore looks for a test using the vector of OLS residuals, $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}$. This raises several diffiqulties. We saw in Chapter 3 that $\boldsymbol{e} = \boldsymbol{M}\boldsymbol{u}$, where $\boldsymbol{M} = \boldsymbol{I} - X(X'X)^{-1}\boldsymbol{X}'$ is symmetric, idempotent of rank $n - k$. Thus the variance-covariance matrix of the e's is

$$\text{var}(\boldsymbol{e}) = E(\boldsymbol{e}\boldsymbol{e}') = \sigma_u^2 \boldsymbol{M}$$

So even if the null hypothesis is true, in that $E(\boldsymbol{u}\boldsymbol{u}') = \sigma_u^2 \boldsymbol{I}$, the OLS residuals will display some autocorrelation, because the off-diagonal terms in $\boldsymbol{M}$ do not vanish. More importantly, $\boldsymbol{M}$ is a function of the sample values of the explanatory variables, which will vary unpredictably from one study to another. This variation makes it impossible to derive an exact finite-sample test on the e's that will be valid for any $\boldsymbol{X}$ matrix that might ever turn up.

6.6.1 Durbin-Watson Test

These problems were treated in a pair of classic and path-breaking articles.[7] The Durbin-Watson test statistic is computed from the vector of OLS residuals $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}$. It is denoted in the literature variously as d or DW and is defined as

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{6.42}$$

Figure 6.1 indicates why d might be expected to measure the extent of first-order autocorrelation. The mean residual is zero, so the residuals will be scattered around the horizontal axis. If the e's are positively autocorrelated, successive values will tend to be close to each other, runs above and below the horizontal axis will occur, and the first differences will tend to be numerically smaller than the residuals themselves. Alternatively, if the e's have a first-order negative autocorrelation, there is a tendency for successive observations to be on opposite sides of the horizontal axis so that first differences tend to be numerically larger than the residuals. Thus d will tend to be "small" for positively autocorrelated e's and "large" for negatively autocorrelated e's. If the e's are random, we have an in-between situation with no tendency for runs above and below the axis or for alternate swings across it, and d will take on an intermediate value.

The Durbin-Watson statistic is closely related to the sample first-order autocorrelation coefficient of the e's. Expanding Eq. (6.42), we have

$$d = \frac{\sum_{t=2}^{n} e_t^2 + \sum_{t=2}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$$

[7] J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least Squares Regression," *Biometrika*, **37**, 1950, 409–428; **38**, 1951, 159–178.

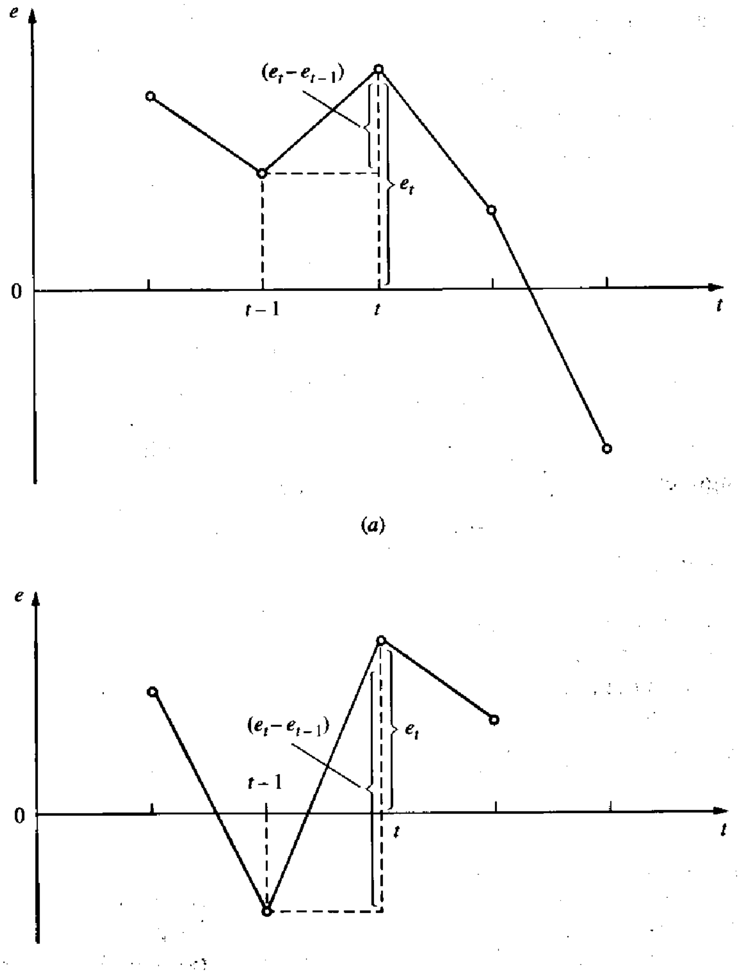


FIGURE 6.1
Autocorrelation patterns: (*a*) Positive autocorrelation; (*b*) negative autocorrelation.

For large *n* the different ranges of summation in numerator and denominator have a diminishing effect and

$$d \simeq 2(1 - \hat{\varphi}) \tag{6.43}$$

where $\hat{\varphi} = \sum e_t e_{t-1} / \sum e_{t-1}^2$ is the coefficient in the OLS regression of e_t on e_{t-1}. Ignoring end-point discrepancies, $\hat{\varphi}$ is seen to be the simple correlation coefficient between e_t and e_{t-1}. Thus, Eq. (6.43) shows heuristically that the range of *d* is from 0 to 4 as well as the following:

$d < 2$ for positive autocorrelation of the *e*'s

$d > 2$ for negative autocorrelation of the *e*'s

$d \simeq 2$ for zero autocorrelation of the *e*'s

The hypothesis under test is, of course, about the properties of the unobservable u's, which will not be reproduced exactly by the OLS residuals; but the foregoing indicators are nonetheless valid in that d will tend to be less (greater) than 2 for positive (negative) autocorrelation of the u's. For a random u series the expected value of d is

$$E(d) = 2 + \frac{2(k-1)}{n-k}$$

where k is the number of coefficients in the regression.

Because any computed d value depends on the associated X matrix, exact critical values of d that will cover all empirical applications cannot be tabulated. Durbin and Watson established upper (d_{U}) and lower (d_{L}) bounds for the critical values. These bounds depend only on the sample size and the number of regressors. They are used to test the hypothesis of zero autocorrelation against the alternative of *positive* first-order autocorrelation. The testing procedure is as follows:

1. If $d < d_{\mathrm{L}}$, reject the hypothesis of nonautocorrelated u in favor of the hypothesis of positive first-order autocorrelation.
2. If $d > d_{\mathrm{U}}$, do not reject the null hypothesis.
3. If $d_{\mathrm{L}} < d < d_{\mathrm{U}}$, the test is inconclusive.

If the value of d exceeds 2, one may wish to test the null hypothesis against the alternative of *negative* first-order autocorrelation. This test is done by calculating $4 - d$ and comparing this statistic with the tabulated critical values as if one were testing for positive autocorrelation. The original DW tables covered sample sizes from 15 to 100, with 5 as the maximum number of regressors. Savin and White have published extended tables for $6 \leq n \leq 200$ and up to 10 regressors.[8] The 5 percent and 1 percent Savin-White bounds are reproduced in Appendix D.

There are two important qualifications to the use of the Durbin-Watson test. First, it is necessary to include a constant term in the regression. Second, it is strictly valid only for a nonstochastic X matrix. Thus it is not applicable when lagged values of the dependent variable appear among the regressors. Indeed, it can be shown that the combination of a lagged Y variable and a positively autocorrelated disturbance term will bias the Durbin-Watson statistic upward and thus give misleading indications.[9] Even when the conditions for the validity of the Durbin-Watson test are satisfied, the inconclusive range is an awkward problem, especially as it becomes fairly large at low degrees of freedom. A conservative practical procedure is to use d_{U} as if it were a conventional critical value and simply reject the null hypothesis if $d < d_{\mathrm{U}}$. The consequences of accepting H_0 when autocorrelation is present are almost certainly more serious than the consequences of incorrectly presuming its presence. It has also been shown that when the regressors are slowly changing series, as many

[8] N. E. Savin and K. J. White, "The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors," *Econometrica*, **45**, 1977, 1989–1996.

[9] M. Nerlove and K. F. Wallis, "Use of the Durbin-Watson Statistic in Inappropriate Situations," *Econometrica*, **34**, 1966, 235–238.

economic series are, the true critical value will be close to the Durbin-Watson upper bound.[10]

When the regression does not contain an intercept term, the upper bound of the conventional Durbin-Watson tables is still valid. However, the lower bound needs to be replaced by d_M. An article by Farebrother provides extensive tables of the 5 percent and 1 percent values of d_M.[11]

6.6.2 The Wallis Test for Fourth-Order Autocorrelation

Wallis has pointed out that many applied studies employ quarterly data, and in such cases one might expect to find *fourth-order* autocorrelation in the disturbance term.[12] The appropriate specification is then

$$u_t = \varphi_4 u_{t-4} + \epsilon_t$$

To test the null hypothesis, H_0: $\varphi_4 = 0$, Wallis proposes a modified Durbin-Watson statistic,

$$d_4 = \frac{\sum_{t=5}^{n}(e_t - e_{t-4})^2}{\sum_{t=1}^{n} e_t^2} \tag{6.44}$$

where the e's are the usual OLS residuals. Wallis derives upper and lower bounds for d_4 under the assumption of a nonstochastic X matrix. The 5 percent points are tabulated in Appendix D. The first table is for use with regressions with an intercept, but without quarterly dummy variables. The second table is for use with regressions incorporating quarterly dummies. Giles and King provide further significance points at 2.5, 1.0, and 0.5 percent levels.[13]

6.6.3 Durbin Tests for a Regression Containing Lagged Values of the Dependent Variable

As has been pointed out, the Durbin-Watson test procedure was derived under the assumption of a nonstochastic X matrix, which is violated by the presence of lagged values of the dependent variable among the regressors. Durbin has derived a large-sample (asymptotic) test for the more general case.[14] It is still a test against

[10] H. Theil and A. L. Nagar, "Testing the Independence of Regression Disturbances," *Journal of the American Statistical Association*, **56**, 1961, 793-806; and E. J. Hannan and R. D. Terrell, "Testing for Serial Correlation after Least Squares Regression," *Econometrica*, **36**, 1968, 133–150.

[11] R. W. Farebrother, "The Durbin-Watson Test for Serial Correlation When There Is No Intercept in the Regression," *Econometrica*, **48**, 1980, 1553–1563.

[12] K. F. Wallis, "Testing for Fourth Order Autocorrelation in Quarterly Regression Equations," *Econometrica*, **40**, 1972, 617–636.

[13] D. E. A. Giles and M. L. King, "Fourth-Order Autocorrelation: Further Significance Points for the Wallis Test," *Journal of Econometrics*, **8**, 1978, 255–259.

[14] J. Durbin, "Testing for Serial Correlation in Least Squares Regression When Some of the Regressors Are Lagged Dependent Variables," *Econometrica*, **38**, 1970, 410–421.

first-order autocorrelation, and one must specify the complete set of regressors. Consider the relation

$$y_t = \beta_1 y_{t-1} + \cdots + \beta_r y_{t-r} + \beta_{r+1} x_{1t} + \cdots + \beta_{r+s} x_{st} + u_t \tag{6.45}$$

with $\qquad u_t = \varphi u_{t-1} + \epsilon_t \qquad |\varphi| < 1 \qquad \text{and} \qquad \epsilon \sim N(0, \sigma_\epsilon^2 I)$

Durbin's basic result is that under the null hypothesis, $H_0\colon \varphi = 0$, the statistic

$$h = \hat{\varphi}\sqrt{\frac{n}{1 - n \cdot \text{var}(b_1)}} \overset{a}{\sim} N(0, 1) \tag{6.46}$$

where n = sample size

$\text{var}(b_1)$ = estimated sampling variance of the coefficient of y_{t-1} in the OLS fit of Eq. (6.45)

$\hat{\varphi} = \sum_{t=2}^{n} e_t e_{t-1} / \sum_{t=2}^{n} e_{t-1}^2$, the estimate of φ from the regression of e_t on e_{t-1}, the e's in turn being the residuals from the OLS regression of Eq. (6.45)

The test procedure is as follows:

1. Fit the OLS regression of Eq. (6.45) and note $\text{var}(b_1)$.
2. From the residuals compute $\hat{\varphi}$ or, if the Durbin-Watson statistic has been computed, use the approximation $\hat{\varphi} = 1 - d/2$.
3. Compute h, and if $h > 1.645$, reject the null hypothesis at the 5 percent level in favor of the hypothesis of positive, first-order autocorrelation.
4. For negative h a similar one-sided test for negative autocorrelation can be carried out.

The test breaks down if $n \cdot \text{var}(b_1) \geq 1$. Durbin showed that an asymptotically equivalent procedure is the following:

1. Estimate the OLS regression [Eq. (6.45)] and obtain the residual e's.
2. Estimate the OLS regression of

$$e_t \text{ on } e_{t-1}, y_{t-1}, \ldots, y_{t-r}, x_{1t}, \ldots, x_{st}$$

3. If the coefficient of e_{t-1} in this regression is significantly different from zero by the usual t test, reject the null hypothesis $H_0\colon \varphi = 0$.

Durbin indicates that this last procedure can be extended to test for an AR(p) disturbance rather than an AR(1) process by simply adding additional lagged e's to the second regression and testing the *joint significance* of the coefficients of the lagged residuals. The AR(p) scheme is

$$u_t = \varphi_1 u_{t-1} + \varphi_2 u_{t-2} + \cdots + \varphi_p u_{t-p} + \epsilon_t \tag{6.47}$$

The null hypothesis would now be

$$H_0\colon \varphi_1 = \varphi_2 = \cdots = \varphi_p = 0$$

The resultant test statistic may be expressed more simply in matrix notation, and this will have the additional advantage of making clear its relation to the Breusch-Godfrey test, which will be explained next.

Let Z denote the $n \times (r+s)$ matrix of the sample data on all the regressors in Eq. (6.45), and $\boldsymbol{e} = \boldsymbol{y} - Z(Z'Z)^{-1}Z'\boldsymbol{y}$ the $n \times 1$ vector of residuals from fitting Eq. (6.45) by OLS. Define

$$E = [\boldsymbol{e}_1 \quad \boldsymbol{e}_2 \quad \cdots \quad \boldsymbol{e}_p] = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ e_1 & 0 & 0 & \cdots & 0 \\ e_2 & e_1 & 0 & \cdots & 0 \\ e_3 & e_2 & e_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_p & e_{p-1} & e_{p-2} & \cdots & e_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_{n-1} & e_{n-2} & e_{n-3} & \cdots & e_{n-p} \end{bmatrix} \tag{6.48}$$

The second-stage regression is then

$$\boldsymbol{e} = [E \quad Z]\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{c} \end{bmatrix} + \boldsymbol{v} \tag{6.49}$$

where $\boldsymbol{a}$ is the vector of estimated coefficients on the lagged residuals and $\boldsymbol{c}$ is the vector of estimated coefficients on the lagged y's and x's. Durbin's suggestion is to test the joint significance of the variables in E. As seen in Chapter 3, this test may easily be done by computing an F statistic based on the difference in the residual sum of squares from a restricted and an unrestricted regression. The relevant restricted regression is $\boldsymbol{e}$ on Z, with ESS $= 0$ and RSS $= \boldsymbol{e}'\boldsymbol{e}$, because $Z'\boldsymbol{e} = \boldsymbol{0}$ from Eq. (6.45). The unrestricted regression is $\boldsymbol{e}$ on $[E\ Z]$. The ESS from this regression is

$$\begin{aligned} \text{ESS} &= \boldsymbol{e}'[E \quad Z]\begin{bmatrix} E'E & E'Z \\ Z'E & Z'Z \end{bmatrix}^{-1}\begin{bmatrix} E' \\ Z' \end{bmatrix}\boldsymbol{e} \\ &= \begin{bmatrix} \boldsymbol{e}'E & \boldsymbol{0} \end{bmatrix}\begin{bmatrix} E'E & E'Z \\ Z'E & Z'Z \end{bmatrix}^{-1}\begin{bmatrix} E'\boldsymbol{e} \\ \boldsymbol{0} \end{bmatrix} \\ &= \boldsymbol{e}'E[E'E - E'Z(Z'Z)^{-1}Z'E]^{-1}E'\boldsymbol{e} \end{aligned} \tag{6.50}$$

Notice that no correction for the mean is required in this expression because $\boldsymbol{e}$ has mean zero.

From Eq. (3.42), the F statistic to test the joint significance of the coefficients on the lagged residuals is then

$$F = \frac{\boldsymbol{e}'E[E'E - E'Z(Z'Z)^{-1}Z'E]^{-1}E'\boldsymbol{e}/p}{\boldsymbol{v}'\boldsymbol{v}/[n-(p+r+s)]} \tag{6.51}$$

However, this test statistic does not have exact, finite sample validity since the regressor matrix in Eq. (6.49) is stochastic. As $n \rightarrow \infty$, $p \cdot F$ tends in distribution to $\chi^2(p)$.[15] Thus an asymptotic test of zero autocorrelation in the disturbances against the alternative of a pth-order autoregression is obtained by computing F as in Eq. (6.51) and referring $p \cdot F$ to $\chi^2(p)$.

[15] See Appendix B.

6.6.4 Breusch-Godfrey Test[16]

These two authors independently built on the work of Durbin to develop LM tests against general autoregressive or moving average disturbance processes. We illustrate the development with a very simple example. Suppose the specified equation is

$$y_t = \beta_1 + \beta_2 x_t + u_t \tag{6.52}$$

with

$$u_t = \beta_3 u_{t-1} + \epsilon_t \tag{6.53}$$

where it is assumed that $|\beta_3| < 1$ and that the ϵ's are independently and identically distributed normal variables with zero mean and variance σ_ϵ^2. Substituting Eq. (6.53) in Eq. (6.52) gives

$$y_t = \beta_1(1-\beta_3) + \beta_2 x_t + \beta_3 y_{t-1} - \beta_2\beta_3 x_{t-1} + \epsilon_t \tag{6.54}$$

We wish to test the hypothesis that $\beta_3 = 0$. Equation (6.54) is nonlinear in the β's. However, if the restriction on β_3 is imposed, it reduces to Eq. (6.52), which is linear in the $\boldsymbol{\beta}$'s, making the LM test attractive.

The sum of squares term in the log-likelihood function for Eq. (6.54) is

$$-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{n}\epsilon_t^2$$

As seen in Chapter 5, the information matrix for this type of regression model is block diagonal, so the $\beta = [\beta_1\ \beta_2\ \beta_3]'$ parameters can be treated separately from the σ_ϵ^2 parameter. The score vector is then

$$s(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} = -\frac{1}{\sigma_\epsilon^2}\sum_{t=1}^{n}\epsilon_t\frac{\partial \epsilon_t}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma_\epsilon^2}\sum_{t=1}^{n}\epsilon_t \mathbf{w}_t$$

where $\mathbf{w}_t = -\partial\epsilon_t/\partial\boldsymbol{\beta}$. The information matrix is

$$\begin{aligned} I(\boldsymbol{\beta}) &= E[s(\boldsymbol{\beta})s'(\boldsymbol{\beta})] \\ &= E\left[\frac{1}{\sigma_\epsilon^4}\left(\sum \epsilon_t \mathbf{w}_t\right)\left(\sum \epsilon_t \mathbf{w}_t\right)'\right] \\ &= E\left[\frac{1}{\sigma_\epsilon^4}\left(\sum \epsilon_t^2 \mathbf{w}_t \mathbf{w}_t' + \sum_{t \neq s} \epsilon_t \epsilon_s \mathbf{w}_t \mathbf{w}_s'\right)\right] \\ &= \frac{1}{\sigma_\epsilon^2}E\left(\sum \mathbf{w}_t \mathbf{w}_t'\right) \end{aligned}$$

where the last line follows from the assumptions about the ϵ's. Asymptotically it makes no difference if $E(\sum \mathbf{w}_t\mathbf{w}_t')$ is replaced by $\sum \mathbf{w}_t\mathbf{w}_t'$. Thus the LM statistic in

[16] T. S. Breusch, "Testing for Autocorrelation in Dynamic Linear Models," *Australian Economic Papers*, **17**, 1978, 334–355; and L. G. Godfrey, "Testing against General Autoregressive and Moving Average Error Models When the Regressors Include Lagged Dependent Variables," *Econometrica*, **46**, 1978, 1293–1302.

Eq. (5.16) becomes

$$\begin{aligned}\text{LM} &= \frac{1}{\tilde{\sigma}_\epsilon^2}\left(\sum \tilde{\epsilon}_t \tilde{\boldsymbol{w}}_t\right)'\left(\sum \tilde{\boldsymbol{w}}_t \tilde{\boldsymbol{w}}_t'\right)^{-1}\left(\sum \tilde{\epsilon}_t \tilde{\boldsymbol{w}}_t\right) \\ &= \frac{1}{\tilde{\sigma}_\epsilon^2}\tilde{\boldsymbol{\epsilon}}'\tilde{W}(\tilde{W}'\tilde{W})^{-1}\tilde{W}'\tilde{\boldsymbol{\epsilon}}\end{aligned} \tag{6.55}$$

where $\tilde{W} = \begin{bmatrix} \cdots & \tilde{\boldsymbol{w}}_1' & \cdots \\ & \vdots & \\ \cdots & \tilde{\boldsymbol{w}}_n' & \cdots \end{bmatrix}$ and $\tilde{\boldsymbol{\epsilon}} = \begin{bmatrix} \tilde{\epsilon}_1 \\ \tilde{\epsilon}_2 \\ \vdots \\ \tilde{\epsilon}_n \end{bmatrix}$

and the tildes indicate that all elements in Eq. (6.55) are evaluated at the restricted estimates, $\tilde{\boldsymbol{\beta}}, \tilde{\sigma}_\epsilon^2 (= \tilde{\boldsymbol{\epsilon}}'\tilde{\boldsymbol{\epsilon}}/n)$. Equation (6.55) shows that LM $= nR^2$, where R^2 is the squared multiple correlation coefficient from the regression of $\tilde{\epsilon}_t$ on $\tilde{\boldsymbol{w}}_t$. It is clear from Eq. (6.54) that imposing the restriction $\beta_3 = 0$ gives $\tilde{\epsilon}_t = y_t - \tilde{\beta}_1 - \tilde{\beta}_2 x_t$, which is the residual from the application of OLS to Eq. (6.52). Further,

$$\boldsymbol{w}_t = \begin{bmatrix} -\dfrac{\partial \epsilon_t}{\partial \beta_1} \\ -\dfrac{\partial \epsilon_t}{\partial \beta_2} \\ -\dfrac{\partial \epsilon_t}{\partial \beta_3} \end{bmatrix} = \begin{bmatrix} 1 - \beta_3 \\ x_t - \beta_3 x_{t-1} \\ y_{t-1} - \beta_1 - \beta_2 x_{t-1} \end{bmatrix}$$

Setting β_3 to zero and replacing β_1 and β_2 by their estimates under the null gives

$$\tilde{\boldsymbol{w}}_t = \begin{bmatrix} 1 \\ x_t \\ \tilde{\epsilon}_{t-1} \end{bmatrix}$$

The test of $\beta_3 = 0$ is therefore obtained in two steps. First apply OLS to Eq. (6.52) to obtain the residuals $\tilde{u}_t$, which we have been accustomed to label e_t. Then regress e_t on $[1 \;\; x_t \;\; e_{t-1}]$ to find R^2. Under H_0, nR^2 is asymptotically $\chi^2(1)$. The second, or auxiliary, regression is exactly the regression of the Durbin procedure. The only difference is in the test procedure. Durbin suggests looking at the significance of the coefficient on e_{t-1}. The Breusch-Godfrey derivation of the LM test gives nR^2 as a test statistic with an asymptotic χ^2 distribution.

This procedure clearly extends to testing for higher orders of autocorrelation. One simply adds further-lagged OLS residuals to the second regression, exactly as shown in the Durbin regression [Eq. (6.49)]. A remarkable feature of the Breusch-Godfrey test is that it also tests against the alternative hypothesis of an MA(p) process for the disturbance.

Finally, note that the Durbin and Breusch-Godfrey procedures are asymptotically equivalent. In general it may be seen that the $\tilde{W}$ matrix in Eq. (6.55) is the $[E \; Z]$ matrix in Eq. (6.49), and $\tilde{\boldsymbol{\epsilon}}$ is $\boldsymbol{e}$. Thus the LM statistic in Eq. (6.55) is, using Eq. (6.50),

$$\mathrm{LM} = \frac{e'E\left[E'E - E'Z(Z'Z)^{-1}Z'E\right]^{-1}E'e}{e'e/n} \tag{6.56}$$

The only difference between the Durbin statistic in Eq. (6.51) and the Breusch-Godfrey statistic in Eq. (6.56) is in the variance terms in the denominator. Breusch shows that these terms have the same probability limit, and so the procedures are asymptotically equivalent.[17]

6.6.5 Box-Pierce-Ljung Statistic

The Box-Pierce Q statistic is based on the squares of the first p autocorrelation coefficients of the OLS residuals.[18] The statistic is defined as

$$Q = n\sum_{j=1}^{p} r_j^2 \tag{6.57}$$

where

$$r_j = \frac{\sum_{t=j+1}^{n} e_t e_{t-j}}{\sum_{t=1}^{n} e_t^2}$$

The limiting distribution of Q was derived under the assumption that the residuals come from an autoregressive AR scheme, or, more generally, from an autoregressive, moving average ARMA scheme fitted to some variable y. Under the hypothesis of zero autocorrelations for the residuals, Q will have an asymptotic χ^2 distribution, with degrees of freedom equal to p minus the number of parameters estimated in fitting the ARMA model. An improved small-sample performance is expected from the revised Ljung-Box statistic,[19]

$$Q' = n(n+2)\sum_{j=1}^{p} r_j^2/(n-j) \tag{6.58}$$

These statistics are sometimes used to test for autocorrelated disturbances in the type of regression equation that we have been considering, but this application is inappropriate because equations such as Eq. (6.54) are not pure AR schemes but have exogenous x variables as well. The effect on the distribution of Q or Q' is unknown.[20]

[17] T. Breusch, ibid., 354.

[18] G. E. P. Box and David A. Pierce, "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models," *Journal of the American Statistical Association*, **65**, 1970, 1509–1526.

[19] G. M. Ljung and G. E. P. Box, "On a Measure of Lack of Fit in Time Series Models," *Biometrika*, **65**, 1978, 297–303.

[20] Hashem Dezhbaksh, "The Inappropriate Use of Serial Correlation Tests in Dynamic Linear Models," *The Review of Economics and Statistics*, **LXXII**, 1990, 126–132.

6.7 ESTIMATION OF RELATIONSHIPS WITH AUTOCORRELATED DISTURBANCES

If one or more of the tests described in the previous section suggest autocorrelated disturbances, what should be done? One possibility is to proceed to a *joint specification* of the relationship, $y = X\beta + u$, and an associated autocorrelation structure,

$$Pu = \epsilon \qquad \text{with} \qquad E(\epsilon\epsilon') = \sigma_\epsilon^2 I \tag{6.59}$$

where P is some nonsingular $n \times n$ matrix that depends, one hopes, on a small number p of unknown parameters. The next step is the joint estimation of all $(k + p + 1)$ parameters. A second, and better, procedure is to start by checking whether the autocorrelation may not be a sign of misspecification in the original relationship. We can, of course, never know the "true" or "correct" relationship. Presumably, most data are generated by extremely complicated processes that defy precise specification and accurate estimation. The target is to obtain as good an approximation as possible to the unknowable complexity. Words like "true" or "correct" may then be used loosely to refer to such an approximation. If a relationship has autocorrelated errors, there is some systematic behavior that is not being modeled but instead is being consigned to the disturbance term. It is desirable to get a comprehensive model of systematic effects and to reduce the errors to white noise. Suppose, for illustration, that the "correct" relationship is

$$y_t = \gamma_1 + \gamma_2 x_t + \gamma_3 x_{t-1} + \gamma_4 y_{t-1} + \epsilon_t \tag{6.60}$$

where the $\{\epsilon_t\}$ are white noise. A researcher's economic theory, however, delivers the proposition that y_t is influenced only by x_t. When this model is fitted to the data, it is not surprising that significant autocorrelation is found in the errors. To take this autocorrelation into account in the estimation of the relationship, the researcher now specifies

$$y_t = \beta_1 + \beta_2 x_t + u_t \qquad \text{and} \qquad u_t = \varphi u_{t-1} + \epsilon_t \tag{6.61}$$

and proceeds to estimate it by, say, GLS. The correct model in Eq. (6.60) involves five parameters, namely, four coefficients and a variance, whereas our researcher's specification has just four parameters. The researcher is thus imposing a possibly invalid restriction on the parameters of the true model. The nature of this restriction may be seen by rewriting Eq. (6.61) in the form

$$y_t = \beta_1(1 - \varphi) + \beta_2 x_t - \varphi\beta_2 x_{t-1} + \varphi y_{t-1} + \epsilon_t \tag{6.62}$$

Comparison of the parameters in Eqs. (6.60) and (6.62) shows that the restriction involved in moving from the former to the latter is

$$\gamma_3 + \gamma_2\gamma_4 = 0 \tag{6.63}$$

This is known as a **common factor** restriction.[21] The origin of the term may be seen by rewriting Eqs. (6.60) and (6.62) using the lag operator.[22] Equation (6.62) becomes

[21] For an extensive discussion, see the Manual for *PCGive*, Version 7, by Jurgen A. Doornik and David F. Hendry, Institute of Economics and Statistics, University of Oxford, UK, 1992.

[22] The lag operator is described in Chapter 7.

$$(1 - \varphi L)y_t = \beta_1(1 - \varphi) + \beta_2(1 - \varphi L)x_t + \epsilon_t$$

showing y_t and x_t to have a common factor in the lag operator. Equation (6.60) gives

$$\begin{aligned}(1 - \gamma_4 L)y_t &= \gamma_1 + (\gamma_2 + \gamma_3 L)x_t + \epsilon_t \\ (1 - \gamma_4 L)y_t &= \gamma_1 + \gamma_2\left(1 + \frac{\gamma_3}{\gamma_2}L\right)x_t + \epsilon_t\end{aligned}$$

If this equation is to have a common factor, the γ's must obey the restriction stated in Eq. (6.63). Thus, one should search for specifications such as Eq. (6.60) with white noise residuals. Tests for common factors can then show whether to reduce such specifications to the more parsimonious specifications like Eq. (6.61).

GLS Estimation. We now assume that the specification $y = X\boldsymbol{\beta} + u$ is as good as we can make it but that, nonetheless, we must allow for an autocorrelation structure, as in Eq. (6.59). Some specific form of autocorrelation must be assumed. By far the most common assumption is an AR(1) process. In that case, as we saw in Eq. (6.35), the variance-covariance matrix for $\boldsymbol{u}$ is

$$\begin{aligned}\operatorname{var}(\boldsymbol{u}) &= \sigma_u^2 \begin{bmatrix} 1 & \varphi & \cdots & \varphi^{n-1} \\ \varphi & 1 & \cdots & \varphi^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi^{n-1} & \varphi^{n-2} & \cdots & 1 \end{bmatrix} \\ &= \frac{\sigma_\epsilon^2}{1-\varphi^2} \begin{bmatrix} 1 & \varphi & \cdots & \varphi^{n-1} \\ \varphi & 1 & \cdots & \varphi^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi^{n-1} & \varphi^{n-2} & \cdots & 1 \end{bmatrix} \qquad (6.64) \\ &= \sigma_\epsilon^2 \boldsymbol{\Omega}\end{aligned}$$

where

$$\boldsymbol{\Omega} = \frac{1}{1-\varphi^2} \begin{bmatrix} 1 & \varphi & \cdots & \varphi^{n-1} \\ \varphi & 1 & \cdots & \varphi^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi^{n-1} & \varphi^{n-2} & \cdots & 1 \end{bmatrix} \qquad (6.65)$$

The inverse matrix, as may readily be verified by multiplying out, is

$$\boldsymbol{\Omega}^{-1} = \begin{bmatrix} 1 & -\varphi & 0 & \cdots & 0 & 0 & 0 \\ -\varphi & 1+\varphi^2 & -\varphi & \cdots & 0 & 0 & 0 \\ 0 & -\varphi & 1+\varphi^2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\varphi & 1+\varphi^2 & -\varphi \\ 0 & 0 & 0 & \cdots & 0 & -\varphi & 1 \end{bmatrix} \qquad (6.66)$$

It can then be seen that the matrix

$$P = \begin{bmatrix} \sqrt{1-\varphi^2} & 0 & 0 & \cdots & 0 & 0 \\ -\varphi & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\varphi & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\varphi & 1 \end{bmatrix} \tag{6.67}$$

satisfies the condition in Eq. (5.23), namely, $\boldsymbol{\Omega}^{-1} = \boldsymbol{P}'\boldsymbol{P}$. If φ be known, there are two equivalent ways of deriving GLS estimates of the $\boldsymbol{\beta}$ vector.[23] One is to substitute φ in Eq. (6.66) and compute $\boldsymbol{b}_{\text{GLS}} = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}$ directly. The alternative is to transform the data by premultiplication by the $\boldsymbol{P}$ matrix and then estimate the OLS regression of $\boldsymbol{y}_*(= \boldsymbol{Py})$ on $\boldsymbol{X}_*(= \boldsymbol{PX})$. This transformation treats the first data point differently from the rest. If, for simplicity, the $\boldsymbol{X}$ matrix contained just a unit vector and a single explanatory variable, the transformed data would be

$$\boldsymbol{y}_* = \begin{bmatrix} (\sqrt{1-\varphi^2})\cdot y_1 \\ y_2 - \varphi y_1 \\ \vdots \\ y_n - \varphi y_{n-1} \end{bmatrix} \qquad \boldsymbol{X}_* = \begin{bmatrix} \sqrt{1-\varphi^2} & (\sqrt{1-\varphi^2})\cdot x_1 \\ 1-\varphi & x_2 - \varphi x_1 \\ \vdots & \vdots \\ 1-\varphi & x_n - \varphi x_{n-1} \end{bmatrix} \tag{6.68}$$

and the GLS estimates would be obtained by regressing $\boldsymbol{y}_*$ on $\boldsymbol{X}_*$, taking care to suppress the constant in the regression package. If, however, the first row in $\boldsymbol{P}$ is dropped, the regression between transformed variables would simply be that of $(y_t - \varphi y_{t-1})$ on a constant and $(x_t - \varphi x_{t-1})$ for the sample points $t = 2, 3, \ldots, n$. The latter regression is obviously not full GLS, and in small samples dropping the first observation can have a marked effect on the coefficient estimate, although asymptotically it is of little importance.

In practice φ is an unknown parameter, which must be estimated along with the other parameters of the model. To give as simple an explanation as possible of the various estimation procedures, we will use the model in Eq. (6.61). As just shown, this model can be rewritten in Eq. (6.62) as

$$y_t = \beta_1(1-\varphi) + \beta_2 x_t - \varphi\beta_2 x_{t-1} + \varphi y_{t-1} + \epsilon_t$$

By the assumption expressed in Eq. (6.59) the disturbance vector $\boldsymbol{\epsilon}$ in this relation is "well behaved," and minimization of $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$ will deliver GLS estimates, subject to the same caveat as before about the treatment of the first observation. However, Eq. (6.62) is *nonlinear* in the three parameters. Thus, **nonlinear least squares** (NLS) is required. With the dramatic advances in personal computing, NLS is readily available and should be used.

In former days, when issues of computability loomed large, much attention was devoted to simple ways of estimating a relation such as Eq. (6.62). The crucial step was to notice that Eq. (6.62) can be rearranged in two equivalent forms as

[23] If necessary, revisit the discussion of GLS in Section 5.4.

$$(y_t - \varphi y_{t-1}) = \beta_1(1-\varphi) + \beta_2(x_t - \varphi x_{t-1}) + \epsilon_t \qquad \textbf{(6.69a)}$$

or

$$(y_t - \beta_1 - \beta_2 x_t) = \varphi(y_{t-1} - \beta_1 - \beta_2 x_{t-1}) + \epsilon_t \qquad \textbf{(6.69b)}$$

If φ were known in Eq. (6.69*a*), the β's could be estimated by straightforward OLS. Similarly, if the β's were known in Eq. (6.69*b*), φ could be estimated by an OLS regression with the intercept suppressed. The seminal paper by Cochrane and Orcutt suggested an **iterative estimation procedure** based on this pair of relations.[24] Start, say, with an estimate or guess $\hat{\varphi}^{(1)}$ of the autocorrelation parameter, and use it to compute the quasi differences $(y_t - \hat{\varphi}^{(1)}y_{t-1})$ and $(x_t - \hat{\varphi}^{(1)}x_{t-1})$. These transformed variables are then used in the OLS regression [Eq. (6.69*a*)], yielding estimated coefficients $b_1^{(1)}$ and $b_2^{(1)}$. These in turn are used to compute the variables in Eq. (6.69*b*); and an OLS regression yields a new estimate $\hat{\varphi}^{(2)}$ of the autocorrelation parameter. The iterations continue until a satisfactory degree of convergence is reached. The Cochrane-Orcutt (C-O) procedure is applied to $t = 2, 3, \ldots, n$, which is equivalent to dropping the first row of the $\boldsymbol{P}$ matrix in Eq. (6.67). Prais and Winsten pointed out that the full $\boldsymbol{P}$ matrix should be used, so that the first observation receives explicit treatment.[25] The concern with iterative procedures is that they may converge to a local minimum and not necessarily to the global minimum. A precaution is to fit equations like Eq. (6.69*a*) for a grid of φ values in steps of 0.1 from, say, -0.9 to 0.9 and then iterate from the regression with the smallest RSS. The same problem exists with NLS, which also uses iteration. It is advisable to start the NLS process with several different coefficient vectors to see if convergence takes place at the same vector.

GLS procedures minimize $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$. However, they do not yield ML estimates, even with special treatment of the first observation. The reason may be seen by referring to the log-likelihood in Eq. (6.16), namely,

$$l = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{V}| - \frac{1}{2}\boldsymbol{u}'\boldsymbol{V}^{-1}\boldsymbol{u}$$

From the relations already defined in Eqs. (6.59), (6.64), and (6.67) it follows that[26]

$$|\boldsymbol{V}| = \sigma_\epsilon^{2n}|\boldsymbol{\Omega}| = \sigma_\epsilon^{2n}(1-\varphi^2)^{-1} \qquad (6.70)$$

and

$$\boldsymbol{u}'\boldsymbol{V}^{-1}\boldsymbol{u} = \frac{1}{\sigma_\epsilon^2}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} \qquad (6.71)$$

Thus,

$$l = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma_\epsilon^2) + \frac{1}{2}\ln(1-\varphi^2) - \frac{1}{2\sigma_\epsilon^2}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} \qquad (6.72)$$

Maximizing the log-likelihood takes account of the term in $\ln(1-\varphi^2)$, which is ignored in the GLS procedure. Beach and MacKinnon drew attention to this point

[24] D. Cochrane and G. H. Orcutt, "Application of Least Squares Regressions to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Association*, **44**, 1949, 32–61.

[25] S. J. Prais and C. B. Winsten, "Trend Estimators and Serial Correlation," *Cowles Commission Discussion Paper*, No. 383, Chicago, 1954.

[26] See Problem 6.6.

and devised an iterative procedure for maximizing Eq. (6.72).[27] Their procedure is available in some software packages. Its advantage is that the estimate of φ is necessarily confined to the unit interval, since $\ln(1 - \varphi^2)$ approaches minus infinity as φ approaches ± 1. If GLS gives an autocorrelation parameter outside the unit interval, the particular specification should be discarded and the model respecified.

More complicated structures may be specified for $\boldsymbol{u}$, such as higher-order AR processes, MA processes, or a combination of both in ARMA schemes. In practice, however, there is normally little sound information on the relevant specification, and the best advice is to develop a rich specification of the original relation so that there is little need for complicated specifications of the disturbance term.

6.8 FORECASTING WITH AUTOCORRELATED DISTURBANCES

Section 3.5 showed how to make point and interval forecasts for the basic linear model with well-behaved disturbances. We now need to see how to update those methods for a linear model with autocorrelated disturbances. As an illustration we will consider the AR(1) case. The specification is

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + u_t \qquad t = 1, 2, \ldots, n \tag{6.73}$$

where $u_t = \varphi u_{t-1} + \epsilon_t$ with $|\varphi| < 1$ and $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma_\epsilon^2 \boldsymbol{I}$ (6.74)

Combining the two relations gives

$$y_{*,t} = \boldsymbol{x}_{*,t}'\boldsymbol{\beta} + \epsilon_t \tag{6.75}$$

where $y_{*,t} = y_t - \varphi y_{t-1}$ and $\boldsymbol{x}_{*,t} = \boldsymbol{x}_t - \varphi \boldsymbol{x}_{t-1}$. This relation satisfies the standard conditions for OLS. Applying OLS to the transformed data, with appropriate treatment of the first observation, then delivers $\boldsymbol{b}_{\text{GLS}}$. From the results in Section 3.5 the optimal point forecast of $y_{*,n+1}$ is then

$$\hat{y}_{*,n+1} = \boldsymbol{x}_{*,n+1}'\boldsymbol{b}_{\text{GLS}} \tag{6.76}$$

which may be rewritten in terms of the original variables as

$$\hat{y}_{n+1} = \boldsymbol{x}_{n+1}'\boldsymbol{b}_{\text{GLS}} + \varphi(y_n - \boldsymbol{x}_n'\boldsymbol{b}_{\text{GLS}}) \tag{6.77}$$

The second term in this forecast is essentially an estimate of the conditional expectation of u_{n+1} because, from Eqs. (6.73) and (6.74),

$$E(u_{n+1} \mid u_n) = \varphi u_n = \varphi(y_n - \boldsymbol{x}_n'\boldsymbol{\beta})$$

and this term is estimated by $\varphi(y_n - \boldsymbol{x}_n'\boldsymbol{b}_{\text{GLS}})$. Again from Section 3.5, the forecast variance is

$$s_f^2 = s_\epsilon^2\left(1 + \boldsymbol{x}_{*,n+1}'(\boldsymbol{X}_*'\boldsymbol{X}_*)^{-1}\boldsymbol{x}_{*,n+1}\right) \tag{6.78}$$

where
$$s_\epsilon^2 = (\boldsymbol{y}_* - \boldsymbol{X}_*\boldsymbol{b}_{\text{GLS}})'(\boldsymbol{y}_* - \boldsymbol{X}_*\boldsymbol{b}_{\text{GLS}})/(n - k) \tag{6.79}$$

as in Eq. (5.27).

[27] C. M. Beach and J. G. MacKinnon, "A Maximum Likelihood Procedure for Regression with Autocorrelated Errors," *Econometrica*, **46**, 1978, 51–58.

The fly in the ointment is that the preceding has assumed the value of φ to be known, which is not usually the case. In practice $\boldsymbol{b}$ and φ have to be jointly estimated. The feasible forecast is then

$$\hat{y}_{n+1} = \boldsymbol{x}'_{n+1}\boldsymbol{b}_{\text{GLS}} + \hat{\varphi}(y_n - \boldsymbol{x}'_n\boldsymbol{b}_{\text{GLS}}\ ) \tag{6.80}$$

The properties of this forecast are no longer known exactly, nor is there any closed-form expression for the forecast variance. The variance in Eq. (6.78) is conditional on φ and does not take account of the uncertainty in estimating that parameter. Some software packages simply report s_ϵ^2 for the forecast variance, which may not be too serious an error if the disturbance variance is much greater than coefficient variance. One might use Eq. (6.78), which is simple enough to compute and attempts to cover most of the coefficient variance. A final possibility is to use **bootstrapping techniques** to establish sampling distributions. These techniques are discussed in Chapter 11.

EXAMPLE 6.2 AN AUTOCORRELATION EXERCISE. This exercise is based on artificial data for the years 1951 to 1990. The X variable is generated by the formula

$$X = 10 + 5 * \text{NRND}$$

with a starting value of 5 in 1950. NRND denotes a randomly distributed, standard normal variable. The Y variable is generated by the formula

$$Y = 2 + 2 * X - 0.5 * X(-1) + 0.7 * Y(-1) + 5 * \text{NRND}$$

The OLS estimation of this specification is shown in Table 6.5. The DW statistic does not reject the hypothesis of zero autocorrelation of the residuals, but the test is unreliable because of the presence of the lagged dependent variable among the regressors. Table 6.6 shows two test statistics for first-order autocorrelation along with the regression on which they are based. The F statistic reported at the top of Table 6.6 is that of the Durbin test in Eq. (6.51). Because we are only testing for first-order autocorrelation, the F

TABLE 6.5

A correctly specified equation

LS // Dependent Variable is Y
Sample: 1951–1990
Included observations: 40

Variable	Coefficient	Std. Error	T-Statistic	Prob.
C	−1.646026	3.299935	−0.498805	0.6210
X	2.024356	0.168530	12.01187	0.0000
X(−1)	−0.355027	0.216876	−1.637004	0.1103
Y(−1)	0.731749	0.066595	10.98803	0.0000

R-squared	0.893465	Mean dependent var	52.58359
Adjusted R-squared	0.884587	S.D. dependent var	15.91960
S.E. of regression	5.408291	Akaike info criterion	3.470506
Sum squared resid	1052.986	Schwarz criterion	3.639394
Log likelihood	−122.1677	F-statistic	100.6387
Durbin-Watson stat	2.181299	Prob(F-statistic)	0.000000

TABLE 6.6

Test of first-order autocorrelation

Breush-Godfrey Serial Correlation LM Test:

F-statistic	0.599158	Probability	0.444097
Obs*R-squared	0.673227	Probability	0.411929

Test Equation:
LS // Dependent Variable is RESID

Variable	Coefficient	Std. Error	T-Statistic	Prob.
C	−0.836639	3.490048	−0.239721	0.8119
X	0.016228	0.170768	0.095032	0.9248
X(−1)	−0.049645	0.227329	−0.218386	0.8284
Y(−1)	0.022589	0.073051	0.309217	0.7590
RESID(−1)	−0.141878	0.183292	−0.774053	0.4441

R-squared	0.016831	Mean dependent var	4.87E−15
Adjusted R-squared	−0.095532	S.D. dependent var	5.196118
S.E. of regression	5.438654	Akaike info criterion	3.503532
Sum squared resid	1035.263	Schwarz criterion	3.714642
Log likelihood	−121.8282	F-statistic	0.149790
Durbin-Watson stat	1.904415	Prob(F-statistic)	0.961856

TABLE 6.7

A misspecified relation

LS // Dependent Variable is Y
Sample: 1951–1990
Included observations: 40

Variable	Coefficient	Std. Error	T-Statistic	Prob.
C	33.79141	4.244113	7.961948	0.0000
X	1.861224	0.371823	5.005674	0.0000

R-squared	0.397368	Mean dependent var	52.58359
Adjusted R-squared	0.381510	S.D. dependent var	15.91960
S.E. of regression	12.51984	Akaike info criterion	5.103335
Sum squared resid	5956.360	Schwarz criterion	5.187779
Log likelihood	−156.8242	F-statistic	25.05677
Durbin-Watson stat	0.446263	Prob(F-statistic)	0.000013

statistic is the square of the t statistic attached to RESID(−1) in the regression; and the Prob values of the F and t statistics are, of course, identical. The other test statistic, nR^2, is the Breusch-Godfrey LM statistic defined in Eq. (6.56). Neither statistic rejects the hypothesis of zero first-order autocorrelation of the disturbance term.

If an investigator mistakenly specified Y as a function of the current X only, he or she would obtain the results shown in Table 6.7. The DW statistic indicates highly significant autocorrelation, which might lead the investigator to proceed with a Cochrane-Orcutt estimation, obtaining the results given in Table 6.8. Our fearless investigator has now achieved a much better R^2 and apparently strong evidence of autocorrelated disturbances. However, as explained in the development leading up to Eq. (6.63), the

TABLE 6.8
A Cochrane-Orcutt estimation

LS // Dependent Variable is Y
Sample: 1951–1990
Included observations: 40
Convergence achieved after 7 iterations

Variable	Coefficient	Std. Error	T-Statistic	Prob.
C	40.09970	5.276770	7.599289	0.0000
X	1.566337	0.212262	7.379250	0.0000
AR(1)	0.744400	0.093639	7.949678	0.0000

R-squared	0.783307	Mean dependent var	52.58359
Adjusted R-squared	0.771594	S.D. dependent var	15.91960
S.E. of regression	7.608265	Akaike info criterion	4.130509
Sum squared resid	2141.771	Schwarz criterion	4.257175
Log likelihood	−136.3677	F-statistic	66.87442
Durbin-Watson stat	1.696594	Prob(F-statistic)	0.000000
Inverted AR Roots	.74		

TABLE 6.9
A test of the Cochrane-Orcutt specification

Wald Test:

Null Hypothesis: C(3) + C(2)*C(4) = 0

F-statistic	36.60186	Probability	0.000001
Chi-square	36.60186	Probability	0.000000

Cochrane-Orcutt specification implies a nonlinear restriction on the parameters of the general relation with X, $X(-1)$, and $Y(-1)$ as regressors. The restriction may be tested on the regression in Table 6.5. The result is shown in Table 6.9. As was to be expected from our prior knowledge, the restriction is decisively rejected.

6.9 AUTOREGRESSIVE CONDITIONAL HETEROSCEDASTICITY (ARCH)

Traditionally, econometricians have been alert to the possibility of heteroscedastic disturbances in cross-section analyses and to autocorrelated disturbances in time series studies. In the former case all pairwise autocorrelations are assumed to be zero, and in the latter case homoscedasticity is assumed. In a seminal paper Engle suggested that heteroscedasticity might also occur in time series contexts.[28] Students

[28] Robert F. Engle, "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, **50**, 1982, 987–1008.

of forecasting, especially in speculative markets such as exchange rates and stock market returns, had noticed that large and small errors tended to occur in clusters. Engle formulated the notion that the recent past might give information about the conditional disturbance variance. He postulated the relation

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_p u_{t-p}^2 \tag{6.81}$$

The conditional disturbance variance is the variance of u_t, conditional on information available at time $t-1$; and it may be expressed here as

$$\begin{aligned} \sigma_t^2 &= \operatorname{var}(u_t \mid u_{t-1}, \ldots, u_{t-p}) \\ &= E(u_t^2 \mid u_{t-1}, \ldots, u_{t-p}) \\ &= E_{t-1}(u_t^2) \end{aligned} \tag{6.82}$$

where E_{t-1} indicates taking an expectation conditional on all information up to the end of period $t-1$. Recent disturbances thus influence the variance of the current disturbance, just as yesterday's earthquake brings a flurry of aftershocks today. A variance such as Eq. (6.81) can arise from a disturbance defined as

$$u_t = \epsilon_t[\alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_p u_{t-p}^2]^{1/2} \tag{6.83}$$

where $\{\epsilon_t\}$ is a white noise series with unit variance.[29] This is an ARCH(p) process. The simplest case is an ARCH(1) process, $u_t = \epsilon_t[\alpha_0 + \alpha_1 u_{t-1}^2]^{1/2}$. Its properties are derived in Appendix 6.3. They are as follows:

1. The u_t have zero mean.
2. The conditional variance is given by $\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2$, which checks with Eq. (6.81).
3. The *unconditional* variance is $\sigma^2 = \alpha_0/(1-\alpha_1)$, which only exists if $\alpha_0 > 0$ and $|\alpha_1| < 1$.
4. The autocovariances are zero.[30]

Testing for ARCH. The obvious test is implied by Eq. (6.81):

1. Fit y to X by OLS and obtain the residuals $\{e_t\}$.
2. Compute the OLS regression, $e_t^2 = \hat{\alpha}_0 + \hat{\alpha}_1 e_{t-1}^2 + \cdots + \hat{\alpha}_p e_{t-p}^2 +$ error.
3. Test the joint significance of $\hat{\alpha}_1, \ldots, \hat{\alpha}_p$.

If these coefficients are significantly different from zero, the assumption of conditionally homoscedastic disturbances is rejected in favor of ARCH disturbances, and the testing procedure can provide a tentative indication of the value of p. One should remember, however, that various specification errors in the original relation can give false indications of ARCH disturbances.

[29] The assumption of a unit variance is for simplicity. Any other variance could be rescaled to unity by suitable adjustment of the other parameters.

[30] This result seems plausible enough for the ARCH(1) process, but implausible for higher-order processes. However, see Appendix 6.3.

Estimation under ARCH. One possible estimation method is a form of feasible GLS. The regression just given in stage 2 is used to provide estimates of the disturbance variances at each sample point, and the original relation is then reestimated by the weighted least-squares procedure that corrects for the heteroscedasticity. This process can come to grief if the estimation procedure at stage 2 yields zero or negative variances. Suitable restrictions on the α parameters, however, can minimize the risk of breakdown. Some investigators impose a set of linearly declining weights on the lagged squared disturbances. If, for example, the value of p were 4, the regression in the foregoing step 2 would take the form

$$e_t^2 = \hat{\alpha}_0 + \hat{\alpha}(0.4e_{t-1}^2 + 0.3e_{t-2}^2 + 0.2e_{t-3}^2 + 0.1e_{t-4}^2) + \text{error}$$

A less restrictive specification of the disturbance equation is available in the GARCH formulation.[31] Bollerslev's suggestion is to replace Eq. (6.81) by

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_p u_{t-p}^2 + \gamma_1 \sigma_{t-1}^2 + \cdots + \gamma_q \sigma_{t-q}^2 \tag{6.84}$$

This is known as the GARCH(p, q) model. It expresses the conditional variance as a linear function of p lagged squared disturbances and q lagged conditional variances. Estimation is difficult for anything other than low values of p and q. In practice the most frequent application is the GARCH(1, 1) model,

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \gamma_1 \sigma_{t-1}^2$$

Substituting successively for the lagged disturbance on the right side gives

$$\sigma_t^2 = \frac{\alpha_0}{1-\gamma_1} + \alpha_1(u_{t-1}^2 + \gamma_1 u_{t-2}^2 + \gamma_1^2 u_{t-3}^2 + \gamma_1^3 u_{t-4}^2 + \cdots)$$

The current variance now depends on all previous squared disturbances; and, provided γ_1 is a positive fraction, the weights decline exponentially.

The asymptotically efficient estimator is maximum likelihood, which gives rise to nonlinear equations requiring iterative treatment; but we will not go into the details here.[32] There has been an explosion of ARCH models in the literature. One of the more important is the ARCH in MEAN, or ARCH-M Model.[33] The crucial feature of the model is the inclusion of the conditional variance as a regressor in a financial equation, thus allowing expected risk on an asset to be reflected in the asset price. A comprehensive set of estimation and testing procedures for ARCH models is available in the 1996 version of EViews.

[31] Tim Bollerslev, "Generalized Autoregressive Conditional Heteroscedasticity," *Journal of Econometrics,* **31**, 1986, 307–327.

[32] See, for example, Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics,* Oxford University Press, 1993, 556–560; or William H. Greene, *Econometric Analysis,* 2nd ed., Macmillan, 1993, 438–442 and 568–577.

[33] Robert F. Engle, David M. Lilien, and Russell P. Robins, "Estimating Time Varying Risk Premia in the Term Structure: The ARCH-M Model," *Econometrica,* **55**, 1987, 391–407.

APPENDIX

APPENDIX 6.1
LM test for multiplicative heteroscedasticity[34]

Consider the model

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + u_t$$

where $\boldsymbol{x}_t' = (1 \; x_{2t} \; x_{3t} \; \cdots \; x_{kt})$. Instead of the usual assumption of homoscedasticity, we now assume

$$Eu_t = 0 \qquad \text{for all } t$$

$$\sigma_t^2 = Eu_t^2 = e^{z_t'\alpha}$$

where $z_t' = (1 \; z_{2t} \; \cdots \; z_{pt})$ is a vector of known variables, possibly including some of the x variables or functions of the x variables, and $\boldsymbol{\alpha} = (\alpha_1 \; \alpha_2 \; \cdots \; \alpha_p)$ is a vector of unknown parameters. The null hypothesis of homoscedasticity then takes the form

$$H_0\!: \alpha_2 = \cdots = \alpha_p = 0$$

in which case $\sigma_t^2 = e^{\alpha_1}$, a constant. By assuming normally distributed disturbances,

$$f(u_t) = \frac{1}{\sqrt{2\pi\sigma_t^2}} e^{-u_t^2/2\sigma_t^2}$$

The log-likelihood for this model is then

$$l = -\frac{n}{2}\ln 2\pi - \frac{1}{2}\sum \ln\sigma_t^2 - \frac{1}{2}\sum \frac{u_t^2}{\sigma_t^2}$$

The information matrix $\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is block diagonal. Thus we need only concentrate on $\partial l/\partial\boldsymbol{\alpha}$ and the submatrix $\boldsymbol{I}_{\alpha\alpha} = -E(\partial^2 l/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}')$. To obtain $\partial l/\partial\boldsymbol{\alpha}$ we need some intermediate results, namely,

$$\frac{\partial\sigma_t^2}{\partial\boldsymbol{\alpha}} = \sigma_t^2 z_t$$

$$\frac{\partial \ln\sigma_t^2}{\partial\boldsymbol{\alpha}} = z_t$$

$$\frac{\partial(\sigma_t^2)^{-1}}{\partial\boldsymbol{\alpha}} = -\frac{z_t}{\sigma_t^2}$$

Then

$$\frac{\partial l}{\partial\boldsymbol{\alpha}} = \frac{1}{2}\sum\left(\frac{u_t^2}{\sigma_t^2} - 1\right)z_t$$

[34] This section is based on L. Godfrey, "Testing for Multiplicative Heteroscedasticity," *Journal of Econometrics*, **8**, 1978, 227–236. T. S. Breusch and A. R. Pagan, "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, **47**, 1979, 1287–1294, shows that the same test procedure holds for more general specifications of heteroscedasticity than the one considered in this section.

and
$$\frac{\partial^2 l}{\partial\alpha\partial\alpha'} = -\frac{1}{2}\sum\frac{u_t^2}{\sigma_t^2}z_t z_t'$$

Taking expectations, we write

$$I_{\alpha\alpha} = -E\left(\frac{\partial^2 l}{\partial\alpha\partial\alpha'}\right) = \frac{1}{2}\sum z_t z_t'$$

because $Eu_t^2 = \sigma_t^2$ for all t. Rewrite $\partial l/\partial\alpha$ as

$$\frac{\partial l}{\partial\alpha} = \frac{1}{2}\sum f_t z_t$$

where
$$f_t = g_t - 1 = \frac{u_t^2}{\sigma_t^2} - 1$$

Then
$$\begin{aligned} s(\alpha)'I_{\alpha\alpha}^{-1}s(\alpha) &= \left(\frac{1}{2}\sum f_t z_t\right)'\left(\frac{1}{2}\sum z_t z_t'\right)^{-1}\left(\frac{1}{2}\sum f_t z_t\right) \\ &= \frac{1}{2}\left(\sum f_t z_t\right)'\left(\sum z_t z_t'\right)^{-1}\left(\sum f_t z_t\right) \end{aligned}$$

This statistic measures one-half the explained sum of squares (ESS) from the regression of f_t on z_t. To obtain the LM test statistic one must evaluate the previous expression at the restricted estimates. In this case

$$\tilde{f}_t = \tilde{g}_t - 1 = \frac{e_t^2}{\tilde{\sigma}^2} - 1$$

where e_t is the residual from the OLS regression of y_t on x_t and $\tilde{\sigma}^2 \sum e_t^2/n$. Since $\tilde{f}$ and $\tilde{g}$ merely differ by a constant, the ESS from the regression of $\tilde{g}$ on z will be equal to the ESS from the regression of $\tilde{f}$ on z. This latter regression is the Godfrey test for multiplicative heteroscedasticity. The *squared* OLS residuals, divided by the estimated residual variance, are regressed on the z variables; under H_0, one-half of the resultant ESS will be asymptotically distributed as $\chi^2(p-1)$.

To derive an asymptotically equivalent test statistic, return to the regression of $\tilde{f}$ on z_t. The $\tilde{f}$ variable has zero mean. Thus ESS $= R^2 \sum \tilde{f}_t^2$, and

$$\sum \tilde{f}_t^2 = \sum\left(\frac{e_t^2}{\tilde{\sigma}^2} - 1\right)^2 = \frac{1}{\tilde{\sigma}^4}\sum e_t^4 - \frac{2}{\tilde{\sigma}^2}\sum e_t^2 + n$$

Dividing by n, we see that

$$\frac{1}{n}\sum \tilde{f}_t^2 = \frac{m_4}{m_2^2} - 2 + 1$$

where m_2 and m_4 denote the second and fourth sample moments about the mean of the OLS residuals. For a normally distributed variable the corresponding population moments obey the relation $\mu_4 = 3\mu_2^2$. Replacing the sample moments by the population equivalents gives $\sum \tilde{f}_t^2 \simeq 2n$. Thus

$$\mathrm{LM} = \tfrac{1}{2}\mathrm{ESS} = nR^2$$

Finally, we note that the process of multiplying the depending variable by a constant, adding a constant to the dependent variable, or both, does not change the regression R^2. Thus the R^2 for the LM statistic may be obtained by just regressing e_t^2 on the z variables. Breusch and Pagan show that if the heteroscedasticity is replaced by the more general formulation $\sigma_t^2 = h(z_t'\boldsymbol{\alpha})$, where $h(\ )$ denotes some unspecified functional form, the same LM test statistic still applies, so this is a very general test for heteroscedasticity.

APPENDIX 6.2
LR test for groupwise homoscedasticity

We start with the log-likelihood given in Eq. (6.16), namely,

$$l = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{V}| - \frac{1}{2}\boldsymbol{u}'\boldsymbol{V}^{-1}\boldsymbol{u}$$

When $\boldsymbol{V} = \sigma^2\boldsymbol{I}$, this becomes

$$l = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\boldsymbol{u}'\boldsymbol{u}$$

As shown in Section 5.2, maximization gives

$$L_{\text{rest}} = L(\hat{\beta}, \hat{\sigma}^2) = (2\pi e)^{-n/2}(\hat{\sigma}^2)^{-n/2}$$

where $\hat{\sigma}^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n$. This is the *restricted* likelihood, since the variances have been assumed equal. Taking logs gives

$$l_{\text{rest}} = -\frac{n}{2}[\ln(2\pi) + 1] - \frac{n}{2}\ln(\hat{\sigma}^2)$$

When groupwise heteroscedasticity is allowed, as in Eq. (6.15), the log-likelihood is

$$l = -\frac{n}{2}\ln(2\pi) - \sum_{i=1}^{g}\frac{n_i}{2}\ln\sigma_i^2 - \frac{1}{2}\sum_{i=1}^{g}\frac{1}{\sigma_i^2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$$

The MLEs of the variances are

$$\hat{\sigma}_i^2 = (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})'(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})/n_i \qquad i = 1, 2, \ldots, g$$

where $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$. Substitution in the log-likelihood gives the *unrestricted* log-likelihood as

$$l_{\text{unrest}} = -\frac{n}{2}[\ln(2\pi) + 1] - \frac{1}{2}\sum_{i=1}^{g} n_i \ln\hat{\sigma}_i^2$$

The LR test statistic is

$$\mathrm{LR} = -2\ln\left(\frac{L_{\text{rest}}}{L_{\text{unrest}}}\right) = 2(l_{\text{unrest}} - l_{\text{rest}})$$

The relevant substitution then gives

$$\mathrm{LR} = n \ln \hat{\sigma}^2 - \sum_{i=1}^{g} n_i \ln \hat{\sigma}_i^2$$

which will be asymptotically as $\chi^2(g - 1)$ under the null hypothesis.

APPENDIX 6.3
Properties of the ARCH(1) process

The process is specified as

$$u_t = \epsilon_t[\alpha_0 + \alpha_1 u_{t-1}^2]^{1/2}$$

with the $\{\epsilon_t\}$ iid(0, 1). The development requires the use of both *conditional* and *unconditional* expectations. $E_{t-1}(u_t)$ denotes the expectation of u_t, conditional on the information available at time $t - 1$. For the first-order process this is equivalent to $E(u_t \mid u_{t-1})$. The unconditional expectation may be found by repeatedly taking conditional expectations period by period. Looking at the conditional means of the process, we write

$$E_{t-1}(u_t) = [\alpha_0 + \alpha_1 u_{t-1}^2]^{1/2} E_{t-1}(\epsilon_t) = 0$$

It then follows that $E_{t-2}[E_{t-1}(u_t)] = E_{t-2}(0) = 0$ and so on for all earlier periods. Thus the unconditional mean is $E(u_t) = 0$.

Turning to the variance, we have

$$u_t^2 = \epsilon_t^2[\alpha_0 + \alpha_1 u_{t-1}^2]$$

Thus, $$E_{t-1}(u_t^2) = \sigma_\epsilon^2[\alpha_0 + \alpha_1 u_{t-1}^2] = \alpha_0 + \alpha_1 u_{t-1}^2$$

Similarly, $$\begin{aligned} E_{t-2}E_{t-1}(u_t^2) &= E_{t-2}[\alpha_0 + \alpha_1 u_{t-1}^2] \\ &= \alpha_0 + \alpha_1 E_{t-2}(u_{t-1}^2) \\ &= \alpha_0 + \alpha_1(\alpha_0 + \alpha_1 u_{t-2}^2) \\ &= \alpha_0 + \alpha_0\alpha_1 + \alpha_1^2 u_{t-2}^2 \end{aligned}$$

Proceeding in this way, we see

$$E_0 \cdots E_{t-2}E_{t-1}(u_t^2) = \alpha_0(1 + \alpha_1 + \alpha_1^2 + \cdots + \alpha_1^{t-1}) + \alpha_1^t u_0^2$$

Provided $|\alpha_1| < 1$, we can take limits as $t \to \infty$, and the unconditional variance is

$$\mathrm{var}(u_t) = \sigma^2 = \frac{\alpha_0}{1 - \alpha_1}$$

which will be positive provided $\alpha_0 > 0$. The process is thus homoscedastic.

The conditional first-order autocovariance is

$$E_{t-1}(u_t u_{t-1}) = u_{t-1}E_{t-1}(u_t) = 0$$

It follows trivially that all higher-order autocovariances are also zero.

In the case of the pth-order process, it is clear that the zero mean condition still holds. The unconditional variance will be a more complicated function of the α parameters. The condition $E_{t-1}(u_t u_{t-1}) = 0$ holds whatever the order of the process, and so the autocovariances are zero for ARCH processes.

PROBLEMS

6.1. Five sample observations are

X	4	1	5	8	2
Y	6	3	12	15	4

Assume a linear model with heteroscedasticity taking the form $\sigma_t^2 = \sigma^2 X_t^2$. Calculate the GLS estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and the corresponding standard errors.

6.2. An investigator estimates a linear relation and the associated standard errors by applying OLS to the data:

X	2	3	1	5	9
Y	4	7	3	9	17

She is subsequently informed that the variance matrix for the disturbances underlying the data is

$$\text{var}(\boldsymbol{u}) = \sigma^2 \cdot \text{diag}\{0.10, 0.05, 0.20, 0.30, 0.15\}$$

Use this information to calculate the correct standard errors for the OLS estimates and compare with those obtained from the conventional formula.

6.3. Sketch the feasible GLS procedure for estimating

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + u_t$$

where the variance of u_t is

$$\sigma_t^2 = e^{\alpha z_t}$$

for some known scalar variable z_t.

6.4. Take another sample of 100 observations from CPS88 and carry out the test procedures illustrated in Example 6.1.

6.5. Using Potexp as a sorting variable, partition the 1000 observations in CPS88 into four groups and carry out the test of groupwise homoscedasticity defined in Eq. (6.21). [*Hint:* It will suffice to estimate the $\boldsymbol{\beta}$ vector separately for each group, as explained in the paragraph leading up to Eq. (6.21), and not proceed to full iteration.]

6.6. Derive the log-likelihood in Eq. (6.72).

6.7. Take the model in Appendix 6.1 and assume that heteroscedasticity has the form

$$\sigma_t^2 = \alpha_0 + \alpha_1 z_{1t} + \cdots + \alpha_p z_{pt}$$

Derive the LM test for $H_0: \alpha_1 = \cdots = \alpha_p = 0$ and compare with the test statistic in Appendix 6.1.

6.8. Prove the assertion in Appendix 6.1 that the process of multiplying the dependent variable by a constant, adding a constant to the dependent variable, or both, does not change the regression R^2. (*Hint:* Define $z = c_1\boldsymbol{y} + c_2\boldsymbol{i}$ where the c's are arbitrary constants and $\boldsymbol{i}$ is a column of ones. Then work in terms of deviations to show $R_{y \cdot X}^2 = R_{z \cdot X}^2$.)

6.9. Derive the results in Eqs. (6.31) and (6.33) for the AR(1) process by the method of iterated expectations illustrated in Appendix 6.3.

6.10. Generate data of your own choosing and experiment with the autocorrelation procedures outlined in Example 6.2.

CHAPTER 7

Univariate Time Series Modeling

This chapter is solely concerned with **time series modeling.** In the univariate case a series is modeled only in terms of its own past values and some disturbance. The general expression is

$$x_t = f(x_{t-1}, x_{t-2}, \ldots, u_t) \tag{7.1}$$

To make Eq. (7.1) operational one must specify three things: the functional form $f(\,)$, the number of lags, and a structure for the disturbance term. If, for example, one specified a linear function with one lag and a white noise disturbance, the result would be the first-order, autoregressive AR(1), process, where for simplicity the intercept is suppressed,

$$x_t = \alpha x_{t-1} + u_t \tag{7.2}$$

This process has already been introduced in Sections 2.5 and 6.6. The general pth-order, AR(p) process is

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} + u_t \tag{7.3}$$

When the disturbance is white noise Eq. (7.3) is a *pure* AR(p) process. This process can be enriched by assuming some more complicated structure for the disturbance term. When u is not assumed to be white noise, the usual alternative specification is a **moving average,** MA(q), process,

$$u_t = \epsilon_t - \beta_1 \epsilon_{t-1} - \cdots - \beta_q \epsilon_{t-q} \tag{7.4}$$

where ϵ is a white noise process. Equation (7.4) specifies a *pure* MA(q) process. Combining Eqs. (7.3) and (7.4) gives a *mixed* autoregressive, moving average, ARMA(p, q) process,

$$x_t = \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} + \epsilon_t - \beta_1 \epsilon_{t-1} - \cdots - \beta_q \epsilon_{t-q} \tag{7.5}$$

7.1 A RATIONALE FOR UNIVARIATE ANALYSIS

At first sight it would not seem very sensible for economists to pay much attention to univariate analysis. After all, economic theory is rich in suggestions for relationships *between* variables. Thus, attempting to explain and forecast a series using only information about the history of that series would appear to be an inefficient procedure, because it ignores the potential information in related series. There are two possible rationales. The first is that a priori information about the possible relationships between series may not be well founded. In such a case a purely *statistical* model relating current to previous values may be a useful summary device and may be used to generate reliable short-term forecasts. Alternatively, if theoretical speculations about the economic structure are well founded, it can be shown that one manifestation of that structure yields equations similar to Eq. (7.5) for each of the endogenous variables in the structure. To illustrate this last point consider the simplest macro model of the elementary textbooks:

$$\begin{aligned} C_t &= \alpha_0 + \alpha_1 Y_t + \alpha_2 C_{t-1} + u_t \\ Y_t &\equiv C_t + I_t \end{aligned} \tag{7.6}$$

where C, I, and Y denote consumption, investment, and national income. Consumption is linearly related to current income and to consumption in the previous period with some disturbance. The second relation is the national income identity. Mathematically this system of two equations in three variables "explains" any two variables in terms of the third variable. In economics we traditionally regard C and Y as the variables whose course is determined by the movements in I and the disturbances. C and Y are then termed **endogenous variables, and I is termed exogenous.** Substituting the second equation in the first gives

$$C_t - \frac{\alpha_2}{1-\alpha_1}C_{t-1} = \frac{\alpha_0}{1-\alpha_1} + \frac{\alpha_1}{1-\alpha_1}I_t + \frac{1}{1-\alpha_1}u_t \tag{7.7a}$$

Some algebra then gives[1]

$$Y_t - \frac{\alpha_2}{1-\alpha_1}Y_{t-1} = \frac{\alpha_0}{1-\alpha_1} + \frac{1}{1-\alpha_1}(I_t - \alpha_2 I_{t-1}) + \frac{1}{1-\alpha_1}u_t \tag{7.7b}$$

Thus, C and Y both have an AR(1) component with the same coefficient on the lagged term. The right side of each equation is an omnibus disturbance term whose properties depend on the behavior of I. If I were a white noise series about some mean, then consumption would follow a pure AR(1) process and income, an ARMA(1,1) process.

The classification into endogenous and exogenous variables is not set in stone but depends on the objectives of the modeler. Suppose we extend the macro model

[1] Lag the income identity one period, multiply by $\alpha_2/(1-\alpha_1)$, and subtract the product from the original identity.

to include an equation for I. The new specification is

$$\begin{aligned} C_t &= \alpha_0 + \alpha_1 Y_t + \alpha_2 C_{t-1} + u_t \\ I_t &= \beta_0 + \beta_1(Y_{t-1} - Y_{t-2}) + v_t \\ Y_t &\equiv C_t + I_t + G_t \end{aligned} \tag{7.8}$$

The consumption function remains the same, but investment is now specified to depend on a lagged change in income levels. Government expenditures G appear as a new exogenous variable. Recasting this model by algebraic substitution is tedious, so we change to a matrix formulation and also introduce the lag operator.

7.1.1 The Lag Operator

The lag operator L, when placed in front of any variable with a time subscript, gives the previous value of the series.[2] Thus,

$$\begin{aligned} L(x_t) &= x_{t-1} \\ L^2(x_t) &= L[L(x_t)] = L(x_{t-1}) = x_{t-2} \\ L^s x_t &= x_{t-s} \\ (1-L)x_t &= x_t - x_{t-1} = \Delta x_t \\ L(1-L)x_t &= x_{t-1} - x_{t-2} = \Delta x_{t-1} \end{aligned} \tag{7.9}$$

where Δ is the first difference operator. In many algebraic manipulations the lag operator may be treated as a scalar. One of the most important operations is taking the inverse of an expression in L. For example, let $A(L) = 1 - \alpha L$ denote a first-order polynomial in L. Consider the multiplication

$$(1-\alpha L)(1 + \alpha L + \alpha^2 L^2 + \alpha^3 L^3 + \cdots + \alpha^p L^p) = 1 - \alpha^{p+1} L^{p+1}$$

As $p \to \infty$, $\alpha^{p+1}L^{p+1} \to 0$ provided $|\alpha| < 1$. We may then write the reciprocal, or **inverse**, of $A(L)$ as

$$A^{-1}(L) = \frac{1}{(1-\alpha L)} = 1 + \alpha L + \alpha^2 L^2 + \alpha^3 L^3 + \cdots \tag{7.10}$$

Using the lag operator, we may rewrite the model in Eq. (7.8) as

$$\begin{bmatrix} (1-\alpha_2 L) & 0 & -\alpha_1 \\ 0 & 1 & -\beta_1 L(1-L) \\ -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} C_t \\ I_t \\ Y_t \end{bmatrix} = \begin{bmatrix} \alpha_0 & 0 \\ \beta_0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} D_t \\ G_t \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \\ 0 \end{bmatrix} \tag{7.11}$$

where D_t is a dummy variable that takes the value of one to accommodate the intercept term. We now write this system in matrix form as

$$\boldsymbol{A}(L)\boldsymbol{x}_t = \boldsymbol{B}\boldsymbol{z}_t + \boldsymbol{w}_t \tag{7.12}$$

[2] In time series literature it is common to find this operator denoted by B (backward shift operator). L is more common in the econometrics literature, and we will adhere to that convention.

where the correspondences should be obvious. $A(L)$ is the 3×3 matrix whose elements are polynomials (some of degree zero) in the lag operator. Its inverse may be written

$$A^{-1}(L) = \frac{1}{|A(L)|}C(L) \tag{7.13}$$

where $|A(L)|$ is the determinant and $C(L)$ is the matrix of cofactors. Combining Eqs. (7.12) and (7.13) gives

$$|A(L)|\boldsymbol{x}_t = \boldsymbol{C}(L)\boldsymbol{B}\boldsymbol{z}_t + \boldsymbol{C}(L)\boldsymbol{w}_t \tag{7.14}$$

The crucial property of Eq. (7.14) is that each endogenous variable in $\boldsymbol{x}_t$ is multiplied by the determinant, which is a scalar polynomial in the lag operator. For this model

$$|A(L)| = (1 - \alpha_1) - (\alpha_2 + \beta_1)L + \beta_1(1 + \alpha_2)L^2 - \alpha_2\beta_1L^3 \tag{7.15}$$

Thus each endogenous variable has the same third-order autoregressive component. The nature of the disturbance term in the AR equation will depend on the exogenous variables on the right-hand side of Eq. (7.14). If a white noise assumption is not appropriate, one should fit a general ARMA model of the type shown in Eq. (7.5).

7.1.2 ARMA Modeling

There are three steps in ARMA modeling:

1. Check the series for stationarity, and, if necessary, transform the series to induce stationarity.
2. From the autocorrelation properties of the transformed series choose a few ARMA specifications for estimation and testing in order to arrive at a preferred specification with white noise residuals.
3. Calculate forecasts over a relevant time horizon from the preferred specification.

We will concentrate on stage 2 first before looking at stage 1. The basic idea is to derive the autocorrelation patterns associated with various low-order AR, MA, and ARMA schemes. Comparing these with the empirical patterns computed from the series under analysis then suggests one or more ARMA specifications for statistical estimation and testing.

7.2 PROPERTIES OF AR, MA, AND ARMA PROCESSES

7.2.1 AR(1) Process

We will first derive the properties of the AR(1) process. The main results were obtained in Section 2.5; but we will rederive them here using a different method, which will have fruitful applications later. The specification is

$$y_t = m + \alpha y_{t-1} + \epsilon_t \tag{7.16}$$

where ϵ is a white noise process. Using the lag operator we rewrite this equation as

$$(1 - \alpha L)y_t = m + \epsilon_t$$

which gives $$y_t = (1 + \alpha L + \alpha^2 L^2 + \cdots)(m + \epsilon_t)$$

Since a constant, like m, has the same value at all periods, application of the lag operator any number of times simply reproduces the constant. Thus we can write

$$y_t = (1 + \alpha + \alpha^2 + \cdots)m + (\epsilon_t + \alpha\epsilon_{t-1} + \alpha^2\epsilon_{t-2} + \cdots)$$

Provided $|\alpha| < 1$ this gives

$$E(y_t) = \frac{m}{1-\alpha} = \mu \tag{7.17}$$

that is, y has a constant unconditional mean, independent of time. As shown in Section 2.5, the same condition on α gives y a constant unconditional variance, namely,

$$\sigma_y^2 = E(y_t - \mu)^2 = \frac{\sigma_\epsilon^2}{1-\alpha^2} \tag{7.18}$$

This variance can be derived in an alternative fashion that also facilitates the derivation of autocovariances. By using Eq. (7.17), it is possible to rewrite Eq. (7.16) as

$$x_t = \alpha x_{t-1} + \epsilon_t \tag{7.19}$$

where $x_t = y_t - \mu$. Squaring both sides of Eq. (7.19) and taking expectations, we find

$$E(x_t^2) = \alpha^2 E(x_{t-1}^2) + E(\epsilon_t^2) + 2\alpha E(x_{t-1}\epsilon_t)$$

The last term on the right-hand side vanishes, since x_{t-1} depends only on ϵ_{t-1}, $\epsilon_{t-2}, \ldots$ and ϵ_t is uncorrelated with all previous values by the white noise assumption. When α satisfies the stationarity condition, $|\alpha| < 1$,

$$\sigma_y^2 = E(x_t^2) = E(x_{t-1}^2) = \cdots$$

and the previous equation becomes

$$\sigma_y^2 = \alpha^2\sigma_y^2 + \sigma_\epsilon^2$$

which reconfirms Eq. (7.18).

The process of multiplying both sides of Eq. (7.19) by x_{t-1} and taking expectations gives

$$E(x_t x_{t-1}) = \alpha E(x_{t-1}^2) + E(x_{t-1}\epsilon_t)$$

After denoting **autocovariance coefficients** by $\gamma_s = E(x_t x_{t-s})$, this last equation gives

$$\gamma_1 = \alpha\gamma_0$$

where γ_0 is another symbol for σ_y^2. In a similar fashion, multiplying Eq. (7.19) by x_{t-2}, followed by taking expectations, gives

$$\gamma_2 = \alpha\gamma_1$$

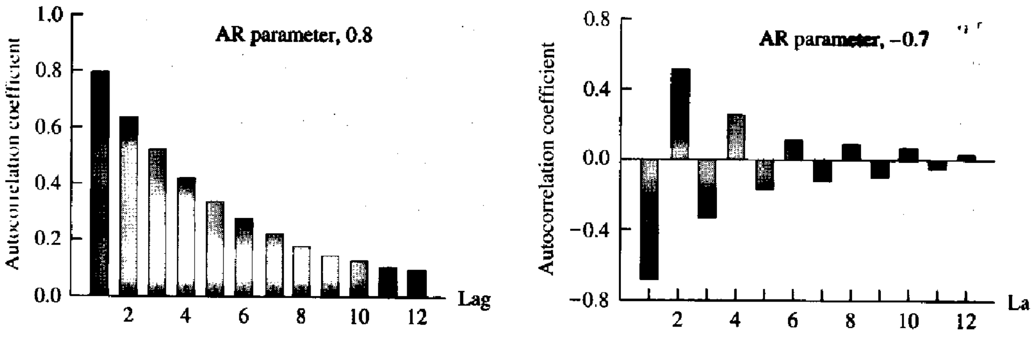


FIGURE 7.1
Correlograms for stationary AR(1) series.

and, in general, $$\gamma_k = \alpha\gamma_{k-1} = \alpha^k\gamma_0 \qquad k = 1, 2, \ldots \tag{7.20}$$

The **autocorrelation coefficients** for a stationary series are defined by

$$\rho_k = \frac{E(x_t x_{t-k})}{\sqrt{\text{var}(x_t)}\sqrt{\text{var}(x_{t-k})}} = \frac{\gamma_k}{\gamma_0} \tag{7.21}$$

The autocovariance and autocorrelation coefficients are symmetrical about lag zero. Thus we need only look at positive lags. The autocorrelation coefficients for the AR(1) process are

$$\rho_k = \alpha\rho_{k-1} = \alpha^k \qquad k = 1, 2, \ldots \tag{7.22}$$

The formula giving the autocorrelation coefficients is known as the **autocorrelation function** of the series, with the abbreviation *acf*, and its graphical representation is known as the *correlogram*. Figure 7.1 shows two correlograms for stationary AR(1) series.

7.2.2 AR(2) Process

The AR(2) process is defined as

$$y_t = m + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \epsilon_t \tag{7.23}$$

By assuming stationarity, the unconditional mean is $\mu = m/(1 - \alpha_1 - \alpha_2)$. Defining $x_t = y_t - \mu$ as before, we may rewrite Eq. (7.23) as

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \epsilon_t \tag{7.24}$$

If we multiply both sides by x_t and take expectations,

$$\gamma_0 = \alpha_1\gamma_1 + \alpha_2\gamma_2 + E(x_t\epsilon_t)$$

From Eq. (7.24) it is clear that $E(x_t\epsilon_t) = \sigma_\epsilon^2$ and so

$$\gamma_0 = \alpha_1\gamma_1 + \alpha_2\gamma_2 + \sigma_\epsilon^2 \tag{7.25}$$

Multiplying Eq. (7.24) by x_{t-1} and x_{t-2} and taking expectations in the usual way, we find

$$\begin{aligned} \gamma_1 &= \alpha_1\gamma_0 + \alpha_2\gamma_1 \\ \gamma_2 &= \alpha_1\gamma_1 + \alpha_2\gamma_0 \end{aligned} \tag{7.26}$$

Substituting Eq. (7.26) in Eq. (7.25) and simplifying, we see that

$$\gamma_0 = \frac{(1-\alpha_2)\sigma_\epsilon^2}{(1+\alpha_2)(1-\alpha_1-\alpha_2)(1+\alpha_1-\alpha_2)}$$

Under stationarity this variance must be a constant, positive number. Requiring each term in parentheses to be positive gives

$$\begin{aligned} \alpha_2 + \alpha_1 &< 1 \\ \alpha_2 - \alpha_1 &< 1 \\ |\alpha_2| &< 1 \end{aligned} \tag{7.27}$$

These are the stationarity conditions for the AR(2) process.

The relations between autocovariances in Eq. (7.26) may be restated in terms of autocorrelation coefficients, namely,

$$\begin{aligned} \rho_1 &= \alpha_1 + \alpha_2\rho_1 \\ \rho_2 &= \alpha_1\rho_1 + \alpha_2 \end{aligned} \tag{7.28}$$

These are the **Yule-Walker** equations for the AR(2) process. Solving for the first two autocorrelation coefficients gives

$$\rho_1 = \frac{\alpha_1}{1-\alpha_2} \qquad \rho_2 = \frac{\alpha_1^2}{1-\alpha_2} + \alpha_2 \tag{7.29}$$

The acf for the AR(2) process is

$$\rho_k = \alpha_1\rho_{k-1} + \alpha_2\rho_{k-2} \qquad k = 3, 4, \ldots \tag{7.30}$$

This is a second-order difference equation with the first two values given in Eq. (7.29). Moreover, the coefficients of the difference equation are those of the AR(2) process. Thus the stationarity conditions ensure that the acf dies out as the lag increases. The acf will be a damped exponential or, if the roots of Eq. (7.30) are complex, a damped sine wave.

Roots of the polynomial in the lag operator

An alternative and enlightening view of the AR(2) process is obtained by recasting it in lag operator notation. Write Eq. (7.24) as

$$A(L)x_t = \epsilon_t$$

where

$$A(L) = 1 - \alpha_1 L - \alpha_2 L^2$$

Express this quadratic as the product of two factors,

$$A(L) = 1 - \alpha_1 L - \alpha_2 L^2 = (1-\lambda_1 L)(1-\lambda_2 L)$$

The connection between the α and the λ parameters is

$$\lambda_1 + \lambda_2 = \alpha_1 \qquad \text{and} \qquad \lambda_1\lambda_2 = -\alpha_2 \tag{7.31}$$

The λ's may be seen as the roots of $\lambda^2 - \alpha_1\lambda - \alpha_2 = 0$, which is the **characteristic equation** of the second-order process. Its roots are

$$\lambda_1, \lambda_2 = \frac{\alpha_1 \pm \sqrt{\alpha_1^2 + 4\alpha_2}}{2}$$

and these satisfy Eq. (7.31). The inverse $A^{-1}(L)$ may be written

$$A^{-1}(L) = \frac{1}{(1-\lambda_1 L)(1-\lambda_2 L)} = \frac{c}{1-\lambda_1 L} + \frac{d}{1-\lambda_2 L}$$

where $c = -\lambda_1/(\lambda_2 - \lambda_1)$ and $d = \lambda_2/(\lambda_2 - \lambda_1)$. Then

$$x_t = A^{-1}(L)\epsilon_t = \frac{c}{1-\lambda_1 L}\epsilon_t + \frac{d}{1-\lambda_2 L}\epsilon_t$$

From the results on the AR(1) process, stationarity of the AR(2) process requires

$$|\lambda_1| < 1 \qquad \text{and} \qquad |\lambda_2| < 1 \tag{7.32}$$

Restating these conditions in terms of the α parameters gives the stationarity conditions already derived in Eq. (7.27).

The AR(2) case allows the possibility of a pair of complex roots, which will occur if $\alpha_1^2 + 4\alpha_2 < 0$. The roots may then be written

$$\lambda_1, \lambda_2 = h \pm vi$$

where h and v are the real numbers $h = \alpha_1/2$, $v = \frac{1}{2}\sqrt{-(\alpha_1^2 + 4\alpha_2)}$, and i is the imaginary number $i = \sqrt{-1}$, giving $i^2 = -1$. The autocorrelation coefficients will now display sine wave fluctuations, which will dampen toward zero provided the complex roots have *moduli* less than one. The absolute value or *modulus* of each complex root is

$$|\lambda_j| = \sqrt{h^2 + v^2} = -\alpha_2 \qquad j = 1,2$$

giving $0 < -\alpha_2 < 1$ as the condition for the correlogram to be a damped sine wave.

For real or complex roots the stationarity condition is that the moduli of the roots should be less than one, or that the roots lie *within the unit circle.* An alternative statement is that the roots of $A(z) = 1 - \alpha_1 z - \alpha_2 z^2$ should lie ***outside the unit circle.*** The roots of $A(z)$ are the values of z that solve the equation

$$A(z) = 1 - \alpha_1 z - \alpha_2 z^2 = (1 - \lambda_1 z)(1 - \lambda_2 z) = 0$$

The roots are obviously $z_j = 1/\lambda_j$ $(j = 1,2)$ so that, if the λ's lie within the unit circle, the z's lie outside the unit circle. The stationarity condition is commonly stated in the literature as the roots of the relevant polynomial in the lag operator lying *outside the unit circle.* Figure 7.2 shows two correlograms for stationary AR(2) series.

Partial autocorrelation function (pacf)

It is sometimes difficult to distinguish between AR processes of different orders solely on the basis of the correlograms. A sharper discrimination is possible on the

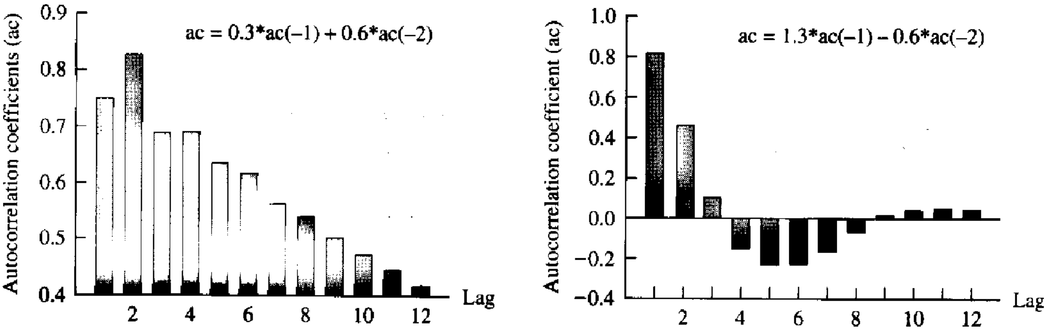


FIGURE 7.2
Correlograms of stationary AR(2) series.

basis of partial correlation coefficients. In an AR(2) process the α_2 parameter is the partial correlation between x_t and x_{t-2} with x_{t-1} held constant. To see this, we recall the definition of a partial correlation coefficient in a three-variable case, given in Eq. (3.15), namely,

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}}$$

These correlation coefficients take a special form for a stationary series. Let 1, 2, and 3 denote x and its first and second lags, respectively. Then

$$r_{12} = \operatorname{corr}(x_t, x_{t-1}) = \operatorname{corr}(x_{t-1}, x_{t-2}) = r_{23} = \rho_1$$

$$r_{13} = \operatorname{corr}(x_t, x_{t-2}) = \rho_2$$

Substitution in the previous formula gives

$$r_{13.2} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$$

If we return to the Yule-Walker equations in (7.28) and solve for α_2 the result is

$$\alpha_2 = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} = r_{13.2}$$

The Yule-Walker equations for an AR(3) process, $x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \alpha_3 x_{t-3} + \epsilon_t$, are

$$\begin{aligned} \rho_1 &= \alpha_1 + \alpha_2\rho_1 + \alpha_3\rho_2 \\ \rho_2 &= \alpha_1\rho_1 + \alpha_2 + \alpha_3\rho_1 \\ \rho_3 &= \alpha_1\rho_2 + \alpha_2\rho_1 + \alpha_3 \end{aligned} \tag{7.33}$$

The α_3 parameter in this setup is the partial correlation between x_t and x_{t-3} with the intervening x's held constant. If, however, the process were only AR(2), the acf in

Eq. (7.30) shows that

$$\rho_3 = \alpha_1\rho_2 + \alpha_2\rho_1$$

Substitution in the third Yule-Walker equation then gives $\alpha_3 = 0$. It can be shown that similar results hold if ever higher-order schemes are incorrectly assumed. Thus the pacf of an AR(2) process will cut off after the second lag. Likewise, the pacf of an AR(3) series will cut off after the third lag. It must be emphasized, however, that these results are for the *population* parameters, which will be imperfectly estimated by sample correlations. Nonetheless, we expect the empirical acf of a stationary AR series to damp toward zero, and the pacf to be approximately zero beyond the order of the process.

7.2.3 MA Processes

The AR(1) process in Eq. (7.19) may be inverted to give

$$x_t = \epsilon_t + \alpha\epsilon_{t-1} + \alpha^2\epsilon_{t-2} + \cdots$$

This is an MA process of infinite order, MA(∞). In a pure MA process a variable is expressed solely in terms of the current and previous white noise disturbances. In practical applications it is only the properties of low-order MA processes that matter. The MA(1) process is

$$x_t = \epsilon_t - \beta_1\epsilon_{t-1}$$

It is simple to show that the autocovariances are

$$\begin{aligned} \gamma_0 &= (1+\beta_1^2)\sigma_\epsilon^2 \\ \gamma_1 &= -\beta_1\sigma_\epsilon^2 \\ \gamma_2 &= \gamma_3 = \cdots = 0 \end{aligned} \tag{7.34}$$

which gives the autocorrelation coefficients as

$$\begin{aligned} \rho_1 &= \frac{-\beta_1}{1+\beta_1^2} \\ \rho_2 &= \rho_3 = \cdots = 0 \end{aligned} \tag{7.35}$$

The MA(1) process may be inverted to give ϵ_t as an infinite series in $x_t, x_{t-1}, \ldots$, namely,

$$\epsilon_t = x_t + \beta_1 x_{t-1} + \beta_1^2 x_{t-2} + \cdots$$

that is,

$$x_t = -\beta_1 x_{t-1} - \beta_1^2 x_{t-2} - \cdots + \epsilon_t \tag{7.36}$$

Because this is an AR(∞) series the partial autocorrelations do not cut off but damp toward zero; but, as we have just seen, the autocorrelations are zero after the first. The properties of a pure MA process are thus the converse of those of a pure AR process. The acf of an MA process cuts off after the order of the process, and the pacf damps toward zero.

Equation (7.36) only makes sense if $|\beta_1| < 1$. If that were not so the implication would be that the most distant x's had the greatest effect on the current x. The condition $|\beta_1| < 1$ is known as the **invertibility condition.** It is similar to the stationarity condition for an AR(1) series, but stationarity of the MA(1) series itself does not impose any condition on β_1. Similar results can be derived for the MA(2) process.[3] In general for a stationary MA(q) process the first q autocorrelation coefficients will be nonzero and the rest zero; the partial autocorrelation coefficients will damp toward zero.

7.2.4 ARMA Processes

The general ARMA(p,q) process is

$$A(L)x_t = B(L)\epsilon_t \tag{7.37}$$

where

$$\begin{aligned} A(L) &= 1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p \\ B(L) &= 1 - \beta_1 L - \beta_2 L^2 - \cdots - \beta_q L^q \end{aligned} \tag{7.38}$$

Stationarity requires the roots of $A(L)$ to lie outside the unit circle, and invertibility places the same condition on the roots of $B(L)$. Given these conditions, the ARMA(p,q) process may alternatively be expressed as a pure AR process of infinite order or as a pure MA process of infinite order, namely,

$$B^{-1}(L)A(L)x_t = \epsilon_t \qquad \text{or} \qquad x_t = A^{-1}(L)B(L)\epsilon_t \tag{7.39}$$

The lowest-order mixed process is the ARMA(1,1),

$$x_t = \alpha x_{t-1} + \epsilon_t - \beta\epsilon_{t-1} \tag{7.40}$$

If we square Eq. (7.40) and take expectations, we find after some manipulation that

$$\sigma_x^2 = \gamma_0 = \frac{1 - 2\alpha\beta + \beta^2}{1 - \alpha^2}\sigma_\epsilon^2 \tag{7.41}$$

Multiplying through Eq. (7.40) by x_{t-1}, then taking expectations, yields

$$\gamma_1 = \alpha\gamma_0 - \beta\sigma_\epsilon^2 = \frac{(\alpha - \beta)(1 - \alpha\beta)}{1 - \alpha^2}\sigma_\epsilon^2 \tag{7.42}$$

Higher-order covariances are given by

$$\gamma_k = \alpha\gamma_{k-1} \qquad k = 2, 3, \ldots \tag{7.43}$$

The autocorrelation function of the ARMA(1,1) process is thus

$$\begin{aligned} \rho_1 &= \frac{(\alpha - \beta)(1 - \alpha\beta)}{1 - 2\alpha\beta + \beta^2} \\ \rho_k &= \alpha\rho_{k-1} \qquad\qquad k = 2, 3, \ldots \end{aligned} \tag{7.44}$$

[3] See Problem 7.2.

TABLE 7.1
Correlation patterns

Process	acf	pacf
AR(p)	Infinite: damps out	Finite: cuts off after lag p
MA(q)	Finite: cuts off after lag q	Infinite: damps out
ARMA	Infinite: damps out	Infinite: damps out

The first coefficient depends on the parameters of both the AR part and the MA part. Subsequent coefficients decline exponentially, with the rate of decline given by the AR parameter. However, by contrast with a pure AR process, the partial autocorrelation coefficients will not cut off but will damp out. This result may be seen intuitively from Eq. (7.39), which shows that all ARMA processes are equivalent to AR processes of *infinite* order.

The correlation patterns of higher-order ARMA processes may be derived in a similar fashion.[4] The expected theoretical patterns for various processes are summarized in Table 7.1. In principle it should be fairly easy to distinguish between pure AR and pure MA processes on the basis of the cutoff in the pacf or acf. Neither cuts off for an ARMA process, so the determination of the order of an ARMA process is often a very difficult and uncertain process. It should also be remembered that these are expected *population* patterns, which may or may not be approximated well by sample estimates.

7.3
TESTING FOR STATIONARITY

Section 7.2 defined the mean, variance, autocovariances, and autocorrelations of a stationary series and derived specific formulae for various low-order ARMA processes. Before calculating sample estimates of these coefficients for any series, one must check whether the series appears to be stationary.[5] Referring to Fig. 2.9, where a stationary AR process and a random walk with drift are shown, we see clearly that computing the mean of the random walk series for various subsets of the data would not be sensible. The results would vary with the particular subset used. This deficiency would be true *a fortiori* for the explosive series shown in Fig. 2.10. If a time series is not stationary it is necessary to look for possible transformations that might induce stationarity. An ARMA model can then be fitted to the transformed series.

There are two principal methods of detecting nonstationarity:

1. Subjective judgment applied to the time series graph of the series and to its correlogram
2. Formal statistical tests for unit roots

[4] See Problem 7.3.

[5] Stationarity was defined in Section 2.5.

7.3.1 Graphical Inspection

Looking again at the three series in Figs. 2.9 and 2.10, we can see that the explosive AR(1) series with a parameter of 1.05 would readily be detected with only a few observations. The other two series with parameters of 0.95 and 1.00 look very similar over some time intervals and somewhat different over others. It is obviously not easy to judge one series to be stationary and the other nonstationary on the basis of visual inspection of the series alone. A more powerful discriminator is the **correlogram.** To illustrate both the uses of the correlogram and the application of unit root tests, we have constructed five artificial series. These are defined as follows:

		Parameter values				
Label	**Series definition**	α	m	δ_0	δ_1	**Type**
Y1	$(1-\alpha L)y_t = m + \epsilon_t$	0.95	1			Stationary
Y2	"	1.00	1			Nonstationary
Y3	"	1.05	1			Explosive
Y4	$(1-\alpha L)(y_t - \delta_0 - \delta_1 t) = \epsilon_t$	0.9		10	0.5	Nonstationary
Y5	"	1.0		10	0.5	Nonstationary

The series were generated with the indicated parameter values and, for all but Y5, a common set of 200 normally distributed random numbers. Y1 is a stationary AR(1) series, Y2 a random walk with drift, and Y3 an explosive AR(1) series. Y4 is the sum of a linear time trend and a stationary AR(1) series. Y5 is another example of a random walk with drift, with a different set of random numbers used in its generation. In each series the first 100 observations were discarded and the last 100 used for calculating acf and pacf coefficients. The *sample* autocorrelations are calculated from the formula

$$r_k = \frac{\sum_{t=k+1}^{n}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$$

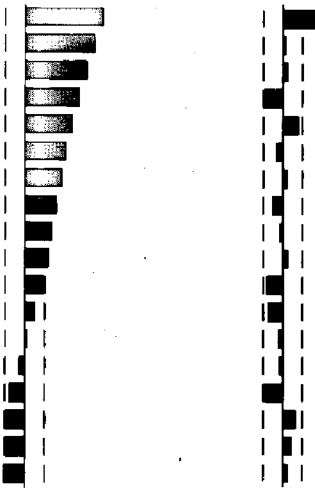
where $\bar{x} = \sum_{t=1}^{n} x_t/n$. For a white noise series, these coefficients have an approximate variance of $1/n$. The *sample* partial autocorrelations are the *last* coefficients in the sequence of ever higher-order AR schemes fitted to the series.

The correlogram for Y1 in Table 7.2 displays the classical pattern for a stationary AR(1) series, with the autocorrelations dying out and only the first partial correlation coefficient being significant. There is, however, one noticeable difference from the theoretical correlogram for an AR(1) series with positive parameter, as shown in Fig. 7.1, and the sample correlogram for Y1 in Table 7.2. All the coefficients in the former are positive, whereas the latter shows some negative, albeit insignificant, autocorrelations. The phenomenon is explained by Kendall and Ord.[6] They cite an article by Anderson that shows that for *any* series the sum of all possible

[6] Sir Maurice Kendall and J. Keith Ord, *Time Series*, 3rd edition, Edward Arnold, 1990, 82–83.

TABLE 7.2
Correlogram of Y1

Sample: 101–200
Included observations: 100

Autocorrelation	Partial correlation		AC	PAC	Q-stat	Prob.
		1	0.882	0.882	80.153	0.000
		2	0.781	0.015	143.67	0.000
		3	0.706	0.062	196.04	0.000
		4	0.593	−0.196	233.45	0.000
		5	0.535	0.177	264.21	0.000
		6	0.474	−0.070	288.58	0.000
		7	0.415	0.047	307.49	0.000
		8	0.361	−0.100	321.93	0.000
		9	0.296	−0.019	331.73	0.000
		10	0.256	0.043	339.18	0.000
		11	0.185	−0.166	343.10	0.000
		12	0.092	−0.145	344.08	0.000
		13	0.021	−0.048	344.13	0.000
		14	−0.055	−0.022	344.49	0.000
		15	−0.146	−0.186	347.05	0.000
		16	−0.185	0.130	351.21	0.000
		17	−0.194	0.089	355.81	0.000
		18	−0.204	0.043	360.97	0.000

The vertical dashed lines represent two standard errors around zero.
AC = Autocorrelation coefficient; PAC = Partial correlation coefficient; Q-stat = Box-Pierce-Ljung statistic [Eq. (6.58)]; Prob. = *P*-value for hypothesis that all autocorrelation coefficients to this point are zero.

autocorrelations is equal to -0.5.[7] The average *r is* thus slightly negative, even for an AR(1) process with positive parameter.

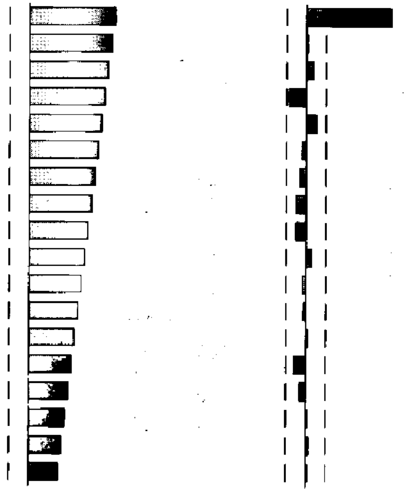
The *sample* autocorrelations and partial correlation coefficients may be computed for a nonstationary series even though their *population* counterparts do not exist. The autocorrelations for the random walk in Table 7.3 decline but do not die out quickly. The pattern, however, is not very dissimilar from that for Y1 because the stationary series has a parameter very close to unity. The autocorrelations for the explosive series in Table 7.4 are almost identical with those of the random walk: the partial correlations, however, are different in that all but the first are essentially zero in the explosive case.

First differences of Y are denoted by DY. The DY2 series is white noise, as shown by Table 7.5. The first difference of the explosive series has a correlogram in Table 7.6 similar to that of the explosive series itself. Taking the first difference of the first difference would still not yield a correlogram that dies out. There is thus a distinction between a series like Y2 that is labeled nonstationary and one like Y3 that is labeled explosive. The distinction is that the nonstationary series can be transformed

[7] O. D. Anderson, "Serial Dependence Properties of Linear Processes," *Journal of the Operational Research Society*, 1980, **31**, 905–917.

TABLE 7.3
Correlogram of Y2

Sample: 101–200
Included observations: 100

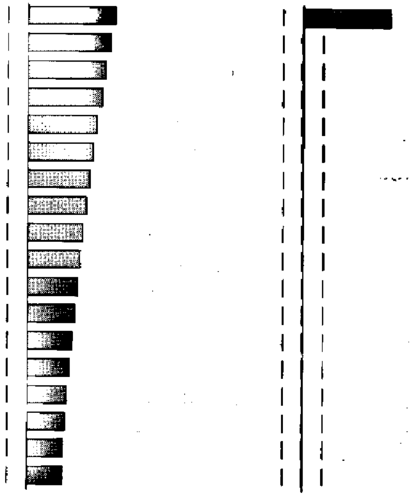


Autocorrelation	Partial correlation		AC	PAC	Q-stat	Prob.
		1	0.958	0.958	94.492	0.000
		2	0.918	0.011	182.22	0.000
		3	0.886	0.068	264.71	0.000
		4	0.841	−0.163	339.84	0.000
		5	0.806	0.094	409.52	0.000
		6	0.770	−0.042	473.90	0.000
		7	0.729	−0.058	532.14	0.000
		8	0.685	−0.084	584.19	0.000
		9	0.637	−0.078	629.70	0.000
		10	0.596	0.057	669.90	0.000
		11	0.554	−0.035	705.07	0.000
		12	0.511	−0.024	735.33	0.000
		13	0.474	0.011	761.61	0.000
		14	0.428	−0.104	783.34	0.000
		15	0.379	−0.056	800.61	0.000
		16	0.338	0.014	814.45	0.000
		17	0.299	0.034	825.41	0.000
		18	0.264	0.018	834.06	0.000

The vertical dashed lines represent two standard errors around zero.
AC = Autocorrelation coefficient; PAC = Partial correlation coefficient; Q-stat = Box-Pierce-Ljung statistic [Eq. (6.58)]; Prob. = P-value for hypothesis that all autocorrelation coefficients to this point are zero.

TABLE 7.4
Correlogram of Y3

Sample: 101–200
Included observations: 100



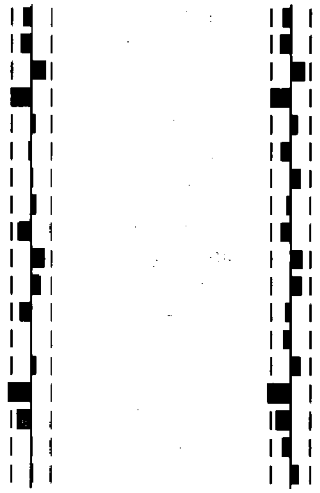
Autocorrelation	Partial correlation		AC	PAC	Q-stat	Prob.
		1	0.946	0.946	92.224	0.000
		2	0.894	−0.006	175.50	0.000
		3	0.845	−0.006	250.58	0.000
		4	0.798	−0.006	318.18	0.000
		5	0.752	−0.006	378.92	0.000
		6	0.709	−0.007	433.41	0.000
		7	0.667	−0.007	482.17	0.000
		8	0.627	−0.007	525.72	0.000
		9	0.588	−0.007	564.51	0.000
		10	0.551	−0.007	598.96	0.000
		11	0.516	−0.007	629.46	0.000
		12	0.482	−0.008	656.38	0.000
		13	0.449	−0.008	680.03	0.000
		14	0.418	−0.008	700.73	0.000
		15	0.388	−0.008	718.75	0.000
		16	0.358	−0.008	734.36	0.000
		17	0.331	−0.008	747.79	0.000
		18	0.304	−0.008	759.26	0.000

The vertical dashed lines represent two standard errors around zero.
AC = Autocorrelation coefficient; PAC = Partial correlation coefficient; Q-stat = Box-Pierce-Ljung statistic [Eq. (6.58)]; Prob. = P-value for hypothesis that all autocorrelation coefficients to this point are zero.

TABLE 7.5

Correlogram of DY2

Sample: 101–200
Included observations: 100



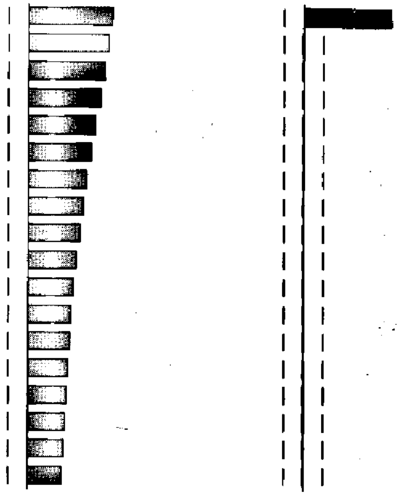
Autocorrelation	Partial correlation		AC	PAC	Q-stat	Prob.
		1	−0.060	−0.060	0.3662	0.545
		2	−0.094	−0.098	1.2928	0.524
		3	0.157	0.147	3.8765	0.275
		4	−0.209	−0.208	8.5106	0.075
		5	0.039	0.057	8.6768	0.123
		6	−0.016	−0.084	8.7040	0.191
		7	0.010	0.089	8.7140	0.274
		8	0.043	−0.030	8.9219	0.349
		9	−0.125	−0.085	10.687	0.298
		10	0.138	0.113	12.850	0.232
		11	0.092	0.096	13.827	0.243
		12	−0.101	−0.043	15.019	0.240
		13	0.002	−0.069	15.019	0.306
		14	0.055	0.084	15.372	0.353
		15	−0.230	−0.226	21.725	0.115
		16	−0.132	−0.146	23.848	0.093
		17	0.020	−0.074	23.897	0.122
		18	0.017	0.073	23.931	0.157

The vertical dashed lines represent two standard errors around zero.
AC = Autocorrelation coefficient; PAC = Partial correlation coefficient; Q-stat = Box-Pierce-Ljung statistic [Eq. (6.58)]; Prob. = P-value for hypothesis that all autocorrelation coefficients to this point are zero.

TABLE 7.6

Correlogram of DY3

Sample: 101–200
Included observations: 100



Autocorrelation	Partial correlation		AC	PAC	Q-stat	Prob.
		1	0.946	0.946	92.223	0.000
		2	0.894	−0.006	175.49	0.000
		3	0.845	−0.005	250.58	0.000
		4	0.798	−0.008	318.17	0.000
		5	0.752	−0.007	378.90	0.000
		6	0.709	−0.006	433.38	0.000
		7	0.667	−0.008	482.14	0.000
		8	0.627	−0.006	525.68	0.000
		9	0.588	−0.007	564.47	0.000
		10	0.551	−0.007	598.92	0.000
		11	0.516	−0.007	629.42	0.000
		12	0.482	−0.007	656.33	0.000
		13	0.449	−0.007	679.99	0.000
		14	0.418	−0.008	700.70	0.000
		15	0.388	−0.008	718.74	0.000
		16	0.359	−0.008	734.36	0.000
		17	0.331	−0.009	747.80	0.000
		18	0.304	−0.008	759.29	0.000

The vertical dashed lines represent two standard errors around zero.
AC = Autocorrelation coefficient; PAC = Partial correlation coefficient; Q-stat = Box-Pierce-Ljung statistic [Eq. (6.58)]; Prob. = P-value for hypothesis that all autocorrelation coefficients to this point are zero.

by first differencing once or a very few times to give a stationary series, which can then be modeled as an ARMA process, whereas this is not true of the explosive series.

7.3.2 Integrated Series

In the older time series literature the type of series we have labeled nonstationary was called **homogeneous nonstationary**.[8] In the more recent literature the series is said to **be integrated.** The **order of integration** is the minimum number of times the series needs to be first differenced to yield a stationary series. Thus Y2 is **integrated of order one,** denoted I(1). A stationary series is then said to be **integrated of order zero,** I(0). Notice that first differencing an I(0) series still yields an I(0) series. For example, a white noise series is the simplest example of an I(0) series; its first difference is a stationary MA(1) series.

7.3.3 Trend Stationary (TS) and Difference Stationary (DS) Series

Y4 is a simple example of a trend stationary series. It may be written

$$\begin{aligned} y_t &= \delta_0 + \delta_1 t + u_t \qquad u_t = \alpha u_{t-1} + \epsilon_t \\ \text{or} \qquad y_t &= [\delta_0(1-\alpha) + \alpha\delta_1] + \delta_1(1-\alpha)t + \alpha y_{t-1} + \epsilon_t \end{aligned} \tag{7.45}$$

where ϵ_t. in our numerical example, is a (Gaussian) white noise series and $|\alpha| < 1$. If y is the log of a variable, Eq. (7.45) asserts that the variable is subject to a constant growth trend and that the deviations from the trend follow a stationary AR(1) process. The graph of Y4 is shown in Fig. 7.3. The underlying linear trend, $10 + 0.5t$, is denoted by Y4HAT: and the actual values are seen to fluctuate around the trend with no obvious tendency for the amplitude of the fluctuations to increase or decrease. For this reason the series is said to be **trend stationary** (TS). The steady increase in the mean level renders the series nonstationary. Its first difference, however, is **stationary. From Eq. (7.45)**

$$\Delta y_t = \delta_1 + \Delta u_t \tag{7.46}$$

where Δu_t is stationary since u_t is stationary.[9] By assumption u_t is ARMA(1,0), and so $\Delta u_t = (1-\alpha L)^{-1}(1-L)\epsilon_t$ is ARMA(1,1); but it is not invertible since the MA polynomial has a unit root.

If the α parameter in Eq. (7.45) is one, so that the autoregressive part of the relation has a unit root, we have the Y5 series, also shown in Fig. 7.3. It fluctuates much like Y4 but strays away from the linear trend. The first difference takes

[8] See G. E. P. Box and G. M. Jenkins, *Time Series Analysis, Forecasting and Control*, revised ed., Holden Day, 1976.

[9] See Problem 7.4.

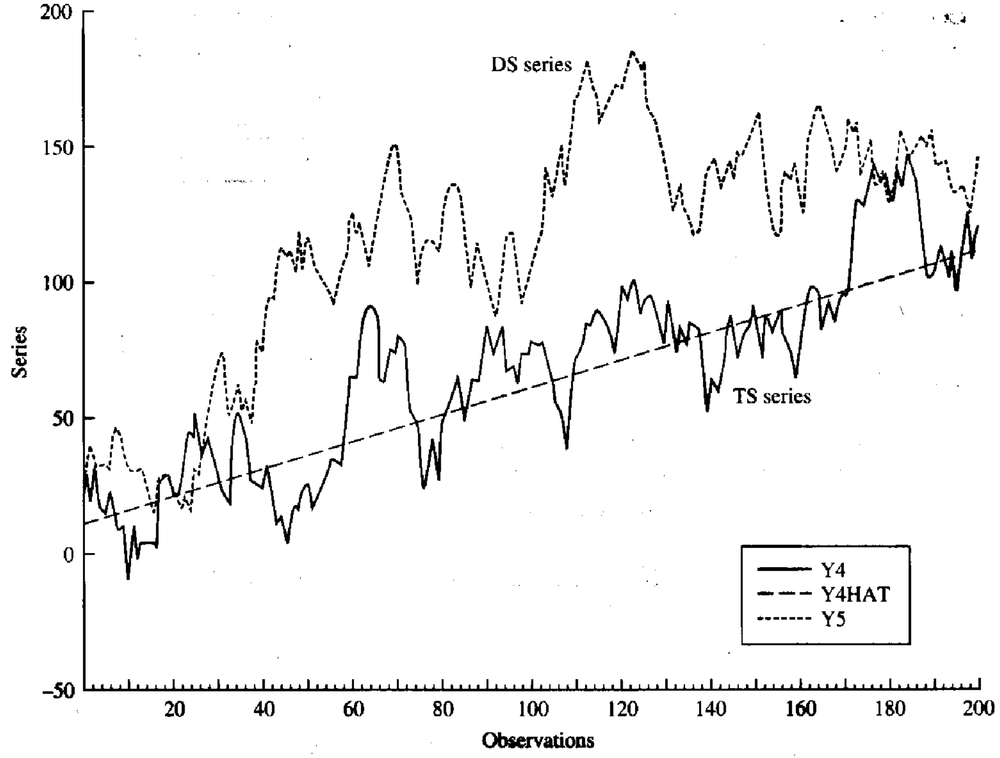


FIGURE 7.3
Trend stationary (TS) and difference stationary (DS) series.

the form

$$\Delta y_t = \delta_1 + \epsilon_t \tag{7.47}$$

Thus Δy is stationary, being a constant plus a white noise series. The y variable is now said to be **difference stationary** (DS).

At first sight Eqs. (7.46) and (7.47) seem to be practically identical, and one might wonder wherein lies the distinction between TS and DS series. There is, in fact, a crucial distinction. If we label ϵ_t the **innovations** or **shocks** to the system, the innovations have a transient, diminishing effect on y in the TS case, and a permanent effect in the DS situation. In Eq. (7.45) u_t measures the deviation of the series from trend in period t. We wish to examine the effect of an innovation ϵ_t on the current and subsequent deviations from trend. By definition

$$u_{t+s} = u_{t-1} + \Delta u_t + \Delta u_{t+1} + \cdots + \Delta u_{t+s}$$

From the assumption in Eq. (7.45),

$$\begin{aligned}\Delta u_t &= \epsilon_t + (\alpha - 1)u_{t-1} \\ &= \epsilon_t + (\alpha - 1)(\epsilon_{t-1} + \alpha\epsilon_{t-2} + \alpha^2\epsilon_{t-3} + \cdots)\end{aligned} \tag{7.48}$$

Let ϵ_t be some given value and set all subsequent innovations to zero. Then, by ignoring the term in u_{t-1} since it is some given constant for the purposes of this analysis, Eq. (7.48) gives

$$\begin{aligned}\Delta u_t &= \epsilon_t \\ \Delta u_{t+1} &= (\alpha - 1)\epsilon_t \\ \Delta u_{t+2} &= \alpha(\alpha - 1)\epsilon_t \\ \Delta u_{t+s} &= \alpha^{s-1}(\alpha - 1)\epsilon_t\end{aligned}$$

Summing these first differences, we find

$$\sum_{j=0}^{s} \Delta u_{t+j} = \epsilon_t\left[1 + (\alpha - 1)\frac{1-\alpha^s}{1-\alpha}\right] = \alpha^s \epsilon_t$$

Thus,

$$u_{t+s} = u_{t-1} + \alpha^s \epsilon_t \tag{7.49}$$

that is, the effect of the ϵ_t innovation on subsequent deviations from trend diminishes toward zero the farther ahead one looks. In the unit root case, $\alpha = 1$ and $\Delta u_t = \epsilon_t$. With all subsequent innovations set at zero as before,

$$u_{t+s} = u_{t-1} + \epsilon_t \tag{7.50}$$

that is, the ϵ_t innovation has a permanent effect on all subsequent deviations from trend.

An alternative and simpler derivation of these results can be obtained by expressing u_t in terms of the innovations. From Eq. (7.45),

$$u_t = \epsilon_t + \alpha\epsilon_{t-1} + \alpha^2\epsilon_{t-2} + \cdots$$

which gives

$$\frac{\partial u_{t+s}}{\partial \epsilon_t} = \alpha^s$$

This result yields Eq. (7.49) or Eq. (7.50), according as $|\alpha| < 1$ or $\alpha = 1$.

The contrast between TS and DS series has been developed in terms of a very simple specification. The same basic contrasts may be developed for more complicated models.[10] In general the model may be written

$$y_t - \delta_0 - \delta_1 t = u_t \qquad A(L)u_t = B(L)\epsilon_t \tag{7.51}$$

where $A(L)$ and $B(L)$ are polynomials of order p and q in the lag operator. When all roots of $A(L)$ lie outside the unit circle, the deviations from trend follow a stationary ARMA(p, q) scheme. If, however, $A(L)$ contains a unit root, the result is a DS model. In this case

$$A(L) = (1 - L)(1 - \lambda_2 L)\cdots(1 - \lambda_p L) = (1 - L)A^*(L)$$

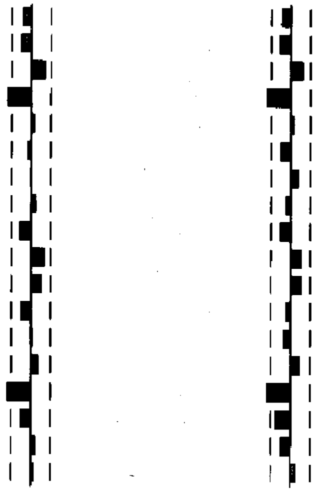
where all $(p - 1)$ roots of $A^*(L)$ lie outside the unit circle. Then Eq. (7.51) becomes

$$A^*(L)(\Delta y_t - \delta_1) = B(L)\epsilon_t, \tag{7.52}$$

[10] See Problem 7.5.

TABLE 7.7
Correlogram of DY4

Sample: 101–200
Included observations: 100

Autocorrelation	Partial correlation		AC	PAC	Q-stat	Prob.
		1	−0.091	−0.091	0.8444	0.358
		2	−0.122	−0.131	2.3834	0.304
		3	0.141	0.120	4.4681	0.215
		4	−0.233	−0.233	10.215	0.037
		5	0.027	0.027	10.290	0.067
		6	−0.030	−0.114	10.386	0.109
		7	−0.001	0.063	10.386	0.168
		8	0.034	−0.052	10.516	0.231
		9	−0.138	−0.110	12.657	0.179
		10	0.137	0.093	14.784	0.140
		11	0.093	0.092	15.768	0.150
		12	−0.104	−0.041	17.026	0.149
		13	0.004	−0.069	17.028	0.198
		14	0.061	0.089	17.470	0.232
		15	−0.230	−0.224	23.826	0.068
		16	−0.129	−0.160	25.831	0.056
		17	0.029	−0.095	25.933	0.076
		18	0.024	0.054	26.006	0.100

The vertical dashed lines represent two standard errors around zero.
AC = Autocorrelation coefficient; PAC = Partial correlation coefficient; Q-stat = Box-Pierce-Ljung statistic [Eq. (6.58)]; Prob. = *P*-value for hypothesis that all autocorrelation coefficients to this point are zero.

so that the first difference of the series can be modeled as a stationary ARMA($p-1, q$) process.

Tables 7.7 and 7.8 give the correlograms for DY4 and DY5. Each lends strong support to the expectation of stationarity. At first sight it may seem surprising that even the low-order autocorrelations of DY4 are insignificant. However, as noted before, this is an ARMA(1,1) series; and it was shown in Eq. (7.44) that

$$\rho_1 = \frac{(\alpha-\beta)(1-\alpha\beta)}{1-2\alpha\beta+\beta^2}$$

When α and β are numerically close, as in the generation of DY4 where $\alpha = 0.9$ and $\beta = 1$, this first autocorrelation will be small and the subsequent autocorrelations still smaller. These correlograms would not enable one to distinguish between a TS and a DS series, which motivates the search for formal statistical tests for unit roots. However, the available tests have low power and so the distinction between the two types of series, although of theoretical importance, may be of little practical significance.

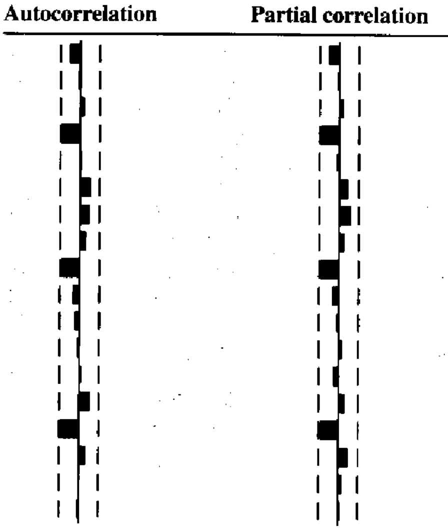
7.3.4 Unit Root Tests

Return to Eq. (7.45), namely,

$$y_t = \delta_0 + \delta_1 t + u_t \qquad u_t = \alpha u_{t-1} + \epsilon_t$$

TABLE 7.8
Correlogram of DY5

Sample: 101–200
Included observations: 100



	AC	PAC	Q-stat	Prob.
1	−0.085	−0.085	0.7405	0.390
2	−0.017	−0.024	0.7702	0.680
3	0.043	0.040	0.9638	0.810
4	−0.166	−0.161	3.8991	0.420
5	0.003	−0.023	3.8999	0.564
6	0.081	0.074	4.6165	0.594
7	0.071	0.099	5.1740	0.639
8	0.055	0.048	5.5110	0.702
9	−0.154	−0.159	8.1562	0.518
10	−0.052	−0.064	8.4664	0.583
11	−0.034	−0.021	8.5977	0.659
12	−0.005	0.017	8.6003	0.737
13	0.021	−0.038	8.6502	0.799
14	0.089	0.056	9.5998	0.791
15	−0.166	−0.159	12.890	0.611
16	0.068	0.085	13.456	0.639
17	−0.000	0.033	13.456	0.705
18	−0.017	0.013	13.490	0.762

The vertical dashed lines represent two standard errors around zero.
AC = Autocorrelation coefficient; PAC = Partial correlation coefficient; Q-stat = Box-Pierce-Ljung statistic [Eq. (6.58)]; **Prob.** = *P*-value for hypothesis that all autocorrelation coefficients to this point are zero.

We require a test of the null hypothesis $H_0\colon \alpha = 1$. As already seen, combining the two equations gives

$$y_t = [\delta_0(1-\alpha) + \alpha\delta_1] + \delta_1(1-\alpha)t + \alpha y_{t-1} + \epsilon_t \tag{7.53}$$

Subtracting y_{t-1} from each side gives a more convenient expression

$$\Delta y_t = [\delta_0(1-\alpha) + \alpha\delta_1] + \delta_1(1-\alpha)t + \gamma y_{t-1} + \epsilon_t \tag{7.54}$$

where $\gamma = \alpha - 1$. The null hypothesis is now $H_0\colon \gamma = 0$. Thus γ will be zero if there is a unit root, and negative under stationary deviations from the trend. This result suggests running an OLS regression on Eq. (7.54) and rejecting the null hypothesis if a significant negative value is found for $\hat{\gamma}$. Recall, however, that the significance test requires the distribution of the test statistic *under the null.* When the null is true, Eq. (7.54) reduces to

$$\Delta y_t = \delta_1 + \epsilon_t \tag{7.55}$$

so that y_t is a random walk with drift and thus nonstationary. The ratio $\hat{\gamma}/\,\text{s.e.}(\hat{\gamma})$ does not then follow the standard t distribution, nor is it asymptotically N(0, 1), because stationarity was required in the derivation of the standard distributions.

The inference problem was solved by Fuller, who obtained limiting distributions for this ratio in several important cases.[11] These distributions were approx-

[11] Wayne A. Fuller, *Introduction to Statistical Time Series*, Wiley, 1976, 366–382.

TABLE 7.9
Asymptotic critical values for unit root tests

Test statistic	1%	2.5%	5%	10%
τ_{nc}	−2.56	−2.23	−1.94	−1.62
τ_c	−3.43	−3.12	−2.86	−2.57
τ_{ct}	−3.96	−3.66	−3.41	−3.13

Reprinted by permission from Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics*, Oxford University Press, 1993, 708.

imated empirically by Dickey.[12] The tests are thus known as Dickey-Fuller (DF) tests. More recently, MacKinnon has derived critical values from a much larger set of replications.[13] MacKinnon has also fitted response surface regressions to these replications, which permit the calculation of Dickey-Fuller critical values for any sample size and for various specifications of regressions like Eq. (7.54). The MacKinnon procedures are now incorporated in EViews software from Quantitative Micro Software. Asymptotic critical values are given in Table 7.9. Definitions of the three test statistics in this table are given shortly.

The unit root test based on Eq. (7.54) attempts to discriminate between series like Y4 and Y5 in our numerical example. It is also often important to discriminate between series like Y1 and Y2, where there is no linear trend. The relevant procedure can be derived by setting δ_1 to zero in Eq. (7.54), giving

$$\Delta y_t = \delta_0(1-\alpha) + \gamma y_{t-1} + \epsilon_t \tag{7.56}$$

Under the null hypothesis this reduces to

$$\Delta y_t = \epsilon_t \tag{7.57}$$

so that y_t is a random walk without drift and nonstationary. The unit root test procedure is to fit Eq. (7.56) by OLS and refer $\hat{\gamma}/\text{s.e.}(\hat{\gamma})$ to the relevant MacKinnon critical value or, more simply, to the appropriate asymptotic value in Table 7.9.

Finally, for processes with zero mean the relevant test regression is

$$\Delta y_t = \gamma y_{t-1} + \epsilon_t \tag{7.58}$$

Under the null this also reduces to Eq. (7.57). There are thus three possible test regressions. Each has Δy as the regressand. In Eq. (7.58) the only regressor is lagged y, in Eq. (7.56) a constant is included in the regressors, and in Eq. (7.54) there is a constant and a time trend in addition to lagged y. Following the notation in Davidson and MacKinnon we denote the three possible test statistics, $\hat{\gamma}/\text{s.e.}(\hat{\gamma})$, by τ_{nc}, τ_c, or τ_{ct}, according to whether they come from Eq. (7.58), Eq. (7.56), or Eq. (7.54).[14] The relevant rows of Table 7.9 are indicated by these symbols.

[12] D. A. Dickey, *Hypothesis Testing for Nonstationary Time Series*, Unpublished manuscript, Iowa State University, Ames, IA, 1975.

[13] James G. MacKinnon, "Critical Values for Cointegration Tests," Chapter 13, *Long-Run Economic Relationships*, eds. R. Engle and C. W. J. Granger, Oxford University Press, 1991.

[14] Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics*, Oxford University Press, 1993, p. 703.

The preceding explanation of the Dickey-Fuller tests has incorporated only an AR(1) process. If this is inadequate ϵ_t will almost certainly be serially correlated, which invalidates the derivation of the DF tests. To investigate the impact of higher-order processes we will look first at a second-order process. Specify

$$y_t = \delta + u_t \qquad u_t = \alpha_1 u_{t-1} + \alpha_2 u_{t-2} + \epsilon_t \tag{7.59}$$

We have not included a trend term so as to keep the exposition simple. If appropriate, a trend term can be added later. The AR polynomial may be written

$$A(L) = 1 - \alpha_1 L - \alpha_2 L^2 = (1 - \lambda_1 L)(1 - \lambda_2 L)$$

If one root, say λ_1, is unity, it follows that

$$A(1) = 1 - \alpha_1 - \alpha_2 = 0 \tag{7.60}$$

where $A(1)$, the result of substituting 1 for L in $A(L)$, gives the sum of the coefficients in $A(L)$. To carry out a unit root test we would examine whether $\hat{\gamma} = 1 - \hat{\alpha}_1 - \hat{\alpha}_2$ differs significantly from zero. The test would be simplified if Eq. (7.59) could be rearranged to have γ as the coefficient of a single variable. Combining the two equations in Eq. (7.59) gives

$$\begin{aligned} y_t &= \delta(1 - \alpha_1 - \alpha_2) + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \epsilon_t \\ &= \delta(1 - \alpha_1 - \alpha_2) + (\alpha_1 + \alpha_2) y_{t-1} - \alpha_2 \Delta y_{t-1} + \epsilon_t \end{aligned}$$

which gives $$\Delta y_t = \delta(1 - \alpha_1 - \alpha_2) - \gamma y_{t-1} - \alpha_2 \Delta y_{t-1} + \epsilon_t \tag{7.61}$$

where $\gamma = 1 - \alpha_1 - \alpha_2 = A(1)$. Comparing Eq. (7.61) with Eq. (7.56), we see that the effect of moving from an AR(1) specification to an AR(2) specification is the addition of a lagged first difference in y to the test regression. This procedure extends in a straightforward fashion to higher-order AR processes. For an AR(p) specification the test regression is

$$\Delta y_t = f(\text{constant, trend}, y_{t-1}, \Delta y_{t-1}, \ldots, \Delta y_{t-p+1}) \tag{7.62}$$

The inclusion of a constant or a constant plus trend is guided by the same considerations as in the AR(1) case. The coefficient of y_{t-1} provides the test of the unit root hypothesis as before. The same critical values can be used as in the AR(1) case. The coefficients of the lagged first differences in y are usually of no specific interest, but hypotheses about them may be tested by conventional t and F statistics. The objective is to include sufficient lagged terms in Eq. (7.62) to yield a white noise residual. Tests based on Eq. (7.62) are known as *augmented Dickey-Fuller* (ADF) tests.

7.3.5 Numerical Example

The preceding Y1 series is a stationary AR(1) scheme with parameter 0.95. Applying a DF test by fitting regression Eq. (7.56) gives the results shown in Table 7.10.[15] The

[15] The EViews output uses the designation ADF for all unit root test statistics, irrespective of whether any lagged first differences are included in the test regression.

TABLE 7.10
Unit root test on Y1

ADF test statistic	−2.450004	1%	Critical value*	−3.4965
		5%	Critical value	−2.8903
		10%	Critical value	−2.5819

*MacKinnon critical values for rejection of hypothesis of a unit root.

Augmented Dickey-Fuller test equation
LS // dependent variable is D(Y1)
Sample: 101–200
Included observations: 100

Variable	Coefficient	Std. error	T-statistic	Prob.
Y1(−1)	−0.115440	0.047118	−2.450004	0.0161
C	3.963090	1.869770	2.119560	0.0366

unit root hypothesis is not rejected even at the 10 percent level. This result is hardly surprising, given that the true AR parameter is close to 1. If the sample is extended to include all 200 data points the DF statistic is −3.42, with a 1 percent critical value of −3.46, which now rejects the unit root hypothesis at about the 1 percent level. If two lagged first differences are added to the test regression with 100 observations, both coefficients are insignificant with t ratios of −0.20 and −0.52. The corresponding ADF statistic is −2.14, which still fails to reject the unit root hypothesis.

Table 7.11 shows the ADF test on Y4, the trend stationary series. The ADF statistic is −2.94, which fails to reject the null hypothesis at the 10 percent level. Again this result is not unexpected since the AR coefficient is 0.9. Low power in statistical tests is an often unavoidable fact of life, with which one must live and not expect to be able to make definitive pronouncements. Failure to reject a null hypothesis justifies at best only a cautious and provisional acceptance. An interesting

TABLE 7.11
Unit root test on Y4

ADF test statistic	−2.941227	1%	Critical value*	−4.0521
		5%	Critical value	−3.4548
		10%	Critical value	−3.2283

*MacKinnon critical values for rejection of hypothesis of a unit root.

Augmented Dickey-Fuller test equation
LS // dependent variable is D(Y4)
Sample: 101–200
Included observations: 100

Variable	Coefficient	Std. error	T-statistic	Prob.
Y4(−1)	−0.162216	0.055152	−2.941227	0.0041
C	0.584690	4.997608	0.116994	0.9071
Trend	0.098689	0.045427	2.172477	0.0323

study by Rudebusch shows that U.S. data on real GNP, which fails to reject the unit root hypothesis, also fails to reject a stationarity hypothesis when the latter is set up as the null.[16]

7.4 IDENTIFICATION, ESTIMATION, AND TESTING OF ARIMA MODELS

7.4.1 Identification

The procedures of Section 7.3 enable one to determine whether a series is stationary or whether it needs to be differenced once, or possibly twice, to yield a stationary series. The procedures of Section 7.2 then provide a tentative decision on the orders of the ARMA process to be fitted to the stationary series. The end result is the **identification** of an **autoregressive, integrated, moving average, ARIMA(p, d, q)** model. The three parameters are these:

$$\begin{aligned} d &= \text{number of differences required for stationarity} \\ p &= \text{order of the AR component} \\ q &= \text{order of the MA component} \end{aligned}$$

Typically d is zero or one, or very occasionally two; and one seeks a **parsimonious** representation with low values of p and q. The difficult choice of the order of p and q may be helped by a numerical procedure suggested by Hannan and Rissanen.[17] The procedure has three steps. In the first step some pure AR processes of fairly high order are estimated by OLS, which is not unreasonable since an unknown ARMA process is equivalent to an infinite AR process. The regression with the smallest value of the Akaike information criterion (AIC) is selected in step two, and the residuals $\{e_t\}$ from this regression are taken as estimates of the unknown ϵ's in an ARMA model. In the final step a number of ARMA models are fitted using these estimated residuals. For instance, if an ARMA (2,1) is fitted, the regression is

$$y_t = m + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + e_t - \beta_1 e_{t-1} + \text{error}$$

Such regressions are fitted by OLS for various values of p and q. The residual variance $\hat{\sigma}^2_{p,q}$ is obtained and the specification chosen that has the lowest value of

$$\ln \hat{\sigma}^2_{p,q} + (p + q) \ln n/n$$

which is the Schwarz criterion. It is important to emphasize that, even though the Hannan-Rissanen procedure yields numerical estimates of an ARMA model, they

[16] Glenn D. Rudebusch, "The Uncertain Unit Root in Real GNP," *American Economic Review*, 1993, **83**, 264–272.

[17] E. J. Hannan and J. Rissanen, "Recursive Estimation of Mixed Autoregressive-Moving Average Order," *Biometrika*, **69**, 1982, 81–94; correction, **70**, 1983, 303.

are the by-product of an identification process and are not meant as final estimates of the relevant coefficients.[18]

7.4.2 Estimation

Software packages typically offer least-squares, whether linear or nonlinear, or maximum likelihood, whether conditional or full, estimation procedures for ARMA models. We will comment only on a few specific cases. Returning to the AR(1) specification in Eq. (7.16), that is,

$$y_t = m + \alpha y_{t-1} + \epsilon_t$$

where ϵ is white noise, OLS is an obvious estimator. The only qualification is that the value of y_1 is taken as given and summations run over $t = 2, 3, \ldots, n$. As seen in Chapter 2, the usual test statistics now only have an asymptotic justification, because of the lagged regressor. OLS may also be seen to be a conditional ML estimator. If we take y_1 as given, the conditional likelihood for the remaining $n - 1$ observations is

$$\begin{aligned} L^* &= p(y_2, y_3, \ldots, y_n \mid y_1) \\ &= p(y_2 \mid y_1)p(y_3 \mid y_2)\cdots p(y_n \mid y_{n-1}) \end{aligned}$$

If we assume Gaussian white noise,

$$p(y_t \mid y_{t-1}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_t - m - \alpha y_{t-1})^2\right]$$

Thus the conditional log-likelihood is

$$l^* = \ln L^* = \text{constant} - \frac{n-1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{n}(y_t - m - \alpha y_{t-1})^2 \quad (7.63)$$

Maximizing with respect to m and α gives the OLS estimates just described.

To obtain *full* ML estimates one must maximize the unconditional likelihood

$$L = p(y_1)L^*$$

Under the assumptions of the AR(1) process

$$y_1 \sim N\left(\frac{m}{1-\alpha}, \frac{\sigma^2}{1-\alpha^2}\right)$$

Thus, $\ln p(y_1) = \text{constant} + \frac{1}{2}\ln(1 - \alpha^2) - \frac{1}{2}\ln\sigma^2 - \frac{1-\alpha^2}{2\sigma^2}\left(y_1 - \frac{m}{1-\alpha}\right)^2$

[18] For a detailed account of the Hannan-Rissanen procedure and some illuminating examples of the identification process see C. W. J. Granger and P. Newbold, *Forecasting Economic Time Series*, 2nd edition, Academic Press, 1986, Chapter 3.

and so the unconditional log-likelihood is

$$\begin{aligned} l &= \ln p(y_1) + l^* \\ &= \text{constant} - \frac{n}{2}\ln\sigma^2 + \tfrac{1}{2}\ln(1-\alpha^2) - \frac{1-\alpha^2}{2\sigma^2}\left(y_1 - \frac{m}{1-\alpha}\right)^2 \\ &\quad - \frac{1}{2\sigma^2}\sum_{t=2}^{n}(y_t - m - \alpha y_{t-1})^2 \end{aligned} \tag{7.64}$$

Taking first derivatives with respect to m and α no longer yields linear equations in these parameters. Thus iterative techniques are required to maximize Eq. (7.64). In small samples the difference between maximizing Eq. (7.63) or Eq. (7.64) may be important, but this difference diminishes with sample size.

Higher-order AR schemes are fitted in a similar fashion. Least squares and conditional maximum likelihood take the first p observations as given. Full maximum likelihood procedures are also available.[19]

The fitting of MA schemes is more complicated. Even the lowest-order processes involve nonlinear methods. For example, the MA(1) scheme is

$$y_t = \mu + \epsilon_t - \beta\epsilon_{t-1}$$

If ϵ_0 is set at zero, then $\epsilon_1 = y_1 - \mu$ and $\epsilon_2 = y_2 - \mu + \beta\epsilon_1$. By proceeding in this fashion, all n values of ϵ can be expressed in terms of μ and β. However, $\sum_{t=1}^{n}\epsilon_t^2$ is a complicated nonlinear function of the parameters, and iterative techniques are required to obtain even conditional estimates. Full ML procedures are again available and are fully described in Hamilton's outstanding treatise. Full ARMA schemes share all the estimation problems of the pure MA processes.

7.4.3 Diagnostic Testing

Identification and estimation have produced an estimated ARIMA model. The next stage in univariate time series modeling is the testing of the resultant equation.

1. One set of tests can be applied to the estimated coefficients of the model, in the manner of Chapter 3. Thus one may test the significance of an included variable or a subset of such variables. One may also test the effect of adding one or more variables to the specification.
2. The residuals of the model also provide important information for testing. If an adequate model has been fitted, the residuals should be approximately white noise.

As shown by Tables 7.2 to 7.8, three crucial aspects of the residuals are their autocorrelations, their partial autocorrelations, and the values of the Box-Pierce-Ljung statistic, which tests the joint significance of subsets of autocorrelation coefficients.

[19] For a comprehensive treatment see James D. Hamilton, *Time Series Analysis*, Princeton, 1994, Chapter 5.

Tables 7.5, 7.7, and 7.8 illustrate the appearance of white noise series compared with Tables 7.2, 7.3, 7.4, and 7.6. If the residuals depart significantly from white noise, the model is unsatisfactory and has to be respecified.

7.5 FORECASTING

The main purpose of fitting ARMA schemes is to project the series forward beyond the sample period. Such projections are sometimes used as a benchmark to compare with forecasts yielded by more complicated multivariate models. In projections or forecasts there are two inevitable sources of error, namely,

- Error due to ignorance of future innovations
- Error due to differences between true and estimated parameter values

In this section we will deal only with the first source of error, illustrating the principles involved with a few low-order processes.

Consider first the AR(1) scheme,

$$y_t - \mu = \alpha(y_{t-1} - \mu) + \epsilon_t \qquad |\alpha| < 1 \qquad \epsilon_t \text{ iid}(0, \sigma^2)$$

which for some purposes is more conveniently written as

$$y_t = (1 - \alpha)\mu + \alpha y_{t-1} + \epsilon_t \tag{7.65}$$

In all that follows we will assume that observations on y are available for periods 1 to n, and that all forecasts are made conditional on information available at time n. Thus

$$\begin{aligned} y_{n+s} &= \text{(unknown) value of } y \text{ in future period } n+s \\ \hat{y}_{n+s} &= \text{forecast of } y_{n+s} \text{ made on the basis of information available at time } n \\ e_{n+s} &= y_{n+s} - \hat{y}_{n+s} = \text{forecast error} \end{aligned}$$

The **mean squared error** (MSE) of a forecast is simply the average, or expected, squared forecast error. This treats positive and negative forecast errors symmetrically and is a widely used criterion for the choice of a forecasting rule. We wish to find a forecasting rule that will minimize MSE. It can be shown that the minimum MSE forecast of y_{n+s} is the **conditional expectation** of y_{n+s}, given information available at time n.[20] As an illustration, consider forecasting y_{n+1} for the AR(1) process. The true value is

$$y_{n+1} = (1 - \alpha)\mu + \alpha y_n + \epsilon_{n+1} \tag{7.66}$$

The minimum MSE forecast is then

$$\hat{y}_{n+1} = E(y_{n+1} \mid y_n) = (1 - \alpha)\mu + \alpha y_n \tag{7.67}$$

[20] Hamilton, ibid., Chapter 4.

The only unknown on the right side of Eq. (7.66) is the innovation in period $n + 1$, and it is replaced by its zero expectation. The forecast in (7.67) may be rearranged as

$$(\hat{y}_{n+1} - \mu) = \alpha(y_n - \mu) \tag{7.68}$$

that is, y_{n+1} is forecast to differ from the mean by a fraction α of the deviation in period n. The forecast error is $e_{n+1} = y_{n+1} - \hat{y}_{n+1} = \epsilon_{n+1}$ and so $\text{var}(e_{n+1}) = \sigma^2$.

Turning to period $n + 2$, we need to express y_{n+2} in terms of y_n and innovations since that time. Substitution in Eq. (7.65) gives

$$\begin{aligned} y_{n+2} &= (1 - \alpha)\mu + \alpha y_{n+1} + \epsilon_{n+2} \\ &= (1 - \alpha)\mu + \alpha[(1 - \alpha)\mu + \alpha y_n + \epsilon_{n+1}] + \epsilon_{n+2} \\ &= (1 - \alpha^2)\mu + \alpha^2 y_n + \alpha\epsilon_{n+1} + \epsilon_{n+2} \end{aligned}$$

The minimum MSE forecast is

$$\hat{y}_{n+2} = (1 - \alpha^2)\mu + \alpha^2 y_n$$

which may be displayed as

$$(\hat{y}_{n+2} - \mu) = \alpha^2(y_n - \mu) = \alpha(\hat{y}_{n+1} - \mu) \tag{7.69}$$

The forecasts from the AR(1) model approach the unconditional mean μ exponentially as the forecast horizon increases. The forecast error variance is $\text{var}(e_{n+2}) = \sigma^2(1 + \alpha^2)$.

Proceeding in this way, we find

$$y_{n+s} = (1 - \alpha^s)\mu + \alpha^s y_n + (\epsilon_{n+s} + \alpha\epsilon_{n+s-1} + \cdots + \alpha^{s-1}\epsilon_{n+1})$$

The forecast is

$$(\hat{y}_{n+s} - \mu) = \alpha^s(y_n - \mu) \tag{7.70}$$

and the forecast error variance is

$$\text{var}(e_{n+s}) = (1 + \alpha^2 + \alpha^4 + \cdots + \alpha^{2(s-1)})\sigma^2 \tag{7.71}$$

Clearly $\qquad \hat{y}_{n+s} \rightarrow \mu \qquad \text{as} \qquad s \rightarrow \infty$

and $\qquad \text{var}(e_{n+s}) \rightarrow \dfrac{\sigma^2}{1 - \alpha^2} = \sigma_y^2 \qquad \text{as} \qquad s \rightarrow \infty$

Thus as the forecast horizon increases, the forecast value tends to the unconditional mean of the process, and the forecast error variance increases toward the unconditional variance of the process.

7.5.1 MA(1) Process

The MA(1) process is

$$y_t = \mu + \epsilon_t - \beta\epsilon_{t-1} \tag{7.72}$$

Again forecasting from period n, we find

$$\hat{y}_{n+1} = \mu - \beta\epsilon_n \tag{7.73}$$

because ϵ_{n+1} is unknown at period n. Implementation of Eq. (7.73), however, does require knowledge of ϵ_n. From Eq. (7.72) it is clear that this in turn requires knowledge of previous values of ϵ. Thus, the strict implementation of Eq. (7.73) requires knowledge of ϵ_0. In practice this is often set at the expected value of zero in order to start the process off. This approximation is obviously of lesser importance as the sample size increases. Clearly $\text{var}(e_{n+1}) = \sigma^2$. Looking two periods ahead, we see that

$$y_{n+2} = \mu + \epsilon_{n+2} - \beta\epsilon_{n+1}$$

and the minimum MSE forecast is

$$\hat{y}_{n+2} = \mu$$

Thus for the MA(1) scheme

$$\hat{y}_{n+s} = \mu \qquad s \geq 2 \tag{7.74}$$

and

$$\text{var}(e_{n+s}) = (1 + \beta^2)\sigma^2 = \sigma_y^2 \qquad s \geq 2 \tag{7.75}$$

From two periods out, the forecast from the MA(1) scheme is simply the unconditional mean of the series, and the forecast error variance is the variance of the series.

7.5.2 ARMA(1,1) Process

As a third example we will combine the AR(1) and MA(1) processes to give the ARMA(1, 1) scheme,

$$(y_t - \mu) = \alpha(y_{t-1} - \mu) + \epsilon_t - \beta\epsilon_{t-1} \tag{7.76}$$

The minimum MSE forecast for period $n + 1$ is then

$$\hat{y}_{n+1} - \mu = \alpha(y_n - \mu) - \beta\epsilon_n$$

This result differs from the AR(1) forecast only by the term in $\beta\epsilon_n$. The forecast error variance is $\text{var}(e_{n+1}) = \sigma^2$. Repeated use of Eq. (7.76) gives

$$(y_{n+2} - \mu) = \alpha^2(y_n - \mu) + \epsilon_{n+2} + (\alpha - \beta)\epsilon_{n+1} - \alpha\beta\epsilon_n$$

The forecast for period $n + 2$ is then

$$(\hat{y}_{n+2} - \mu) = \alpha^2(y_n - \mu) - \alpha\beta\epsilon_n = \alpha(\hat{y}_{n+1} - \mu) \tag{7.77}$$

Thus, as in the AR(1) case, successive forecasts deviate from the mean in a declining exponential fashion. The forecast error variance is $\text{var}(e_{n+2}) = \sigma^2[1 + (\alpha - \beta)^2]$. By continuing in this way it may be shown that

$$(\hat{y}_{n+s} - \mu) = \alpha^s(y_n - \mu) - \alpha^{s-1}\beta\epsilon_n \tag{7.78}$$

The forecast thus tends to the unconditional mean as the forecast horizon increases. Likewise, it may be shown that[21]

$$\operatorname{var}(e_{n+s}) \to \sigma^2\left(\frac{1 - 2\alpha\beta + \beta^2}{1 - \alpha^2}\right) \qquad \text{as} \qquad s \to \infty \tag{7.79}$$

As shown in Eq. (7.41), this limiting variance is the variance of the y series.

7.5.3 ARIMA(1,1,0) Process

As a final illustration, consider a series z whose first differences follow an AR(1) scheme:

$$\begin{aligned} z_t - z_{t-1} &= y_t \\ (y_t - \mu) &= \alpha(y_{t-1} - \mu) + \epsilon_t \end{aligned} \tag{7.80}$$

From Eq. (7.80) we can write

$$\begin{aligned} z_{n+s} &= z_n + y_{n+1} + \cdots + y_{n+s} \\ &= (z_n + s\mu) + (y_{n+1} - \mu) + \cdots + (y_{n+s} - \mu) \end{aligned}$$

Continuous substitution for the $(y_t - \mu)$ terms gives

$$z_{n+s} = z_n + s\mu + \frac{\alpha(1 - \alpha^s)}{1 - \alpha}(y_n - \mu) + e_{n+s} \tag{7.81}$$

where

$$\begin{aligned} e_{n+s} &= \epsilon_{n+s} + (1 + \alpha)\epsilon_{n+s-1} + (1 + \alpha + \alpha^2)\epsilon_{n+s-2} + \cdots \\ &\quad + (1 + \alpha + \alpha^2 + \cdots + \alpha^{s-1})\epsilon_{n+1} \end{aligned} \tag{7.82}$$

The forecasts are given by the first three terms on the right side of Eq. (7.81), two of which increase with the forecast horizon, s. Notice, however, that the term in the initial value, z_n, does not fade away. From Eq. (7.82) the forecast error variance is

$$\operatorname{var}(e_{n+s}) = \sigma^2\left[1 + (1 + \alpha)^2 + (1 + \alpha + \alpha^2)^2 + \cdots + (1 + \alpha + \cdots + \alpha^{s-1})^2\right] \tag{7.83}$$

This variance increases monotonically with s. The forecasts of a nonstationary series become ever more imprecise as the forecast horizon increases.

All the formulae in this section are based on the assumption that the parameters of the process are known precisely. In practice they are replaced by sample estimates. The point forecasts will still be MSE asymptotically, but the estimated forecast error variances will understate the true values because the formulae do not allow for coefficient error. The EViews software, however, calculates the error variances correctly by allowing for coefficient uncertainty as well as ignorance of future innovations.

[21] See Problem 7.8.

7.6 SEASONALITY

Many series that are measured at regular intervals within a year may well display seasonal regularities. Construction activity will be lower in winter months, unemployment tends to peak in the summer, and so on. In other words a variable may be more closely related to its value in the same quarter (month, week, etc.) of the previous year than to its value in the immediately preceding quarter. Thus, for quarterly data we would be led to specify

$$x_t = \phi x_{t-4} + u_t \tag{7.84}$$

If the u series were white noise, the acf of this process would consist of exponentially declining spikes (under the usual stationarity assumption) at lags 4, 8, 12, etc. The intervening autocorrelations would all be zero. Most economic series, however, display some continuity over *adjacent* periods. Thus a white noise assumption in Eq. (7.84) is inappropriate. Suppose therefore that we specify an AR(1) scheme for u,

$$u_t = \alpha u_{t-1} + \epsilon_t \tag{7.85}$$

where ϵ is a white noise series. Combining the two relations gives

$$(1 - \alpha L)(1 - \phi L^4)x_t = \epsilon_t \tag{7.86}$$

This is a simple example of an **autoregressive multiplicative seasonal model.** The shorthand designation is AR(1) × SAR(1). On multiplying out, the equation may be rewritten as

$$x_t = \alpha x_{t-1} + \phi x_{t-4} - \alpha\phi x_{t-5} + \epsilon_t \tag{7.87}$$

This is seen to be a special case of a general AR(5) process, with two coefficients set to zero and a nonlinear relation between the remaining three coefficients. Because of the fifth order we expect the pacf to cut off after lag 5. The derivation of the theoretical acf is complicated. A typical pattern is shown by the empirical correlogram in Table 7.12. This has been generated from Eq. (7.87) by setting $\alpha = \phi = 0.8$. The autocorrelations decline with increasing lags, but there are relative peaks at lags 4, 8, 12, and 16. The partial autocorrelations are essentially zero after lag 5, with a distinctive positive spike at lag 1 and a distinctive negative spike at lag 5.

In a similar fashion one may specify **moving average multiplicative seasonal models.** An MA(1) × SMA(1) model would be

$$x_t = (1 - \beta L)(1 - \theta L^4)\epsilon_t \tag{7.88}$$

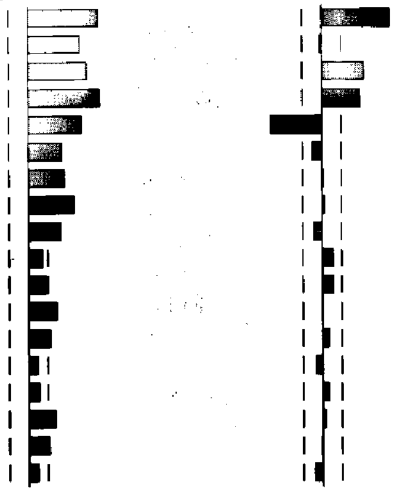
Here one expects the autocorrelations to cut off after lag 5 and the partial autocorrelations to damp away.

More generally one may specify mixed models combining AR, SAR, MA, and SMA components. It is extremely difficult to make tentative judgments on the orders of such models.[22] Finally, one should note that the usual preliminary check for

[22] Walter Vandaele, *Applied Time Series and Box-Jenkins Models*, Academic Press, 1983, contains many illustrative correlograms for various schemes and some detailed empirical analyses of real data. Study of such material helps develop the judgment so necessary in univariate time series analysis.

TABLE 7.12
Correlogram of seasonal series

Sample: 6–100
Included observations: 95



Autocorrelation	Partial correlation		AC	PAC	Q-stat	Prob.
		1	0.765	0.765	57.363	0.000
		2	0.579	−0.015	90.590	0.000
		3	0.639	0.484	131.49	0.000
		4	0.782	0.432	193.41	0.000
		5	0.577	−0.562	227.49	0.000
		6	0.365	−0.100	241.27	0.000
		7	0.379	0.011	256.31	0.000
		8	0.498	0.024	282.61	0.000
		9	0.341	−0.083	295.04	0.000
		10	0.164	0.122	297.94	0.000
		11	0.194	0.114	302.05	0.000
		12	0.322	0.002	313.58	0.000
		13	0.233	0.071	319.68	0.000
		14	0.084	−0.071	320.49	0.000
		15	0.127	0.064	322.36	0.000
		16	0.268	0.027	330.72	0.000
		17	0.223	−0.003	336.60	0.000
		18	0.078	−0.093	337.32	0.000

The vertical dashed lines represent two standard errors around zero.
AC = Autocorrelation coefficient; PAC = Partial correlation coefficient; Q = Box-Pierce-Ljung statistic [Eq. (6.58)]; Prob. = P-value for hypothesis that all autocorrelation coefficients to this point are zero.

stationarity is required before fitting these models. In some cases both first-order differencing and seasonal differencing may be required to induce stationarity. If z denotes a quarterly series, the appropriate differencing may be $(1-L)(1-L^4)z_t$ in order to yield a stationary series for analysis.

7.7 A NUMERICAL EXAMPLE: MONTHLY HOUSING STARTS

Figure 7.4 shows housing starts in the United States. The series is *Total New Private Housing Units Started* (thousands, not seasonally adjusted) from the DRI/McGraw Hill data set. The Citibase label for the series is HS6FR. First of all, we check for stationarity to see if we should construct a model in the levels of the series. Visual inspection does not give any strong indication of nonstationarity, and this is confirmed by formal tests. An augmented Dickey-Fuller test gives the result shown in Table 7.13. The unit root hypothesis is strongly rejected. Table 7.14 shows the correlogram, which tells a similar story. This looks very similar to the correlogram of an autoregressive multiplicative seasonal series in Table 7.12. The autocorrelation coefficients decline and then rise to a relative peak at lag 12 before declining substantially at higher lags, and the partial autocorrelations show a positive spike at lag 1 and a negative spike at lag 13. These patterns suggest an AR(1)×SAR(12) as a first

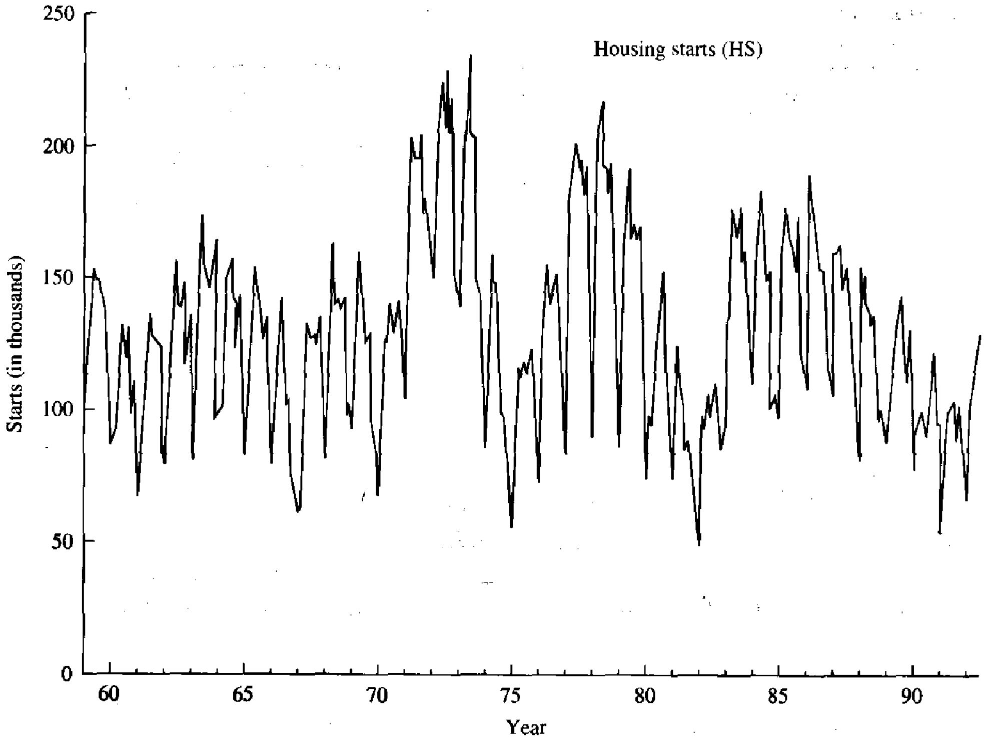


FIGURE 7.4
Total new private housing units started; no seasonal adjustment.

approximation to the series.[23] We will use observations from 1959:01 to 1984:12 for fitting, and the remaining observations from 1985:01 to 1992:04 for out-of-sample forecast tests. Table 7.15 shows the results of fitting $(1 - \alpha L)(1 - \phi L^{12})HS_t$, with allowance for a nonzero intercept. All coefficients are highly significant. The regression accounts for 86 percent of the variance of housing starts. There is, however, substantial residual variation: The standard error of the regression is more than 11 percent of the mean value of the dependent variable.

Let us look at the forecasting performance of the equation. Using only data up to 1984:12, we forecast the remaining 88 observations up to 1992:04. The results are shown in Fig. 7.5. The forecast is reasonably good for the first 12 months, picking up the seasonal pattern and only somewhat underestimating the high level of activity. Both features, however, get progressively worse. The forecast seasonal pattern diminishes as it must, because the autoregressive coefficients are numerically less than one and forecast innovations are set at zero. The forecast underpredicts the high activity of the middle and late 1980s and overpredicts the low activity of the early

[23] Notice that we have previously described this model as AR(1) × SAR(1), where 1 indicates the order of each component. SAR(12) is computer-oriented notation, telling the computer that the lag in the first-order seasonal component is 12 months.

TABLE 7.13

ADF Test on HS

ADF test statistic	−4.969008	1%	Critical value*	−3.4490
		5%	Critical value	−2.8691
		10%	Critical value	−2.5708

*MacKinnon critical values for rejection of hypothesis of a unit root.

Augmented Dickey-Fuller test equation
LS // dependent variable is D(HS)
Sample: 1959:06–1992:04
Included observations: 395
Excluded observations: 0 after adjusting endpoints

Variable	Coefficient	Std. error	T-statistic	Prob.
HS(−1)	−0.148240	0.029833	−4.969008	0.0000
D(HS(−1))	0.219352	0.049430	4.437661	0.0000
D(HS(−2))	0.111332	0.049647	2.242480	0.0255
D(HS(−3))	−0.065359	0.049837	−1.311461	0.1905
D(HS(−4))	−0.195488	0.049835	−3.922725	0.0001
C	18.73706	3.901098	4.803023	0.0000

R-squared	0.198754	Mean dependent var	−0.108354
Adjusted R-square	0.188455	S.D. dependent var	19.94195
S.E. of regression	17.96486	Akaike info criterion	5.791908
Sum squared resid	125544.3	Schwartz criterion	5.852347
Log likelihood	−1698.383	F-statistic	19.29878
Durbin-Watson stat	1.956461	Prob(F-statistic)	0.000000

1990s. The **mean absolute percentage error** of the forecasts is 25.2 percent. All actual values, however, lie well within the forecast confidence limits. A Chow forecast test for the 88 forecasts returns an F value of 0.66, with a corresponding P value of 0.99, so the hypothesis of a stable AR scheme for this specification is not rejected. A final check on this model is to test whether the residuals are white noise, or at least nonautocorrelated, for this is an objective of univariate modeling. Table 7.16 gives the correlogram of the residuals. These residuals display significant autocorrelation and partial autocorrelation coefficients at lag 1 and also at lags 11, 12, and 13. Thus a more complicated model is required.

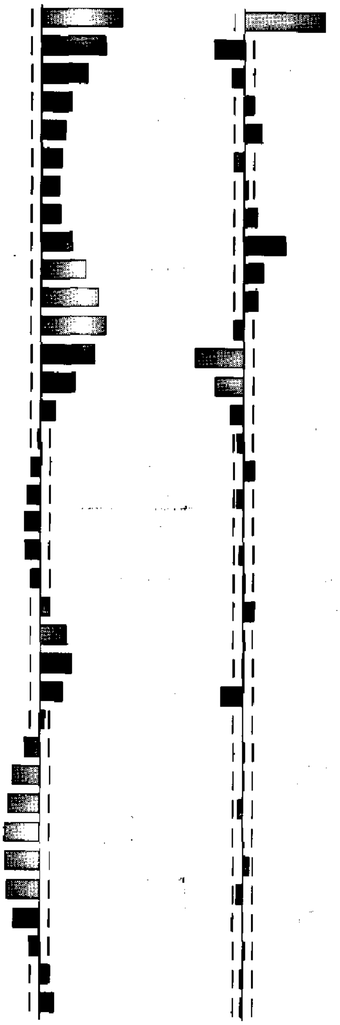
Let us try a mixed model, incorporating some MA terms as well as the AR terms,

$$(1-\alpha L)(1-\phi L^{12})\text{HS}_t = (1-\beta L)(1-\theta L^{12})\epsilon_t \tag{7.89}$$

Table 7.17 shows the results of fitting this specification. Both autoregressive terms are highly significant, as is the seasonal MA term. The forecasts from this model, shown in Fig. 7.6, are a substantial improvement over those given by the purely autoregressive scheme. The seasonal pattern in the forecasts is now well sustained over the forecast horizon, largely because the SAR(12) coefficient in Table 7.17 is much larger than the corresponding coefficient in Table 7.15. As a consequence the forecasts are now very good for the first three or four years, as compared with just the first year in Fig. 7.5. The residuals from the mixed scheme are much closer to a white noise series than those from the autoregressive scheme, as the reader should verify by

TABLE 7.14
Correlogram of HS

Sample: 1959:01–1992:04
Included observations: 400



Autocorrelation	Partial correlation		AC	PAC	Q-stat	Prob.
		1	0.859	0.859	297.25	0.000
		2	0.660	−0.295	473.38	0.000
		3	0.453	−0.115	556.67	0.000
		4	0.299	0.083	593.07	0.000
		5	0.237	0.189	615.92	0.000
		6	0.197	−0.106	631.82	0.000
		7	0.183	0.038	645.58	0.000
		8	0.195	0.134	661.23	0.000
		9	0.300	0.437	698.28	0.000
		10	0.455	0.204	783.57	0.000
		11	0.611	0.149	938.09	0.000
		12	0.680	−0.096	1129.7	0.000
		13	0.560	−0.485	1259.8	0.000
		14	0.355	−0.276	1312.2	0.000
		15	0.138	−0.110	1320.2	0.000
		16	−0.023	−0.066	1320.4	0.000
		17	−0.091	0.109	1323.9	0.000
		18	−0.140	−0.059	1332.1	0.000
		19	−0.162	0.004	1343.2	0.000
		20	−0.154	−0.050	1353.2	0.000
		21	−0.065	0.015	1355.0	0.000
		22	0.096	0.104	1358.9	0.000
		23	0.250	0.024	1385.5	0.000
		24	0.318	−0.006	1428.9	0.000
		25	0.210	−0.210	1447.8	0.000
		26	0.029	0.003	1448.1	0.000
		27	−0.165	0.021	1459.9	0.000
		28	−0.302	0.011	1499.4	0.000
		29	−0.361	−0.046	1556.0	0.000
		30	−0.395	0.014	1623.7	0.000
		31	−0.396	0.061	1692.1	0.000
		32	−0.377	−0.061	1754.2	0.000
		33	−0.273	0.006	1786.8	0.000
		34	−0.102	0.028	1791.3	0.000
		35	0.061	−0.037	1793.0	0.000
		36	0.137	−0.031	1801.3	0.000

The vertical dashed lines represent two standard errors around zero.
AC = Autocorrelation coefficient; PAC = Partial correlation coefficient; Q = Box-Pierce-Ljung statistic [Eq. (6.58)]; Prob. = P-value for hypothesis that all autocorrelation coefficients to this point are zero.

computing the relevant correlogram. The specification in Eq. (7.89) is not meant to be the last word on this series. The reader should experiment with other specifications. In view of the high SAR(12) coefficient in Table 7.17 it would be interesting to fit an ARMA model to the seasonal differences, $\Delta_{12}HS_t = HS_t - HS_{t-12}$, of the housing series.

TABLE 7.15
AR multiplicative seasonal model

LS // dependent variable is HS
Sample: 1960:02–1984:12
Included observations: 299
Excluded observations: 0 after adjusting endpoints
Convergence achieved after 5 iterations

Variable	Coefficient	Std. error	T-statistic	Prob.
C	130.9373	19.92150	6.572662	0.0000
AR(1)	0.867224	0.028964	29.94124	0.0000
SAR(12)	0.678202	0.043346	15.64631	0.0000

R-squared	0.860238	Mean dependent var	128.9087
Adjusted R-squared	0.859294	S.D. dependent var	39.33173
S.E. of regression	14.75365	Akaike info criterion	5.392964
Sum squared resid	64430.39	Schwartz criterion	5.430092
Log likelihood	−1227.511	F-statistic	910.9449
Durbin-Watson stat	2.259440	Prob(F-statistic)	0.000000

Inverted AR Roots	.97	.87	.84 + .48i	.84
−.48i				
.48 + .84i	.48 − .84i	.00 + .97i	−.00 − .97i	
−.48 + .84i	−.48 − .84i	−.84 − .48i	−.84 + .48i	
−.97				

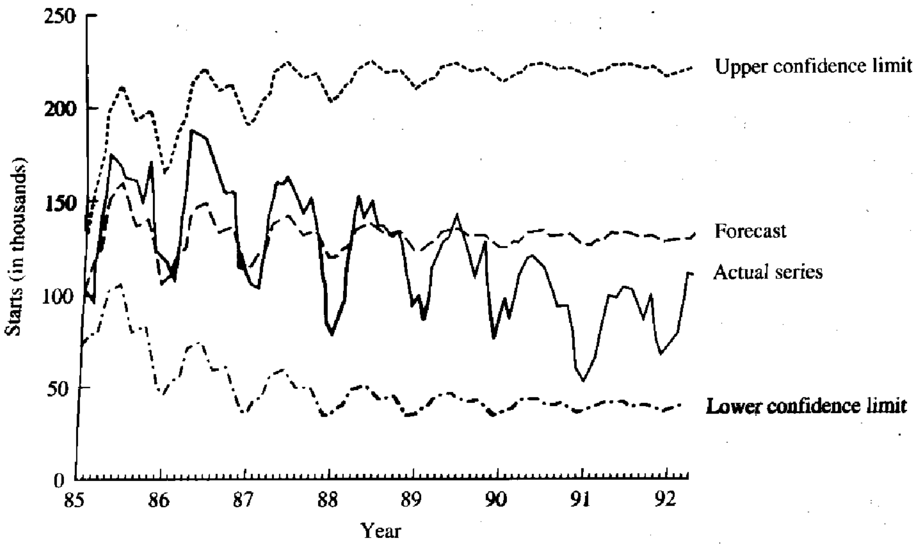
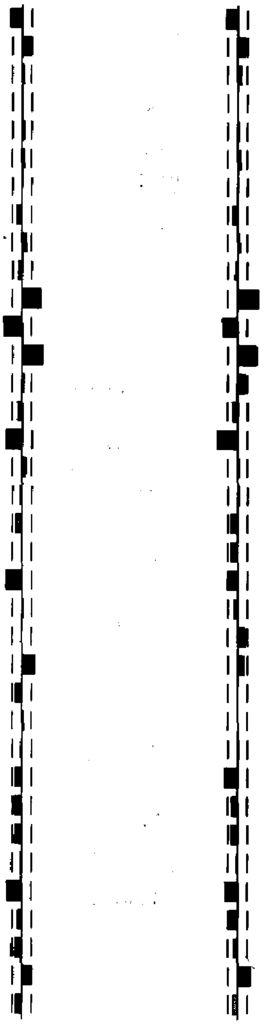


FIGURE 7.5
Forecasts of housing starts.

TABLE 7.16
Correlogram of residuals

Sample: 1960:02–1984:12
Included observations: 299
Q-statistic probabilities adjusted for two ARMA term(s)



Autocorrelation	Partial correlation		AC	PAC	Q-stat	Prob.
		1	−0.130	−0.130	5.1382	
		2	0.091	0.075	7.6553	
		3	0.020	0.042	7.7787	0.005
		4	0.004	0.004	7.7828	0.020
		5	0.027	0.023	7.9984	0.046
		6	0.025	0.030	8.1942	0.085
		7	0.005	0.008	8.2031	0.145
		8	−0.052	−0.058	9.0266	0.172
		9	0.049	0.033	9.7615	0.202
		10	−0.040	−0.023	10.260	0.247
		11	0.200	0.193	22.815	0.007
		12	−0.194	−0.156	34.597	0.000
		13	0.218	0.170	49.585	0.000
		14	0.023	0.076	49.747	0.000
		15	−0.042	−0.055	50.304	0.000
		16	−0.145	−0.211	57.033	0.000
		17	0.047	0.034	57.733	0.000
		18	−0.017	0.016	57.825	0.000
		19	−0.065	−0.061	59.169	0.000
		20	−0.001	−0.062	59.170	0.000
		21	−0.161	−0.111	67.566	0.000
		22	0.010	−0.044	67.600	0.000
		23	−0.007	0.074	67.617	0.000
		24	0.135	0.057	73.541	0.000
		25	−0.067	0.004	75.027	0.000
		26	0.017	0.003	75.127	0.000
		27	−0.005	0.025	75.137	0.000
		28	−0.065	−0.134	76.548	0.000
		29	−0.088	−0.055	79.154	0.000
		30	−0.074	−0.059	80.983	0.000
		31	−0.004	−0.025	80.988	0.000
		32	−0.150	−0.128	88.596	0.000
		33	−0.048	−0.088	89.389	0.000
		34	−0.069	−0.032	90.989	0.000
		35	0.093	0.129	93.933	0.000
		36	−0.064	−0.046	95.331	0.000

The vertical dashed lines represent two standard errors around zero.
AC = Autocorrelation coefficient; PAC = Partial correlation coefficient; Q = Box-Pierce-Ljung statistic [Eq. (6.58)]; Prob. = *P*-value for hypothesis that all autocorrelation coefficients to this point are zero.

TABLE 7.17
A mixed multiplicative model

LS // dependent variable is HS
Sample: 1960:02–1984:12
Included observations: 299
Excluded observations: 0 after adjusting endpoints
Convergence achieved after 12 iterations

Variable	Coefficient	Std. error	T-statistic	Prob.
C	395.9867	715.1165	0.553737	0.5802
AR(1)	0.945961	0.019845	47.66749	0.0000
SAR(12)	0.996715	0.006046	164.8596	0.0000
MA(1)	−0.167314	0.052647	−3.178023	0.0016
SMA(12)	−0.928458	0.018316	−50.69104	0.0000

R-squared	0.904386	Mean dependent var	128.9087
Adjusted R-squared	0.903085	S.D. dependent var	39.33173
S.E. of regression	12.24445	Akaike info criterion	5.026727
Sum squared resid	44078.42	Schwartz criterion	5.088607
Log likelihood	−1170.758	F-statistic	695.2121
Durbin-Watson stat	2.073135	Prob(F-statistic)	0.000000

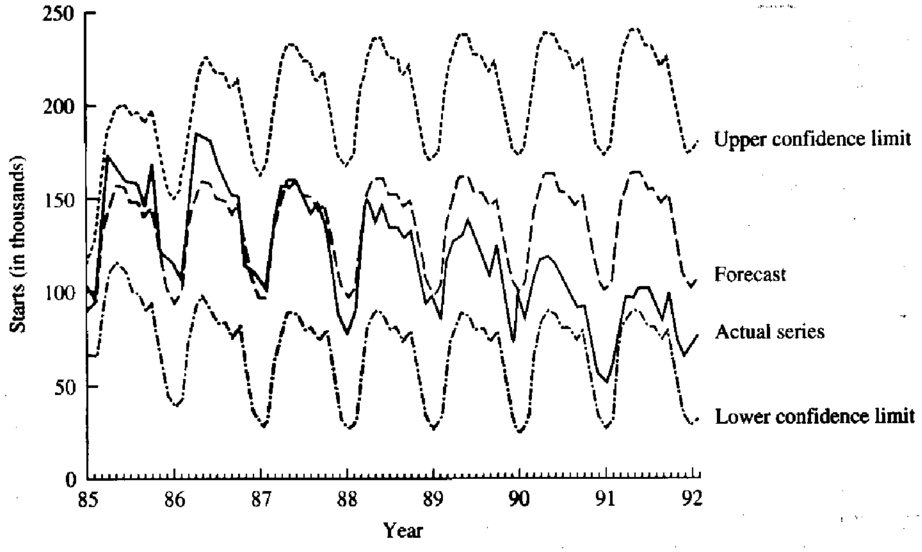


FIGURE 7.6
Forecasts from mixed model.

PROBLEMS

7.1. A demand/supply model is specified as

$$D\colon P_t = \alpha_0 + \alpha_1 Q_t + \alpha_2 Y_t + u_t$$
$$S\colon Q_t = \beta_0 + \beta_1 P_{t-1} + \beta_2(P_{t-1} - P_{t-2}) + v_t$$

where P indicates price, Q quantity, and Y income. Derive the AR equations for P and Q, determining the order and showing that the AR coefficients are identical.

7.2. Derive the acf and pacf for the MA(2) scheme, $u_t = \epsilon_t - \beta_1\epsilon_{t-1} - \beta_2\epsilon_{t-2}$.

7.3. Derive the acf for ARMA(2,1) and ARMA(2,2) processes.

7.4. Derive the mean, variance, and autocorrelations of Δu_t in Eq. (7.46).

7.5. (*a*) Evaluate $\partial u_{t+s}/\partial\epsilon_t$ for $|\alpha| < 1$and $\alpha = 1$ in the model $(1 - \alpha L)u_t = \epsilon_t - \beta\epsilon_{t-1}$.
(*b*) Show that the effect of a unit impulse in ϵ_t on u_{t+s} in the model $(1 - L)(1 - \alpha L)u_t = \epsilon_t$ is $(1 - \alpha^{s+1})/(1 - \alpha)$.

7.6. Carry out appropriate unit root tests on the Y2 and Y5 series.

7.7. A macroeconomist postulates that the log of U.S. real GNP can be represented by

$$A(L)(y_t - \delta_0 - \delta_1 t) = \epsilon_t$$

where $$A(L) = 1 - \alpha_1 L - \alpha_2 L^2$$

An OLS fit yields

$$y_t = -0.321 + 0.0030t + 1.335y_{t-1} - 0.401y_{t-2} + u_t$$

Determine the values of α_1, α_2, δ_0, and δ_1.
Compute the roots of the characteristic equation.
What is the estimated value of $A(1)$?
An alternative specification fitted to the same data yields

$$\Delta y_t = 0.003 + 0.369\Delta y_{t-1} + v_t$$

What are the roots of this equation?

(Regressions from Rudebusch, *op. cit.*)

7.8. Prove the results for the ARMA(1,1) process stated in Eqs. (7.78) and (7.79).

7.9. Prove the results for the ARIMA(1,1,0) process stated in Eqs. (7.81) and (7.83).

7.10. Try other ARMA schemes of your choice for the housing data of Section 7.7. In particular, try fitting an ARMA model to the seasonal differences of the series, and compare with the mixed model given in the text. It might also be interesting to fit a model to

$$y_t = (1 - L)(1 - L^{12})\mathrm{HS}_t$$

CHAPTER 8

Autoregressive Distributed Lag Relationships

The multiple regression equation has already been studied in Chapter 3, where it was introduced as

$$y_t = \beta_1 + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + u_t \qquad t = 1, 2, \ldots, n$$

In that chapter no specific attention was paid to whether the sample data were of time series or cross-section form. Here we will concentrate specifically on time series data. In Chapter 7 the only regressors considered were lagged values of the dependent variable. Now the regressors may include *lagged* values of the dependent variable and *current* and *lagged* values of one or more explanatory variables. Such a relation is called an **autoregressive distributed lag** (ADL) relation. The theoretical properties of ADL schemes will be outlined in the next section and problems of estimation, testing, and applications in subsequent sections.

8.1 AUTOREGRESSIVE DISTRIBUTED LAG RELATIONS

The simplest example of an ADL scheme is

$$y_t = m + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \epsilon_t \tag{8.1}$$

This is labeled ADL(1,1) since the dependent variable and the single explanatory variable are each lagged once. The ϵ series is presumed to be white noise. Inverting the lag polynomial in y gives

$$y_t = (1 + \alpha_1 + \alpha_1^2 + \cdots)m + (1 + \alpha_1 L + \alpha_1^2 L^2 + \cdots)(\beta_0 x_t + \beta_1 x_{t-1} + \epsilon_t)$$

Thus the current value of y depends on the current and all previous values of x and ϵ. Alternatively, this relation shows that the current value of x has an effect on the current and *future* values of y. Taking partial derivatives, we write

$$\frac{\partial y_t}{\partial x_t} = \beta_0$$

$$\frac{\partial y_{t+1}}{\partial x_t} = \beta_1 + \alpha_1\beta_0$$

$$\frac{\partial y_{t+2}}{\partial x_t} = \alpha_1\beta_1 + \alpha_1^2\beta_0$$

$$\vdots$$

The simple lags in Eq. (8.1) imply a set of dynamic responses in y to any given change in x. There is an immediate response, followed by short-run, medium-run, and long-run responses. The long-run effect of a unit change in x_t is obtained by summing the partial derivatives; provided the stability condition $|\alpha_1| < 1$ is satisfied, the sum is $(\beta_0 + \beta_1)/(1 - \alpha_1)$. Suppose that x is held constant at some level $\bar{x}$ indefinitely. Then, given the stability condition and setting the innovations at their expected value of zero, the foregoing relation shows that y will tend to a constant value $\bar{y}$, given by

$$\bar{y} = \frac{m}{1-\alpha_1} + \frac{\beta_0 + \beta_1}{1-\alpha_1}\bar{x} \tag{8.2}$$

This is a **static equilibrium** equation. A simpler alternative derivation is obtained by replacing all values of y and x in Eq. (8.1) by their respective long-run values and setting the innovation to zero.

8.1.1 A Constant Elasticity Relation

If y and x are the natural logarithms of Y and X, Eq. (8.2) implies a constant elasticity equilibrium relation

$$Y = AX^{\gamma} \tag{8.3}$$

or, in log form,

$$y = a + \gamma x$$

where

$$a = \frac{m}{1-\alpha_1} \qquad \gamma = \frac{\beta_0 + \beta_1}{1-\alpha_1} \tag{8.4}$$

8.1.2 Reparameterization

The properties of ADL relations can often be simply revealed by reparameterizing the equation. As an example, replace y_t by $y_{t-1} + \Delta y_t$ and x_t by $x_{t-1} + \Delta x_t$ in Eq. (8.1). The result is

$$\Delta y_t = m + \beta_0\Delta x_t - (1-\alpha_1)y_{t-1} + (\beta_0 + \beta_1)x_{t-1} + \epsilon_t \tag{8.5}$$

By using Eq. (8.4), Eq. (8.5) may be rearranged to give

$$\Delta y_t = \beta_0\Delta x_t - (1-\alpha_1)[y_{t-1} - a - \gamma x_{t-1}] + \epsilon_t \tag{8.6}$$

This formulation is an example of an **error correction model** (ECM). The current change in y is seen to be the sum of two components. The first is proportional to the current change in x, and the second is a partial correction for the extent to which y_{t-1} deviated from the equilibrium value corresponding to x_{t-1}. This deviation, or *equilibrium error,* is shown by the term in square brackets. If it is positive, there is a downward correction in the current period, given the stability condition on α_1. Conversely, a negative error produces an upward correction. In a static equilibrium Δx and Δy will each be zero. Making this substitution in Eq. (8.6) is yet another way of deriving the static equilibrium equation, Eq. (8.2).

The parameters in Eq. (8.5) could be estimated by running the OLS regression of Δy_t on a constant Δx_t, y_{t-1}, and x_{t-1}. From the four estimated coefficients and their variance-covariance matrix, one could derive estimates of the four parameters in Eq. (8.1), namely, m, α_1, β_0, β_1, and their standard errors. Alternatively, one could estimate these parameters directly by applying OLS to Eq. (8.1). As shown in Appendix 8.1, **the two procedures give identical results.** This important property is due to the fact that the move from Eq. (8.1) to Eq. (8.5) involves only linear, nonsingular transformations of the variables and does not impose any restrictions.

8.1.3 Dynamic Equilibrium

Instead of the static assumption, suppose that X grows at a steady rate g so that $\Delta x_t = g$ for all t. Given a constant elasticity of γ, the steady growth rate in Y will be γg. Substituting in Eq. (8.6) gives the dynamic equilibrium as

$$y = \frac{m - (\gamma - \beta_0)g}{1 - \alpha_1} + \gamma x \tag{8.7}$$

or

$$Y = AX^{\gamma} \qquad A = \exp\left[\frac{m - (\gamma - \beta_0)g}{1 - \alpha_1}\right] \tag{8.8}$$

Thus the multiplicative constant differs between the static and dynamic equilibrium cases. When there is zero growth, Eq. (8.7) reverts to Eq. (8.2).

8.1.4 Unit Elasticity

Under the constant elasticity specification in Eq. (8.8) the equilibrium ratio Y/X varies with the level of X. If the elasticity were a positive fraction, the ratio would go to zero for infinitely large X and, conversely, would increase without bound if the elasticity were greater than one. If, say, X represented total income and Y were expenditure on a consumption commodity or group of commodities, such implications would be implausible. A more plausible assumption would be a unit elasticity. Thus in some cases it may be desirable to test for unit elasticity and perhaps impose it on the estimation process. The hypothesis is

$$H_0: \gamma = \frac{\beta_0 + \beta_1}{1 - \alpha_1} = 1 \qquad \text{that is,} \qquad H_0: \alpha_1 + \beta_0 + \beta_1 = 1$$

The test may be carried out by estimating Eq. (8.1) and testing the appropriate linear restriction on the coefficients. Alternatively, one could estimate Eq. (8.5) and then test whether the sum of the coefficients on y_{t-1} and x_{t-1} is zero. An even simpler possibility is a further reparameterization of Eq. (8.5), which focuses attention on just a single coefficient. Add and subtract $(1-\alpha_1)x_{t-1}$ on the right-hand side of Eq. (8.5). The result is

$$\Delta y_t = m + \beta_0 \Delta x_t - (1-\alpha_1)(y_{t-1} - x_{t-1}) + (\beta_0 + \beta_1 + \alpha_1 - 1)x_{t-1} + \epsilon_t \tag{8.9}$$

The unit elasticity hypothesis is then tested by running the OLS regression of Δy_t on a constant, Δx_t, $(y_{t-1} - x_{t-1})$, and x_{t-1}. If the coefficient of x_{t-1} is significantly different from zero, the hypothesis of unit elasticity is rejected. If the hypothesis is not rejected, one may wish to impose it on the estimation process. Equation (8.9) then simplifies to

$$\Delta y_t = m + \beta_0 \Delta x_t - (1-\alpha_1)(y_{t-1} - x_{t-1}) + \epsilon_t \tag{8.10}$$

8.1.5 Generalizations

The ADL(p, q) scheme gives a richer lag structure that still retains the specification of just one explanatory variable:

$$A(L)y_t = m + B(L)x_t + \epsilon_t \tag{8.11}$$

with

$$A(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p$$
$$B(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_q L^q$$

As an illustration we will look at the $p = q = 2$ case,

$$y_t = m + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \epsilon_t \tag{8.12}$$

If we assume the variables to be in logarithmic form as before, the constant elasticity is

$$\gamma = \frac{B(1)}{A(1)} = \frac{\beta_0 + \beta_1 + \beta_2}{1 - \alpha_1 - \alpha_2}$$

The reparameterization of Eq. (8.12) may be based on period $t-1$ or period $t-2$. For the former we make the substitutions

$$y_t = y_{t-1} + \Delta y_t \qquad y_{t-2} = y_{t-1} - \Delta y_{t-1}$$
$$x_t = x_{t-1} + \Delta x_t \qquad x_{t-2} = x_{t-1} - \Delta x_{t-1}$$

Putting these expressions in Eq. (8.12) gives

$$\Delta y_t = m - \alpha_2 \Delta y_{t-1} + \beta_0 \Delta x_t - \beta_2 \Delta x_{t-1} - (1 - \alpha_1 - \alpha_2)(y_{t-1} - \gamma x_{t-1}) + \epsilon_t \tag{8.13}$$

The error correction term relates to period $t-1$, and all other variables are either current or lagged first differences. This formulation readily yields the dynamic equilibrium relation as

$$y = \frac{m + (\beta_0 - \beta_2 - \alpha_2\gamma - \gamma)g}{1 - \alpha_1 - \alpha_2} + \gamma x$$

Setting g to zero gives the static equilibrium relation and setting $\gamma = 1$ gives the unit elasticity relation.

Adding more right-hand-side variables gives the general ADL($p, q_1, q_2, \ldots, q_k$) scheme,

$$A(L)y_t = m + B_1(L)x_{1t} + B_2(L)x_{2t} + \cdots + B_k(L)x_{kt} + \epsilon_t \qquad (8.14)$$

where the orders of the lag polynomials are $p, q_1, q_2, \ldots, q_k$.

8.2 SPECIFICATION AND TESTING

The crucial practical question is how to implement an equation like (8.14). Specifically, what variables should appear as regressors, and what should be the orders of the lag polynomials? One looks to economic theory for guidance on the variables to be included. but theorists do not always speak with one voice, or, if they do, the message may be too general to be useful. For example, the theory of consumer behavior would suggest that the demand for a specific good should depend *inter alia* on the prices of *all* items in the consumption basket, threatening the impossible situation of having more variables than observation points. The situation is even more difficult with respect to lag specification, where there is basically little chance of theoretical guidance. In practice, there is an inevitable interaction between theory and data, with different specifications being discarded or modified in the light of empirical results. The question then becomes how best to conduct such specification searches. There is no procedure that commands general acceptance. Some critics throw up their hands in horror and decry all such activities as "data mining," unlikely to produce results of any value. Other practitioners are devotees of particular approaches, but the great army of empirical researchers often looks in vain for guidance. The power of the modern PC and the accessibility of data bases both exacerbate and alleviate the problem. The bewildering array of feasible computational procedures aggravates the problem of choice. On the other hand, provided a person knows what procedures he or she wants, it is now very easy and quite enjoyable to implement them with all the modern bells and whistles.

8.2.1 General to Simple and Vice Versa

For a long time in the development of econometrics, especially with limited computational resources, it was fairly common to start out with reasonably simple specifications of relations like Eq. (8.14) and then possibly to expand them by adding variables or lags or both as might seem suitable. This approach may be classified as **simple to general**. If a specification gave autocorrelated residuals, a Cochrane-Orcutt specification, usually with just a single autoregressive parameter, might be added as a palliative to the equation and the estimation left there. A serious defect of such an approach is that an excessively simple specification will likely give misleading information even about the effects of the variables actually included in the specification. For example, suppose the specified model is

$$y_t = \beta x_t + u_t \tag{8.15}$$

and the "true" model is

$$y_t = \beta x_t + \gamma z_t + v_t \tag{8.16}$$

The investigator would estimate β from Eq. (8.15) as

$$\begin{aligned} b = \frac{\sum yx}{\sum x^2} &= \frac{\sum(\beta x + \gamma z + v)x}{\sum x^2} \\ &= \beta + \gamma\frac{\sum zx}{\sum x^2} + \frac{\sum vx}{\sum x^2} \end{aligned}$$

If we make the usual assumptions that x and z are nonstochastic and that v is white noise, it follows that

$$E(b) = \beta + \gamma b_{zx}$$

where b_{zx} is the slope of the regression of z on x. **Thus b is biased and inconsistent, unless the sample observations on z and x have zero correlation. Omitting a relevant variable from the estimated specification invalidates the Gauss-Markov theorem.** One must not blithely assume that computing an OLS regression necessarily delivers best linear unbiased estimates! Furthermore, $\text{var}(b) = \sigma_v^2/\sum x^2$, but the pseudodisturbance in Eq. (8.15) is $u_t = \gamma z_t + v_t$. Thus the s^2 estimated from Eq. (8.15) will likely overestimate σ_v^2.

Suppose, on the other hand, that Eq. (8.15) is the correct specification but Eq. (8.16) is estimated. Now, instead of omitting a relevant variable, we include an irrelevant variable. OLS estimation gives

$$\hat{\beta} = \frac{1}{D}\left[\sum z^2 \sum yx - \sum xz \sum yz\right]$$

$$\hat{\gamma} = \frac{1}{D}\left[\sum x^2 \sum yz - \sum xz \sum yx\right]$$

where $D = \sum x^2 \sum z^2 - (\sum xz)^2$. On the assumption that Eq. (8.15) is the correct model, $E(\sum yx) = \beta \sum x^2$ and $E(\sum yz) = \beta \sum xz$. Thus,

$$E(\hat{\beta}) = \beta \qquad \text{and} \qquad E(\hat{\gamma}) = 0$$

Thus OLS now provides unbiased estimates of the population coefficients. It can also be shown that, in this case, the OLS residuals yield the usual unbiased estimate of the disturbance variance.[1]

[1] See Problem 8.4.

These two cases suggest that omitting relevant variables is more serious than including irrelevant ones because in the former case the coefficients will be biased, the disturbance variance overestimated, and conventional inference procedures rendered invalid, whereas in the latter the coefficients will be unbiased, the disturbance variance properly estimated, and the inference procedures valid. The resultant strategy is to start with a very catholic specification both in terms of included variables and lag structure. That specification should then be subjected to the various tests outlined in Chapters 4 and 6 for autocorrelation, heteroscedasticity, parameter constancy, etc. If the specification survives these tests, the second stage is to investigate whether various **reductions** are valid. With quarterly data a general specification might have included lags up to the fifth order. One might test whether all lags of a given order could be regarded as zero, or whether all coefficients on a given variable might be treated as zero, or whether other restrictions might be imposed. This **general to simple** approach is basically due to the work of David Hendry and his associates.[2] A numerical illustration with the gasoline data will be given in Section 8.4.

8.2.2 Estimation and Testing

Having specified the initial ADL equation, the next question is how the equation should be estimated and tested. The focus of attention is on a *single* equation, but can we ignore the generating process of the regressors in that equation? To put it in other words, do we have to formulate a *multiequation* model in order to get a proper analysis of just a single equation? To give the simplest possible explanation of the basic issues we will consider just a bivariate relation between y_t and x_t. We assume that the **data generation process** (DGP) for these variables can be approximated by some bivariate probability distribution (pdf), denoted by $f(y_t, x_t)$ for $t = 1, \ldots, n$. Such a pdf can always be factorized into the product of a *marginal* and a *conditional* density as, for example,

$$f(y_t, x_t) = f(x_t)f(y_t \mid x_t) \tag{8.17}$$

Let us further assume the pdf to be bivariate normal; that is,

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} \sim IN(\boldsymbol{\mu}, \boldsymbol{\Omega}) \qquad t = 1, \ldots, n \tag{8.18}$$

where $\sim IN$ reads "independently and normally distributed" and

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \boldsymbol{\Omega} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

[2] For a typical example of the approach see David F. Hendry and Neil R. Ericsson, "Modeling the Demand for Narrow Money in the United Kingdom and the United States," *European Economic Review*, **35**, 1991, 833–886. Comprehensive summaries of the approach are available in Neil R. Ericsson *et al.*, "PC-GIVE and David Hendry's Econometric Methodology," *Revista de Econometria*, **10**, 1990, 7–117; and in Christopher L. Gilbert, "Professor Hendry's Econometric Methodology," *Oxford Bulletin of Economics and Statistics*, **48**, 1986, 283–307.

give the mean vector and the variance-covariance matrix. The variables are thus allowed to be contemporaneously correlated ($\sigma_{12} \neq 0$), but the independence assumption implies zero autocorrelations. It was shown in Chapter 1 that y_t and x_t **each have** marginal distributions, which are univariate normal, that is,

$$y_t \sim N(\mu_1, \sigma_{11}) \qquad x_t \sim N(\mu_2, \sigma_{22})$$

and also that the distribution of y_t, *conditional on* x_t, is univariate normal, namely.

$$y_t \mid x_t \sim N[\alpha + \beta x_t, \sigma_{11}(1 - \rho^2)] \tag{8.19}$$

where $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$ is the correlation between y_t and x_t, and

$$\alpha = \mu_1 - \beta\mu_2 \qquad \beta = \rho\frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}} = \frac{\sigma_{12}}{\sigma_{22}} \tag{8.20}$$

The joint density can be factorized as the product of the marginal density for x_t and the conditional density for y_t given x_t. From Eq. (8.19) we may write

$$y_t = \alpha + \beta x_t + u_t \tag{8.21}$$

The "disturbance" u has zero mean, since $E(y_t \mid x_t) = \alpha + \beta x_t$. From Eq. (8.19), it also has a constant variance. For estimation purposes, however, its most important property is that it is stochastically independent of x. It is worth making a detailed derivation of this last property, as the method is also useful in more complicated cases. Rewrite Eq. (8.18) as

$$\begin{aligned} y_t &= \mu_1 + \epsilon_{1t} \\ x_t &= \mu_2 + \epsilon_{2t} \\ \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix} &\sim IN(\mathbf{0}, \mathbf{\Omega}) \end{aligned} \tag{8.22}$$

Multiply the second equation by σ_{12}/σ_{22} and subtract the result from the first. This gives

$$y_t - \frac{\sigma_{12}}{\sigma_{22}}x_t = \left(\mu_1 - \frac{\sigma_{12}}{\sigma_{22}}\mu_2\right) + \left(\epsilon_{1t} - \frac{\sigma_{12}}{\sigma_{22}}\epsilon_{2t}\right) \tag{8.23}$$

This expression is clearly Eq. (8.21) with the parameters defined in Eq. (8.20). Thus u_t in Eq. (8.21) is defined by

$$u_t = \epsilon_{1t} - \frac{\sigma_{12}}{\sigma_{22}}\epsilon_{2t} \tag{8.24}$$

It follows that $\operatorname{var}(u_t) = \sigma_{11} - \sigma_{12}^2/\sigma_{22} = \sigma_{11}(1-\rho^2)$ as given in Eq. (8.19). Further,

$$E(x_t u_t) = E(x_t\epsilon_{1t}) - \frac{\sigma_{12}}{\sigma_{22}}E(x_t\epsilon_{2t})$$

Using Eq. (8.22), we find $E(x_t\epsilon_{1t}) = \sigma_{12}$ and $E(x_t\epsilon_{2t}) = \sigma_{22}$. Thus, $E(x_t u_t) = 0$. Because the ϵ's have zero autocorrelations, all lagged covariances between x and u

are zero. Given normality, zero covariances imply stochastic independence. Using $x_t \| u_t$ to state "x_t and u_t are stochastically independent," we have

$$x_t \| u_{t+s} \qquad \text{for all } s \tag{8.25}$$

In the language of the Cowles Commission, x_t is said to be **exogenous** in Eq. (8.21).[3] More recently this condition is said to define **strict exogeneity** to distinguish it from other types of exogeneity, which will be introduced shortly. A related but less stringent condition is

$$x_t \| u_{t+s} \qquad \text{for all } s \geq 0 \tag{8.26}$$

that is, x_t is independent of the current and all future disturbances but not of past disturbances. Again in Cowles Commission terminology x_t is **predetermined** in Eq. (8.21).

The bivariate model in Eq. (8.18) can thus be reparameterized as

$$\begin{aligned} y_t &= \alpha + \beta x_t + u_t \\ x_t &= \mu_2 + \epsilon_{2t} \end{aligned} \tag{8.27}$$

From Eq. (8.24) it follows directly that $E(u_t\epsilon_{2t}) = 0$. Thus $[y_t \; x_t]'$ has mean vector $[\alpha + \beta\mu_2 \;\; \mu_2]'$ and variance-covariance matrix

$$\text{var}\begin{bmatrix} u_t \\ \epsilon_{2t} \end{bmatrix} = \begin{bmatrix} \sigma_{11} - \sigma_{12}^2/\sigma_{22} & 0 \\ 0 & \sigma_{22} \end{bmatrix} \tag{8.28}$$

The first equation in Eq. (8.27) satisfies all the requirements for the classical inference procedures described in previous chapters, namely, zero mean, homoscedastic, serially uncorrelated disturbances which are also distributed independently of the regressor(s). The conditional equation may thus be analyzed on its own. The marginal distribution of the regressor contains no information relevant to the parameters of the conditional equation.

Notice that we might alternatively have reparameterized the bivariate normal distribution as

$$\begin{aligned} x_t &= \gamma + \delta y_t + v_t \\ y_t &= \mu_1 + \epsilon_{1t} \end{aligned}$$

with $\gamma = \mu_2 - \delta\mu_1, \delta = \sigma_{12}/\sigma_{11}$, and $v_t = \epsilon_{2t} - \delta\epsilon_{1t}$. Now y_t would be exogenous in the conditional equation, which this time could be analyzed independently of the marginal equation for y.

The bivariate (multivariate) normal distribution is not a plausible DGP for economic variables, which typically display strong autocorrelation patterns rather than the implied zero autocorrelations. It is time to move to more realistic DGPs where the exogeneity issue becomes more complicated.

[3] For the simplest, but not always simple, exposition of the Cowles Commission approach, see William C. Hood and Tjalling C. Koopmans, *Studies in Econometric Method*, Wiley, 1953.

8.2.3 Exogeneity

The modern treatment of this subject extends and develops the Cowles Commission approach. The classic reference is the article by Engle, Hendry, and Richard, hereinafter referred to as EHR.[4] The basic elements in the EHR treatment will be explained in terms of a bivariate DGP used in their exposition. Let the bivariate DGP be

$$y_t = \beta x_t + \epsilon_{1t} \tag{8.29a}$$

$$x_t = \alpha_1 x_{t-1} + \alpha_2 y_{t-1} + \epsilon_{2t} \tag{8.29b}$$

Now both y and x will be autocorrelated. We still retain the assumption of normally and serially independent disturbances, that is,

$$\begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix} \sim IN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right] \tag{8.30}$$

Suppose that the focus of interest is Eq. (8.29a). The crucial question is the "exogeneity" of x. The message of EHR is that the question is ill-defined. It all depends on why Eq. (8.29a) is being analyzed. Three main purposes are distinguished:

1. To make inferences about one or more **parameters of interest**
2. To **forecast** y conditional on x
3. To test whether the relation in Eq. (8.29a) is **structurally invariant** to changes in the marginal distribution of x

Corresponding to these three purposes are **three types of exogeneity**, namely, **weak**, **strong**, and **super** exogeneity.

Weak exogeneity

In general any joint density can be factorized as the product of a marginal distribution of one or more variables and a conditional distribution of a scalar variable y on those variables. Let $\boldsymbol{\lambda}_1$ denote the parameters of the conditional distribution and $\boldsymbol{\lambda}_2$, the parameters of the marginal distribution. These parameters will be functions of the parameters $\boldsymbol{\theta}$ of the original joint density (DGP). Let $\boldsymbol{\psi}$ denote the parameters of interest. If the conditioning variables are weakly exogenous for $\boldsymbol{\psi}$, then inferences about $\boldsymbol{\psi}$ from the conditional distribution will be equivalent to inferences from the joint distribution. In other words, the marginal distribution of the conditioning variables contains no relevant information and may be ignored in the analysis.

Given the factorization into marginal and conditional densities, two conditions have to be satisfied for weak exogeneity to hold, namely,

$$\boldsymbol{\psi} = f(\boldsymbol{\lambda}_1) \tag{8.31}$$

[4] Robert F. Engle, David F. Hendry, and Jean-Francois Richard, "Exogeneity," *Econometrica*, **51**, 1983, 277–304.

that is, the parameters of interest can be expressed uniquely in terms of the parameters of the conditional distribution, and

$$\boldsymbol{\lambda}_1 \text{ and } \boldsymbol{\lambda}_2 \text{ are variation-free} \tag{8.32}$$

Variation-free means, loosely, that any parameter in $\boldsymbol{\lambda}_1$ is free to assume any value in its admissible range, irrespective of the values taken by parameters in $\boldsymbol{\lambda}_2$, and vice versa. There must be no cross restrictions, whether equalities or inequalities, between elements in the two sets.

These concepts may be illustrated by the model in Eq. (8.29). The process of multiplying Eq. (8.29*b*) by σ_{12}/σ_{22} and subtracting the result from Eq. (8.29*a*) gives a conditional equation

$$y_t = \delta_0 x_t + \delta_1 x_{t-1} + \delta_2 y_{t-1} + u_t \tag{8.33}$$

where

$$\begin{aligned} \delta_0 &= \beta + \frac{\sigma_{12}}{\sigma_{22}} \\ \delta_1 &= -\alpha_1 \frac{\sigma_{12}}{\sigma_{22}} \\ \delta_2 &= -\alpha_2 \frac{\sigma_{12}}{\sigma_{22}} \end{aligned} \tag{8.34}$$

The disturbance u_t in the conditional equation is the same as that already defined in Eq. (8.24), with the properties stated in Eqs. (8.25) and (8.28). The DGP in Eq. (8.29) can thus be reparameterized with Eq. (8.33) defining the conditional equation and Eq. (8.29*b*) the marginal equation. The various parameter sets are

$$\begin{aligned} \boldsymbol{\theta} &= (\beta, \alpha_1, \alpha_2, \sigma_{11}, \sigma_{12}, \sigma_{22}) \\ \boldsymbol{\lambda}_1 &= (\delta_0, \delta_1, \delta_2, \sigma_u^2) \qquad \boldsymbol{\lambda}_2 = (\alpha_1, \alpha_2, \sigma_{22}) \end{aligned} \tag{8.35}$$

Suppose that β is the parameter of interest. First, check if the condition in Eq. (8.31) is satisfied. Using Eq. (8.34) to express β in terms of the λ parameters gives

$$\beta = \delta_0 + \frac{\delta_1}{\alpha_1} = \delta_0 + \frac{\delta_2}{\alpha_2}$$

Thus β cannot be expressed solely in terms of $\boldsymbol{\lambda}_1$, and the first condition fails. Moreover, the fact that there are two equivalent expressions for β implies a cross restriction between elements of $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, namely,

$$\alpha_2\delta_1 - \alpha_1\delta_2 = 0$$

Thus, the parameters of the conditional and marginal distributions are not variation-free. Both conditions fail. The variable x_t is not weakly exogenous for β.

If, however, the two disturbances in Eq. (8.29) were independent ($\sigma_{12} = 0$), the parameter setup is

$$\begin{aligned} \boldsymbol{\theta} &= (\beta, \alpha_1, \alpha_2, \sigma_{11}, \sigma_{22}) \\ \boldsymbol{\lambda}_1 &= (\beta, \sigma_{11}) \qquad \boldsymbol{\lambda}_2 = (\alpha_1, \alpha_2, \sigma_{22}) \end{aligned} \tag{8.36}$$

Now ψ contains only the single element β, which in turn is in $\boldsymbol{\lambda}_1$, so the condition in Eq. (8.31) is satisfied. The variation-free condition is also clearly satisfied, so in this case x_t is weakly exogenous for β.

As another illustration, reinstate the assumption that $\sigma_{12} \neq 0$, but suppose that δ_0 is now the parameter of interest. Reference to Eq. (8.35) shows that the condition in Eq. (8.31) is satisfied but the variation-free condition is not, for Eq. (8.34) shows the cross restriction between elements of $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ already noted. The independence of x_t and u_t, however, means that δ_0, δ_1, and δ_2 could be *consistently* estimated by applying OLS to the conditional Eq. (8.33). But these estimates would not be fully efficient because they ignore the information in the cross restrictions. Thus the failure of weak exogeneity does not necessarily imply that inference from the conditional distribution is impossible or invalid. Depending on the parameters of interest, it may merely mean that the inference is not fully efficient.

As a final illustration, suppose $\alpha_2 = 0$ so that lagged y plays no role in the generation of x. If β is the parameter of interest, the first two equations in Eq. (8.34), which still hold, show that $\beta = \delta_0 + \delta_1/\alpha_1$, so the condition in Eq. (8.31) is not satisfied. The cross restriction has disappeared, so the parameters are now variation-free; but we cannot make inferences about β from the parameters of the conditional equation alone. If δ_0 were the parameter of interest, then Eqs. (8.31) and (8.32) would both be satisfied, so x_t would now be weakly exogenous for δ_0.

Strong exogeneity

If x_t is weakly exogenous for β and, *in addition*, y does not **Granger cause** x, then x is said to be strongly exogenous for β. Granger causality or noncausality is concerned with whether lagged values of y do or do not improve on the explanation of x obtainable from only lagged values of x itself.[5] A simple test is to regress x on lagged values of itself and lagged values of y. If the latter are jointly insignificant, y is said not to Granger cause x. If one or more lagged y values are significant then y is said to Granger cause x. The test, however, is often very sensitive to the number of lags included in the specification. Changing lag length can result in changed conclusions. If strong exogeneity holds, β may be estimated from the conditional distribution alone and used to make forecasts of y conditional on forecasts of x, the latter in turn being derived from the past history of x alone.

Super exogeneity

Super exogeneity holds if the parameters of the conditional distribution are invariant to changes in the marginal distribution of the conditioning variables. Suppose that in Eq. (8.29) y is GNP and x is the money stock. Equation (8.29*b*) might then represent a decision rule for the monetary authorities, setting the current money stock in response to last period's GNP and money stock; and Eq. (8.29*a*) would describe how economic agents set GNP in response to the money stock. Much attention has been given to the Lucas suggestion that the estimation of Eq. (8.29*a*) under one

[5] C. W. J. Granger, "Investigating Causal Relations by Econometric Methods and Cross-Spectral Methods," *Econometrica*, 1969, 37, 424–438.

monetary regime does not necessarily give valid information of how agents will behave under a different regime.[6] If x is super exogenous for β the Lucas critique would not apply; switches of monetary regime would not affect the estimation of Eq. (8.29*a*) nor undermine the validity of forecasts made from it.

8.2.4 Exogeneity Tests

Referring to the model in Eq. (8.29), we see there that weak exogeneity of x_t for β requires $\sigma_{12} = 0$. Thus, we are in the somewhat unfortunate position where the main advantage of weak exogeneity is that one can ignore the marginal distribution, yet the test for valid weak exogeneity requires the modeling of both the marginal and conditional distributions. Engle has developed a general LM test for weak exogeneity.[7] The general procedure tests the joint hypothesis that y_t does not appear in the marginal equation(s) for the conditioning variable(s) and that an appropriate submatrix of the disturbance covariance matrix is zero. In the model of Eq. (8.29) it has been assumed that y_t does not appear in the marginal Eq. (8.29*b*), that there is only one marginal equation, and that hence there is only one element in the relevant submatrix. The null hypothesis is thus $H_0\colon \sigma_{12} = 0$. In this very simple example the LM test becomes equally simple. It is based on the residuals from Eqs. (8.29*a*) and (8.29*b*). Under the null Eq. (8.29*a*) is the conditional equation, distributed independently of the marginal Eq. (8.29*b*). Thus, under the null, each equation may be efficiently estimated by OLS. Let the resultant residuals from Eqs. (8.29*a*) and (8.29*b*) be denoted by e_y and e_x, respectively. For simplicity in writing the model no intercepts have been shown, but in estimating the equations one will include a constant term, unless there is good a priori reason not to do so. The LM test statistic is constructed as follows:

- Regress e_y on a constant, x, and e_x.
- Under H_0, nR^2 from this regression is asymptotically distributed as $\chi^2(1)$.
- Reject H_0 if nR^2 exceeds a preselected critical value.

In this bivariate case an alternative version of the test is to run a regression of e_x on a constant, lagged x, lagged y, and e_y. In finite samples the R^2 will differ in the two regressions, but they are equivalent asymptotically.

Still another version of the test, which involves the calculation of just one set of residuals, is to replace the first regression for the foregoing LM test statistic by a regression of y on a constant, x, and e_x and then to test whether the coefficient of e_x is significantly different from zero. The t test for this coefficient is asymptotically equivalent to the test based on nR^2. If both regressions are estimated, the coefficient of e_x and its estimated standard error will be found to be the same, whether e_y or

[6] R. E. Lucas, Jr., "Econometric Policy Evaluation; A Critique," in Vol. 1 of the Carnegie-Rochester Conferences on Public Policy, supplementary series to the *Journal of Monetary Economics*, eds. K. Brunner and A. Meltzer, North-Holland, 1976, 19–46.

[7] Robert F. Engle, "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics," Chapter 13, *Handbook of Econometrics*, eds. Zvi Griliches and Michael D. Intriligator, North-Holland, 1984.

y is used as the regressand.[8] There is a similar regression that requires only the computation of e_y.

Tests for strong exogeneity require a test for weak exogeneity and a test for Granger causality. The latter test procedure has already been described. Super exogeneity will hold if there is weak exogeneity and the parameters of the conditional distribution ($\boldsymbol{\lambda}_1$) can be shown to be constant, while those of the marginal distribution ($\boldsymbol{\lambda}_2$) are not. The first step is to subject the conditional distribution to rigorous testing for parameter constancy. If that can be established, attention shifts to the marginal distribution. If the parameters of the marginal distribution are also found to be stable, this sample provides no proof of super exogeneity but merely a presumption due to the constancy of $\boldsymbol{\lambda}_1$. Should $\boldsymbol{\lambda}_2$ vary over time, the prescription is to search for dummies or other variables that might model this variation. These variables are then added to the conditional equation. Should they be jointly insignificant, this result is taken as evidence that the parameters of the conditional process are invariant to changes in the parameters of the marginal process.[9] It seems clear that there may be difficulties in deciding when to conclude that the parameters of the marginal distribution are unstable and when to search for variables to model that instability. Presumably in some cases it may be possible to expand the specification of the marginal distribution to include those variables and possibly find the new structure to be stable.

8.2.5 The Wu-Hausman Test[10]

Rewrite the model in Eq. (8.29) in vector form as

$$\boldsymbol{y} = \boldsymbol{x}\beta + \boldsymbol{\epsilon}_1$$

$$\boldsymbol{x} = \boldsymbol{x}_{-1}\alpha_1 + \boldsymbol{y}_{-1}\alpha_2 + \boldsymbol{\epsilon}_2$$

where $\boldsymbol{x}_{-1} = [x_{n-1}\ x_{n-2}\ \cdots\ x_0]'$ and $\boldsymbol{y}_{-1} = [y_{n-1}\ y_{n-2}\ \cdots\ y_0]'$. As we have seen, if $\sigma_{12} \neq 0$, then ϵ_{2t} affects both x_t and ϵ_{1t}. Thus x_t and ϵ_{1t} are correlated in the first equation. In this case x_t satisfies neither the Cowles Commission criterion in Eq. (8.25) for exogeneity nor the condition for being predetermined in Eq. (8.26). Consequently the OLS estimator of β is biased and inconsistent. Instead of deriving a direct test of σ_{12} the Wu-Hausman procedure concentrates on the first equation and tests

$$H_0:\quad \text{plim}\left(\frac{1}{n}\boldsymbol{x}'\boldsymbol{\epsilon}_1\right) = 0$$

against the alternative

[8] See Problem 8.6.

[9] For an illustration of tests for super exogeneity see Neil R. Ericsson et al., op. cit.

[10] J. A. Hausman, "Specification Tests in Econometrics," *Econometrica*, **46**, 1978, 1251–1271; and D. Wu, "Alternative Tests of Independence between Stochastic Regressors and Disturbances," *Econometrica*, **41**, 1973, 733–750. See also Section 10.6.2.

$$H_1: \quad \operatorname{plim}\left(\frac{1}{n}x'\epsilon_1\right) \neq 0$$

The basic idea is to contrast two possible estimators of β. Under H_0 the OLS estimate, $\hat{\beta}_0 = (x'x)^{-1}x'y$, will be consistent and asymptotically efficient. Under H_1 it will be inconsistent. Suppose that we can find an instrument z that satisfies

$$\operatorname{plim}\left(\frac{1}{n}z'x\right) \neq 0 \qquad \text{and} \qquad \operatorname{plim}\left(\frac{1}{n}z'\epsilon_1\right) = 0$$

Then an instrumental variable estimator, $\hat{\beta}_1 = (z'x)^{-1}z'y$, can be constructed. This estimator will be consistent *under both hypotheses*. Under H_0 both estimators are consistent, so the difference between them should vanish asymptotically. Denote the difference by $\hat{q} = \hat{\beta}_1 - \hat{\beta}_0$. Then under H_0

$$\frac{\hat{q}}{\text{s.e.}(\hat{q})} \overset{a}{\sim} N(0, 1) \qquad \text{or} \qquad \frac{\hat{q}^2}{\text{var}(\hat{q})} \overset{a}{\sim} \chi^2(1)$$

Hausman has shown that $\text{var}(\hat{q}) = \text{var}(\hat{\beta}_1) - \text{var}(\hat{\beta}_0)$. Thus the asymptotic test of H_0 is based on

$$\frac{\hat{q}^2}{\text{var}(\hat{\beta}_1) - \text{var}(\hat{\beta}_0)} \overset{a}{\sim} \chi^2(1) \tag{8.37}$$

To facilitate further developments it is helpful to recall the discussion of 2SLS estimators in Chapter 5 and to rewrite the instrumental variable estimator as

$$\hat{\beta}_1 = (\hat{x}'\hat{x})^{-1}\hat{x}'y \qquad \text{where} \qquad \hat{x} = z(z'z)^{-1}z'x = P_z x$$

The first stage consists of the OLS regression of x on z to give the regression values $\hat{x}$. The second stage consists of the OLS regression of y on $\hat{x}$. The residuals from the first-stage regression may be written as $v = x - \hat{x}$. Since v is orthogonal to $\hat{x}$ by construction, it follows that $\hat{x}'\hat{x} = \hat{x}'x$. The difference $\hat{q}$ may now be expressed as

$$\begin{aligned}\hat{q} &= (\hat{x}'\hat{x})^{-1}\hat{x}'y - (x'x)^{-1}x'y \\ &= (\hat{x}'\hat{x})^{-1}[\hat{x}'y - (\hat{x}'\hat{x})(x'x)^{-1}x'y] \\ &= (\hat{x}'\hat{x})^{-1}[\hat{x}'M_x y]\end{aligned} \tag{8.38}$$

where $M_x = I - x(x'x)^{-1}x'$. The last line of Eq. (8.38) indicates that $\hat{q}$ will only go to zero in the limit if $\hat{x}'M_x y$ goes to zero. That expression, in turn, suggests a regression involving $\hat{x}$, x, and y. Consider the regression, sometimes referred to as an artificial or auxiliary regression,

$$y = x\beta + \hat{x}\delta + u$$

The coefficient of $\hat{\delta}$ in this regression is

$$\hat{\delta} = (\hat{x}'M_x\hat{x})^{-1}\hat{x}'M_x y \tag{8.39}$$

with variance given by

$$\text{var}(\hat{\delta}) = \sigma^2(\hat{x}'M_x\hat{x})^{-1} \tag{8.40}$$

The null hypothesis now becomes $H_0\colon \delta = 0$, and an asymptotic test is based on

$$\frac{\delta^2}{\mathrm{var}(\hat{\delta})} \overset{a}{\sim} \chi^2(1) \tag{8.41}$$

It can be shown that the test statistics in Eqs. (8.37) and (8.41) are identical.[11] The choice between them is a matter of computational ease, and in many cases the regression form of the test is simpler. A final point to note is the estimation of σ^2. One estimate may be obtained by substituting $\hat{\beta}_0$ for β in Eq. (8.29a) and another by substituting $\hat{\beta}_1$ for β. A third possibility is to use the residuals from the regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$. All three estimates will be consistent under the null, but will vary in finite samples.

The Wu-Hausman test has been explained in terms of an extremely simple equation. More generally, let the equation be

$$\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{u} \tag{8.42}$$

where X_1 is $n \times k_1$ and X_2 is $n \times k_2$. Suppose it is thought that X_1 may be correlated with $\boldsymbol{u}$, but X_2 is not. The null hypothesis is thus

$$H_0: \quad \mathrm{plim}\left(\frac{1}{n}X_1'\boldsymbol{u}\right) = \boldsymbol{0}$$

Suppose further that there is a matrix of instruments Z_1 of order $n \times l(\geq k_1)$ such that in the limit Z_1 is correlated with X_1 but not with $\boldsymbol{u}$. Define $Z = [Z_1 \; X_2]$ and regress X_1 on Z to obtain $\hat{X}_1 = Z(Z'Z)^{-1}Z'X_1 = P_zX_1$. The null hypothesis is then tested by testing the significance of $\boldsymbol{\delta}$ in the regression

$$\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \hat{X}_1\boldsymbol{\delta} + \boldsymbol{v} \tag{8.43}$$

Strictly speaking, the valid asymptotic test involves a quadratic form in $\hat{\boldsymbol{\delta}}$ and its covariance matrix. Under H_0, $\hat{\boldsymbol{\delta}}'[\mathrm{var}(\hat{\boldsymbol{\delta}})]^{-1}\hat{\boldsymbol{\delta}} \overset{a}{\sim} \chi^2(k_1)$. In practice the conventional F test is often employed. The alternative form of the test could also be derived by finding the IV and OLS estimates of the β parameters in Eq. (8.42); forming $\hat{q}$, the vector of contrasts; and testing the relevant quadratic form in $\hat{q}$. As in the simpler case, the two test statistics are identical, but the regression test is usually the simpler to apply. A more detailed discussion is given in Section 10.6.2.

Once weak exogeneity has been established (or, often in practice, just assumed), ADL equations like Eq. (8.14) are usually estimated by OLS. Provided the x's satisfy the usual stationarity assumptions and the parameters of $A(L)$ satisfy the usual stability assumptions, the standard inference procedures are asymptotically valid. The discussion in Chapter 7, however, has shown that these results may no longer hold in the presence of nonstationary regressors, and we now turn to this topic.

8.3 NONSTATIONARY REGRESSORS

Suppose that in the model represented by Eq. (8.29) the marginal equation is

$$x_t = x_{t-1} + \epsilon_{2t}$$

[11] See Appendix 8.2.

The x variable is now a random walk or, in the language of Chapter 7, integrated of order one, I(1). From Eq. (8.29*a*) the y variable is also I(1). The discussion in Chapter 7 showed that in regressions with nonstationary variables the usual t statistics have **nonstandard distributions,** and consequently the use of the standard tables may give seriously misleading inferences. A related problem is the possibility of finding **spurious regressions.**

The spurious regression problem was laid bare in a classic paper by Yule in 1926.[12] Painstakingly shuffling two decks of playing cards, he drew cards at random to generate series of random numbers. Then on a hand calculator he laboriously computed hundreds of correlation coefficients between (independent) random series and tabulated the resulting distribution. The distribution was approximately symmetrical and unimodal about the "true" value of zero, because he was essentially correlating independent white noise series. Letting ϵ_t represent a white noise series, he then generated x_t from the formula $x_t - x_{t-1} = \epsilon_t$. Thus he now had pairs of *independent* I(1) variables. The distribution of the correlation coefficients between pairs of x series was like an inverted saucer, with a flat top and fairly high densities toward the extremes at $+1$ and -1. Finally he generated y series from $y_t - y_{t-1} = x_t$. Thus, y is an I(2) series. The distribution of the correlation coefficients between pairs of y series was now U-shaped with almost all of the mass concentrated at the extremes. Yule's paper is an early example of a Monte Carlo investigation and is a landmark in the statistical literature. The clear message of the paper is that "statistically significant" correlations may easily be found between independent nonstationary series.

Half a century later, and with the benefit of modern computing power, the issue has been explored further by Granger and Newbold.[13] They generated 100 pairs of independent random walks, that is, independent I(1) variables, and fitted two-variable, linear OLS regressions. The conventional t statistic was calculated to test the significance of each regression. Their main finding was that in 77 of the 100 simulations t had a numerical value in excess of 2, thus leading to incorrect rejection of the null hypothesis of no relationship in more than three-quarters of all cases. They also found very low DW statistics, which tends to cause the conventional formulae to underestimate standard errors and thus overstate t values. However, further simulations involving reestimation with a Cochrane-Orcutt AR(1) correction reduced but did not eliminate the probability of making incorrect inferences.[14] Adding further random walk explanatory variables only increased the percentage of wrong inferences, as is shown in Table 8.1.

It is clear that regressions of independent random walk variables are almost certain to produce incorrect inferences. Phillips has tackled the issue theoretically and has shown that in regressions of independent random walks the regression coefficients do not converge to constants with increasing sample size, as in the standard case. Further, the usual t ratio does not possess a limiting distribution but diverges

[12] G. Udny Yule, "Why Do We Sometimes Get Nonsense Correlations between Time-Series?," *Journal of the Royal Statistical Society,* Series A, **89**, 1926, 1–69.

[13] C. W. J. Granger and P. Newbold, "Spurious Regressions in Econometrics," *Journal of Econometrics,* **2**, 1974, 111–120.

[14] C. W. J. Granger and P. Newbold, *Forecasting Economic Time Series,* Academic Press, 1977, 208–214.

TABLE 8.1
Regressions of random walk variables

Number of explanatory variables	Precentage of times H_0 rejected	Average DW statistic	Average R^2
1	76	0.32	0.26
2	78	0.46	0.34
3	93	0.55	0.46
4	95	0.74	0.55
5	96	0.88	0.59

Source: Granger and Newbold, "Spurious Regressions in Econometrics," *Journal of Econometrics,* **2**, 1974, 116.

with increasing sample size, thus increasing the probability of incorrect inferences as the sample size increases.[15]

One should not overreact to these seemingly alarming results. First, low DW statistics are now usually taken as indications of seriously misspecified relations, so that one abandons the relation and works on a respecification. Second, these results are all derived from regressions involving only *current* variables. If y and x are independent random walks and y_t is regressed on x_t, then a double specification error has been committed; a relevant variable y_{t-1} has been excluded and an irrelevant variable x_t has been included. Finally, as the discussion in Section 8.2 showed, conventional formulations of ADL relationships, especially in the general to simple approach, involve many lagged terms, so possible nightmares from correlating just current variables are less likely to arise. However, an important real problem still remains. What inference procedures can be used if some or all of the variables in an ADL specification are nonstationary?

The problem has been addressed in an important article by Sims, Stock, and Watson.[16] The technical level of their paper is beyond that assumed in this book, so only the barest summary is given here.[17] Basically the presence of I(1) variables implies that most, if not all, test statistics will have nonstandard distributions. Thus, one cannot automatically refer conventional test statistics to the standard t, $N(0, 1)$, F, or χ^2 tables. However, there are exceptions to this general rule. If, in some reparameterization, a parameter can be expressed as the coefficient of a mean-zero, I(0) variable, the conventional test statistics on that coefficient are asymptotically valid. Further, if in a reparameterization a subset of parameters can be expressed as coefficients of mean-zero, I(0) variables, then conventional tests on that subset are asymptotically valid. We will illustrate with the ADL(1,1) model defined in Eq. (8.1), namely,

$$y_t = m + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \epsilon_t \tag{8.1}$$

[15] P. C. B. Phillips, "Understanding Spurious Regressions in Econometrics," *Journal of Econometrics,* **33**, 1986, 311–340.

[16] C. A. Sims, J. H. Stock, and M. W. Watson, "Inference in Linear Time Series Models with Some Unit Roots," *Econometrica,* **58**, 1990, 113–144.

[17] For a comprehensive summary see A. Banerjee, J. J. Dolado, J. W. Galbraith, and D. F. Hendry, *Co-Integration, Error-Correction, and the Econometric Analysis of Non-Stationary Data,* Oxford University Press, 1993, Chapter 6, and especially the example on pages 188–189.

As we have already seen, y_t can then be expressed as an (infinite) sum of current and lagged x's and current and lagged ϵ's. Suppose that x follows a random walk, $x_t = x_{t-1} + \eta_t$, and is thus I(1). The ϵ's have been assumed to be white noise and thus I(0), and so y_t is a combination of I(1) and I(0) variables and is itself I(1), because in general linear combinations of I(1) variables are also I(1).

As a simple illustration of this rule suppose that (a different) y is defined as an arbitrary linear combination of two I(1) variables,

$$y_t = ax_t + bz_t$$
$$x_t = x_{t-1} + u_t$$
$$z_t = z_{t-1} + v_t$$

where a and b are arbitrary constants and u and v, white noise disturbances. The process of first differencing the first equation and substituting from the second and third gives

$$y_t = y_{t-1} + (au_t + bv_t)$$

Thus, y is also a random walk and I(1).

All the variables in Eq. (8.1) are thus I(1), and it might seem that nonstandard inference procedures must be found. However, this conclusion is premature because we have not examined possible reparameterizations. One such reparameterization is given in Eq. (8.6), namely,

$$\Delta y_t = \beta_0 \Delta x_t - (1 - \alpha_1)[y_{t-1} - a - \gamma x_{t-1}] + \epsilon_t \tag{8.6}$$

where $\quad a = \dfrac{m}{1-\alpha_1} \quad$ and $\quad \gamma = \dfrac{\beta_0 + \beta_1}{1 - \alpha_1}$

In Eq. (8.6) Δy_t and Δx_t are each I(0), under the assumption that x is I(1), but what about the term in square brackets? In the discussion of Eq. (8.6) it was described as the deviation of y_{t-1} from the static equilibrium value corresponding to x_{t-1}. A static equilibrium value for y is found by holding x constant over time. That concept is not very meaningful in the current context, where x is presumed to follow a random walk. The deviation, however, has a very important role to play in the current context. Let us define it as

$$z_t = y_t - \frac{m}{1-\alpha_1} - \frac{\beta_0 + \beta_1}{1-\alpha_1} x_t \tag{8.44}$$

Now subtract $[m/(1 - \alpha_1) + (\beta_0 + \beta_1)x_t/(1 - \alpha_1)]$ from both sides of Eq. (8.1). The result is

$$\begin{aligned} z_t &= \frac{-\alpha_1 m}{1-\alpha_1} + \alpha_1 y_{t-1} - \frac{\alpha_1\beta_0 + \beta_1}{1-\alpha_1} x_t + \beta_1 x_{t-1} + \epsilon_t \\ &= \alpha_1 z_{t-1} + v_t \end{aligned} \tag{8.45}$$

where $\quad v_t = \epsilon_t - \dfrac{\alpha_1\beta_0 + \beta_1}{1-\alpha_1}\eta_t$

and η_t is the disturbance in the random walk for x. Provided the stationarity condition $|\alpha_1| < 1$ is satisfied, z is a stationary AR(1) process and thus I(0). As shown in Eq. (8.44), z_t is a linear combination of two I(1) variables, y_t and x_t, but this linear combination is I(0). In this case y_t and x_t are said to be **cointegrated:** This linear combination of the two I(1) variables is a mean-zero I(0) variable. As seen in Chapters 2 and 7, I(1) variables display a tendency to "wander." When two I(1) variables are cointegrated, however, they will tend to "wander together." The zero mean and constant variance of z_t prevent them from drifting too far apart. Equation (8.44) is then called a **cointegrating relation.** Equation (8.6) may be rewritten as

$$\Delta y_t = \beta_0 \Delta x_t - (1 - \alpha_1)z_{t-1} + \epsilon_t \tag{8.46}$$

All three variables in this equation are mean-zero I(0), and so inference about β_0 and α_1, separately or jointly, can proceed in standard fashion. The equation is also said to be **balanced** because all variables are integrated to the same degree. The remaining parameter in Eq. (8.1) is β_1. It too can be shown as the coefficient of a mean-zero, I(0) variable in another reparameterization of Eq. (8.1), namely,

$$\Delta y_t = m - (1 - \alpha_1)y_{t-1} + (\beta_0 + \beta_1)x_t - \beta_1 \Delta x_t + \epsilon_t \tag{8.47}$$

The t statistic on β_1 is asymptotically N(0, 1) since Δx_t is a mean-zero, I(0) variable. This regression might seem unbalanced because it contains two I(0) variables and two I(1) variables. However, the I(1) variables are cointegrated, even though the time subscripts are not an exact match. Thus the presence of *two* I(1) variables allows the possibility of a linear combination being I(0), giving the same order of integration on each side of the equation.

Returning to Eq. (8.1),

$$y_t = m + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \epsilon_t \tag{8.1}$$

we find that the reparameterizations in Eqs. (8.46) and (8.47) show that all three slope parameters can appear as coefficients of mean-zero, I(0) variables and thus have t statistics that are asymptotically N(0, 1). The crucial point is that for estimation purposes **none of the reparameterizations need be actually carried out.** The parameters can be estimated and tested by applying OLS to Eq. (8.1). As shown in Appendix 8.1 identical estimates and test statistics result from the specification in Eq. (8.1) or any nonsingular linear transformation. Despite the standard results for the parameters individually, joint tests involving all three parameters are nonstandard, since there is no single reparameterization that shows all three as coefficients of mean-zero, I(0) variables.

Estimation and testing of the cointegrating equation

On continuing with Eq. (8.1) two questions arise. First, how should the cointegrating relation be estimated? Second, how should one test if there really is a cointegrating relation? On estimation, one possibility is to estimate the ADL Eq. (8.1) and then compute the parameters of the relation shown in Eq. (8.44) from the parameters of the ADL equation. The first suggestion in the literature, however, was simply and surprisingly to fit a linear regression of y_t on a constant and x_t.[18] One

[18] R. F. Engle and C. W. J. Granger, "Cointegration, Error Correction: Representation, Estimation, and Testing," *Econometrica*, **55**, 1987, 251–276.

says *surprising* because it is clear from Eqs. (8.44) and (8.45) that the disturbance z_t in such a regression will be correlated with the regressor, which conventionally suggests inconsistency. However, the conventional argument no longer goes through when the regressor is I(1). Applying OLS to Eq. (8.44) gives

$$\hat{\gamma} = \gamma + \frac{\sum(x_t - \bar{x})z_t}{\sum(x_t - \bar{x})^2}$$

Consistency requires $$\frac{\text{plim}\left(\frac{1}{n}\sum(x_t - \bar{x})z_t\right)}{\text{plim}\left(\frac{1}{n}\sum(x_t - \bar{x})^2\right)} = 0$$

When x is stationary and correlated with z, the numerator and denominator are each nonzero constants and so the condition fails. However, when x is I(1) its variance increases without limit with the sample size. The covariance in the numerator still tends to a finite constant, and so the condition holds. Moreover, the rate at which the OLS estimate approaches the population parameter is faster than in the conventional stationary case. Thus $\hat{\gamma}$ is said to be a **superconsistent** estimator of γ.[19] The estimated relation could then be used to obtain the OLS residuals, $\hat{z}_t$. These in turn could be substituted for z_t in Eq. (8.46) and OLS used to estimate α_1 and β_0.

Asymptotic results often provide but cold comfort for practical econometricians, who perforce live in a finite-sample world. There is evidence that superconsistent estimates may have substantial finite-sample biases and that estimating the ADL relation and then solving for the parameters of the cointegrating relation may give less biased estimates of these parameters.[20] The comparison, however, is hardly fair, since the latter approach assumes some knowledge of the ADL relation, whereas the Engle-Granger regression of y_t on x_t requires no such information.

A defining feature of a cointegrating relation is that the error in such a relation should be I(0). Thus it is desirable to test whether the equilibrium error process has a unit root. Unit root tests were described in Chapter 7 and some appropriate nonstandard critical values presented. Those critical values cannot be used for the present purpose, for they were applicable to the *actual* values of the process being tested, and here we only have the *estimated* values, $\hat{z}_t$. Relevant critical values for cointegrating tests are available from a comprehensive Monte Carlo simulation by MacKinnon.[21] From Eq. (8.45) the test regression is

$$\Delta\hat{z}_t = (\alpha_1 - 1)\hat{z}_{t-1} + v_t \tag{8.48}$$

The t statistic given by the ratio of $(\hat{\alpha}_1 - 1)$ to its standard error is then referred to the MacKinnon critical values. The null hypothesis of no cointegration is H_0: $\alpha_1 - 1 = 0$. Significant negative values would lead to rejection of the null. Table 1 in the

[19] James H. Stock, "Asymptotic Properties of Least Squares Estimators of Cointegrating Vectors," *Econometrica*, **55**, 1987, 1035–1056.

[20] See Banerjee et al, op. cit., Chapter 7.

[21] James G. MacKinnon, "Critical Values for Cointegration Tests," Chapter 13, *Long-Run Economic Relationships*, eds. R. F. Engle and C. W. J. Granger, Oxford University Press, 1991.

TABLE 8.2
Asymptotic critical values for cointegration tests

k*	Test statistic†	1%	5%	10%
2	*c*	−3.90	−3.34	−3.04
	ct	−4.32	−3.78	−3.50
3	*c*	−4.29	−3.74	−3.45
	ct	−4.66	−4.12	−3.84
4	*c*	−4.64	−4.10	−3.81
	ct	−4.97	−4.43	−4.15
5	*c*	−4.96	−4.42	−4.13
	ct	−5.25	−4.72	−4.43
6	*c*	−5.25	−4.71	−4.42
	ct	−5.52	−4.98	−4.70

Reprinted by permission from Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics*, Oxford University Press, 1993, 722.

*The heading k indicates the number of variables in the estimated cointegrating equation.

†The letters *c* and *ct* indicate whether that equation contains a constant or a constant plus a linear time trend.

MacKinnon reference permits the calculation of critical values for different sample sizes, different significance levels, and different numbers of variables in the cointegrating equation. Our exposition so far has involved just two variables in the ADL relation, and hence just two variables in the cointegrating equation. In practice one is normally estimating a multivariate cointegrating equation. The MacKinnon test assumes that an intercept term has been used in the cointegrating equation, and it also allows for the possibility of a linear time trend also having been included. Table 8.2 gives the asymptotic critical values.

When one moves beyond the two-variable case the possibility immediately arises of there being more than one cointegrating relation. Suppose there are $k(> 2)$ variables, all of which are I(1). A cointegrating relation is a linear combination of these variables that is I(0). Clearly there may be $0, 1, 2, \ldots, k - 1$ cointegrating relations. Clearly also there will be problems with the Engle-Granger regression approach. Do we obtain different cointegrating relations simply by changing the direction of error minimization in a k-variable regression? Chapter 9 will present a more general approach to both testing for the number of cointegrating relations and estimating the parameters of these relations.

8.4
A NUMERICAL EXAMPLE

We return to the data on personal gasoline consumption, which has been briefly considered in Chapters 1 and 4. The variables are as follows:

$$Y = \text{log of real per capita expenditure on gasoline}$$

$$X2 = \text{log of the real price of gasoline}$$

$$X3 = \text{log of real per capita disposable personal income}$$

The data are quarterly, seasonally adjusted series and cover the period 1959.1 to 1990.4. Figure 8.1 shows the series over the full sample period. The middle panel shows the dramatic changes in the real price of gasoline with a rise of 29 percent from 1973.3 to 1974.2, an even greater rise of 60 percent from 1978.3 to 1980.2, a very substantial decline in the first half of the 1980s, and a rebound in the later 1980s. Real expenditure rose steadily throughout the 1960s and early 1970s and then declined without ever regaining its earlier peak. These series obviously present a formidable empirical challenge to any demand analyst.

8.4.1 Stationarity

First one looks at the stationarity of the series. Visual inspection suggests that income is nonstationary. Price and expenditure each display structural breaks associated with the occurrence of the oil price shocks. The conventional test, assuming an intercept and four lags, does not reject the unit root hypothesis in any of the three cases, as is shown in Table 8.3.[22] Perron has argued that structural breaks invalidate conventional unit root tests.[23] He has developed a test procedure that allows for one known structural break consisting of a change in level and/or a change in growth rate and has also provided relevant critical values. However, application of the Perron procedure to expenditure and price does not reject the unit root hypothesis.[24] The hypothesis of a unit root in the first differences is rejected for all three series, so we conclude that they are all I(1). Next we look for the possible existence of a cointegrating relationship.

8.4.2 Cointegration

Estimation of the Engle-Granger cointegrating relationship gives the results shown in Table 8.4. These imply a long-run price elasticity of -0.15 and a long-run income elasticity of 0.71. However, one must test whether this represents a cointegrating relation. There are two important cointegrating tests. The first is to test the residuals from this relation for stationarity. The second is to fit a general ADL specification to these three variables and see whether one can solve for a meaningful long-run relationship.

The residuals from the regression in Table 8.4 are shown in Fig. 8.2. This series is clearly nonstationary: there is a dramatic inverse spike at 1974.1. Applying the regression in Eq. (8.48) to these residuals gives an ADF statistic of -1.34, which does not reject the unit root hypothesis. There is thus strong evidence that we do not have a cointegrating relation.

[22] In this and subsequent tables the variables are printed in uppercase form. They do, however, correspond to the variables in lowercase type in the text.

[23] Pierre Perron, "The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis," *Econometrica*, **57**, 1989, 1361–1401.

[24] See Problem 8.8.

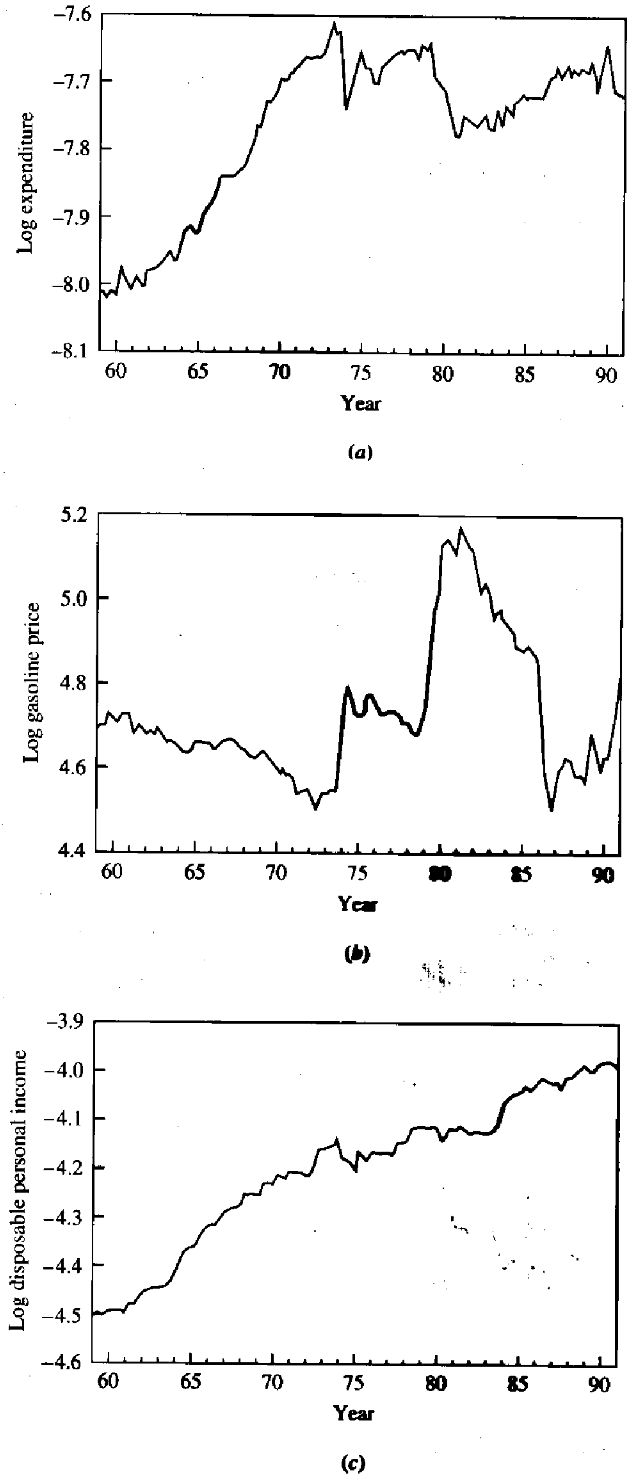


FIGURE 8.1
Personal gasoline consumption in the United States, 1959.1–1990.4: (*a*) Real per capita expenditure on gasoline; (*b*) real price of gasoline; (*c*) real per capita disposable income.

TABLE 8.3
ADF values for Y, X2, and X3

Y	X2	X3
−2.45	−1.79	−1.94

The 1 percent, 5 percent, and 10 percent MacKinnon critical values, obtained from the EViews output, are −3.48, −2.88, and −2.58, respectively.

TABLE 8.4
A cointegrating relation?

LS // Dependent Variable is Y
Sample: 1959:1–1990:4
Included observations: 128

Variable	Coefficient	Std. Error	T-Statistic	Prob.
X2	−0.150285	0.031231	−4.812013	0.0000
X3	0.705614	0.033735	20.91648	0.0000
C	−4.093340	0.224756	−18.21239	0.0000

R-squared	0.777862	Mean dependent var	−7.763027
Adjusted R-squared	0.774308	S.D. dependent var	0.120187
S.E. of regression	0.057097	Akaike info criterion	−5.702846
Sum squared resid	0.407510	Schwarz criterion	−5.636002
Log likelihood	186.3580	F-statistic	218.8572
Durbin-Watson stat	0.084511	Prob(F-statistic)	0.000000

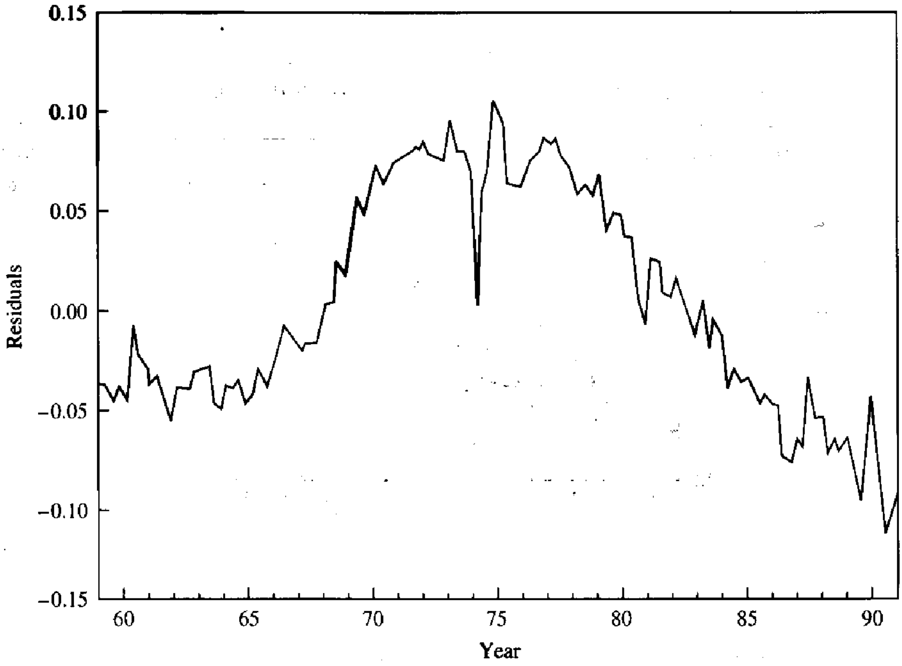


FIGURE 8.2
Residuals from regression in Table 8.4.

This conclusion is confirmed by fitting an ADL model to the data. Results are shown in Table 8.5. The ADL relation may be written as

$$A(L)y_t = m + B_2(L)x_{2t} + B_3(L)x_{3t} + u_t$$

By holding the variables constant, the implied long-run relation is

$$\bar{y} = \frac{m}{A(1)} + \frac{B_2(1)}{A(1)}\bar{x}_2 + \frac{B_3(1)}{A(1)}\bar{x}_3 \tag{8.49}$$

where replacing L by 1 in a lag polynomial gives the sum of the coefficients in that polynomial. Clearly, estimates of the coefficients in Eq. (8.49) can be obtained from the estimated ADL relation, but the exercise is meaningless if these three sums are not significantly different from zero. Thus we can test for cointegration by testing whether $A(1)$, $B_2(1)$, and $B_3(1)$ are zero. Testing that $A(1)$ is zero is equivalent to testing that the sum of the coefficients on the lagged Y terms is equal to 1. The actual

TABLE 8.5

An ADL model of expenditure, price, and income

LS // Dependent Variable is Y
Sample: 1960:2–1990:4
Included observations: 123
Excluded observations: 0 after adjusting endpoints

Variable	Coefficient	Std. Error	T-Statistic	Prob.
X2	−0.267647	0.037874	−7.066703	0.0000
X2(−1)	0.262873	0.069291	3.793737	0.0002
X2(−2)	−0.017411	0.075414	−0.230867	0.8179
X2(−3)	−0.072094	0.077389	−0.931585	0.3537
X2(−4)	0.014402	0.077210	0.186524	0.8524
X2(−5)	0.058206	0.046340	1.256058	0.2119
X3	0.292764	0.158824	1.843326	0.0681
X3(−1)	−0.162176	0.220228	−0.736400	0.4631
X3(−2)	−0.049270	0.214372	−0.229835	0.8187
X3(−3)	0.010409	0.213133	0.048838	0.9611
X3(−4)	0.084917	0.210132	0.404110	0.6870
X3(−5)	−0.198967	0.153118	−1.299434	0.1966
Y(−1)	0.660572	0.096063	6.876439	0.0000
Y(−2)	0.067018	0.114535	0.585131	0.5597
Y(−3)	−0.023578	0.117094	−0.201359	0.8408
Y(−4)	0.132194	0.119013	1.110747	0.2692
Y(−5)	0.163124	0.101384	1.608975	0.1106
C	0.005543	0.126043	0.043975	0.9650

R-squared	0.984158	Mean dependent var	−7.752780
Adjusted R-squared	0.981593	S.D. dependent var	0.111024
S.E. of regression	0.015063	Akaike info criterion	−8.256599
Sum squared resid	0.023823	Schwarz criterion	−7.845060
Log likelihood	351.2514	F-statistic	383.7051
Durbin-Watson stat	1.922386	Prob(F-statistic)	0.000000

sum is 0.999, and the P value for the null hypothesis is 0.98. Thus, $A(1)$ is effectively zero and the long-term relation breaks down. The other two sums both turn out to be -0.022, with P values of 0.04 and 0.31. There is thus no evidence of a cointegrating relation between these three variables.[25]

8.4.3 A Respecified Relationship

The foregoing "simpleminded" approach has paid no attention to the special characteristics of the market for gasoline. First, consumption is mediated through appropriate equipment (cars). Dramatic price increases set in motion lengthy and expensive changes in the type of new equipment produced and also led the federal government to set fuel efficiency targets for car fleets in future years. Second, gasoline is desired not for its own sake but rather as a means of producing "mileage." To model a demand function for miles, let us define the following variables:

$$\begin{aligned} Y &= \text{real per capita expenditure on gasoline ("gallons")} \\ X_2 &= \text{real price of a gallon} \\ X_3 &= \text{real per capita income} \\ X_4 &= \text{miles per gallon} \\ M &= \text{miles per capita} = Y \cdot X_4 \\ \text{RPM} &= \text{real price of a mile} = X_2/X_4 \end{aligned}$$

A demand function for miles might be formulated as

$$M = K(\text{RPM})^{\beta_2}(X_3)^{\beta_3} = K\left(\frac{X_2}{X_4}\right)^{\beta_2} X_3^{\beta_3}$$

The implied demand function for "gallons" is then

$$Y = KX_2^{\beta_2}X_3^{\beta_3}X_4^{-(1+\beta_2)} \tag{8.50}$$

Converting to logs then adds a new variable, x_4 = log of miles per gallon, to the previous specification.

Annual data on travel and fuel consumption are available from the U.S. Department of Transportation, Federal Highway Administration, from which a miles per gallon (mpg) series may be computed. We have taken mpg for "all passenger vehicles." There is also a series for "passenger cars." The movements of the two series are practically identical, since cars consume 99 percent of the gasoline used by the combined group of cars, motorcycles, and buses. From 1959 to 1973 the mpg figure declined by 6 percent, accompanied by a slowly declining real price and rising

[25] Since the variables are nonstationary and there does not appear to be a cointegrating relation, the stated P values are suspect. However, the estimated sums are all practically zero, and it is unlikely that nonstandard distributions would change the conclusion in the text.

TABLE 8.6
A cointegrating relation

LS // Dependent Variable is Y
Sample: 1959:1–1990:4
Included observations: 128

Variable	Coefficient	Std. Error	T-Statistic	Prob.
X2	−0.138561	0.010985	−12.61399	0.0000
X3	0.998547	0.015403	64.82624	0.0000
X4	−0.518128	0.017390	−29.79491	0.0000
C	−1.514535	0.117185	−12.92429	0.0000

R-squared	0.972774	Mean dependent var	−7.763027
Adjusted R-squared	0.972116	S.D. dependent var	0.120187
S.E. of regression	0.020069	Akaike info criterion	−7.786363
Sum squared resid	0.049945	Schwarz criterion	−7.697238
Log likelihood	350.7031	F-statistic	1476.851
Durbin-Watson stat	0.741016	Prob(F-statistic)	0.000000

income. For obvious reasons the series was slow to rise after the price shocks of the 1970s. It rose by under 6 percent from 1973 to 1978 and by a further 21 percent between 1978 and 1983. It continued rising throughout the price declines of the 1980s, and by 1990 mpg was almost 50 percent greater than in 1959. We have converted the annual data to quarterly form by linear interpolation and taken logs to produce the x_4 series for incorporation in the statistical analysis.

Looking first for a possible cointegrating relation gives the results in Table 8.6. Figure 8.3 shows the actual and fitted series and the residuals. These residuals look fairly stationary in mean levels compared with those in Fig. 8.2, but there is still a pronounced inverse spike at 1974.1. Regressing the first difference of the residuals on the lagged residual gives an ADF statistic of −5.39. The asymptotic 1 percent critical value from Table 8.2 for $k = 4$ is −4.64. The 1 percent critical value from MacKinnon's table for this sample size is −4.78. Thus, the hypothesis of a unit root in the residuals is rejected, and by contrast with Table 8.4 we may have a cointegrating relation.

Inserting a dummy variable that takes on the value of 1 in 1974.1 and 0 elsewhere gives the regression output shown in Fig. 8.4. The spike in the residuals has been removed. The ADF statistic from these residuals is −5.11, so the hypothesis of a unit root in the residuals is still rejected. The cointegrating elasticities in this regression are identical to two decimal places with those given in Table 8.6.

8.4.4 A General ADL Relation

In fitting a general ADL relation to this expanded data set there is a question of whether to include a dummy variable for the 1974.1 spike. We have chosen to include it and leave it as an exercise for the reader to carry out the same analysis excluding the dummy variable. The result, using lags up to the fifth quarter, is given in

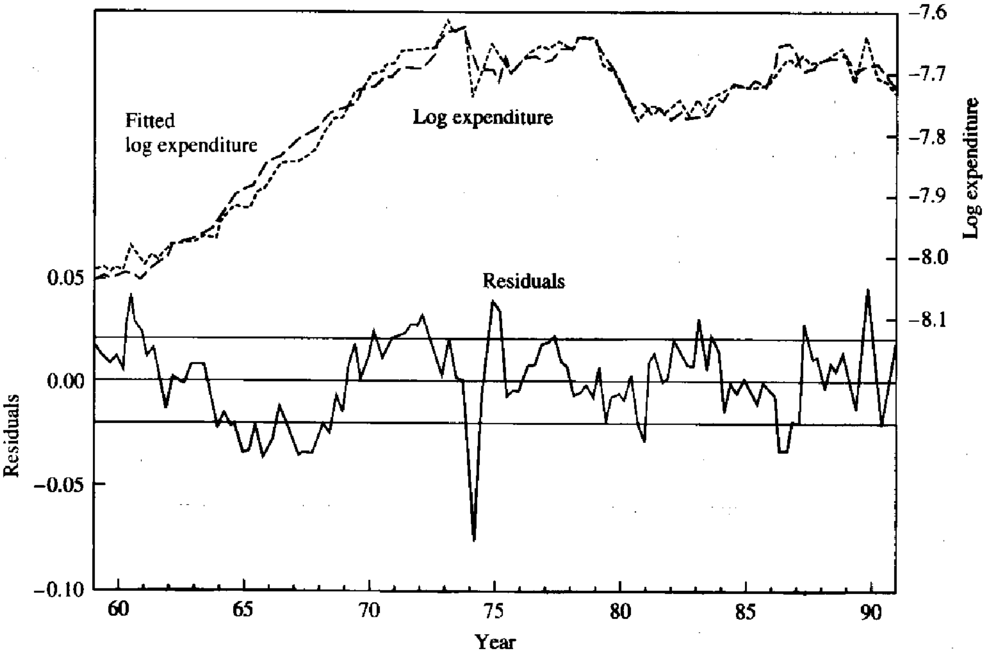


FIGURE 8.3
Regression output from Table 8.6.

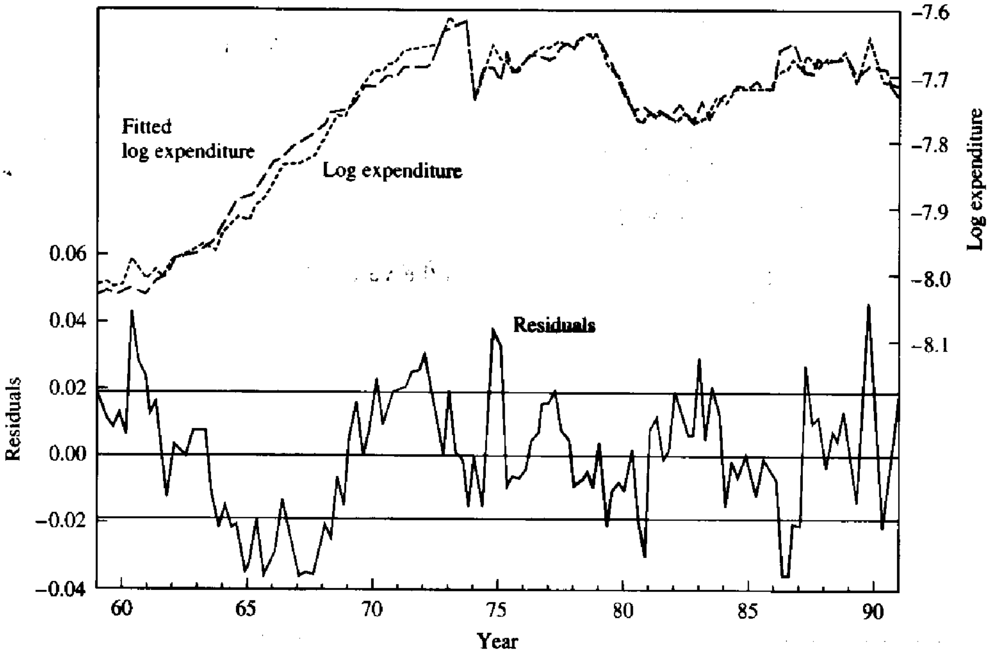


FIGURE 8.4
A cointegrating relation with a dummy variable.

TABLE 8.7

An ADL model of expenditure, price, income, and mpg

LS // Dependent Variable is Y
Sample: 1960:2–1990:4
Included observations: 123 after adjusting endpoints

Variable	Coefficient	Std. error	T-statistic	Prob.
X2	−0.198597	0.032181	−6.171346	0.0000
X2(−1)	0.189293	0.054372	3.481461	0.0007
X2(−2)	−0.016542	0.058558	−0.282483	0.7782
X2(−3)	−0.115526	0.060759	−1.901377	0.0602
X2(−4)	0.066485	0.061770	1.076328	0.2844
X2(−5)	0.029012	0.036749	0.789468	0.4317
X3	0.164873	0.129176	1.276347	0.2048
X3(−1)	0.145563	0.181069	0.803907	0.4234
X3(−2)	−0.169946	0.173293	−0.980687	0.3292
X3(−3)	0.066797	0.167245	0.399397	0.6905
X3(−4)	0.045524	0.162186	0.280691	0.7795
X3(−5)	−0.004198	0.122806	−0.034182	0.9728
X4	−1.557670	0.585739	−2.659324	0.0091
X4(−1)	2.697054	1.241118	2.173085	0.0322
X4(−2)	−2.278796	1.381349	−1.649689	0.1022
X4(−3)	1.170965	1.372563	0.853123	0.3957
X4(−4)	0.204391	1.256976	0.162606	0.8712
X4(−5)	−0.375389	0.621491	−0.604013	0.5472
Y(−1)	0.581024	0.082101	7.076947	0.0000
Y(−2)	0.014630	0.091902	0.159190	0.8738
Y(−3)	−0.166262	0.094471	−1.759933	0.0815
Y(−4)	0.297403	0.098549	3.017817	0.0032
Y(−5)	0.023884	0.083523	0.285956	0.7755
DUM	−0.093403	0.012909	−7.235528	0.0000
C	−0.290653	0.129715	−2.240701	0.0273

R-squared	0.991490	Mean dependent var	−7.752780
Adjusted R-squared	0.989406	S.D. dependent var	0.111024
S.E. of regression	0.011427	Akaike info criterion	−8.764246
Sum squared resid	0.012797	Schwarz criterion	−8.192664
Log likelihood	389.4717	F-statistic	475.7689
Durbin-Watson stat	1.872773	Prob(F-statistic)	0.000000

Table 8.7. The sums of coefficients are

$$\begin{aligned} A(1) &= 0.2493 & B_2(1) &= -0.0459 \\ B_3(1) &= 0.2486 & B_4(1) &= -0.1394 \end{aligned} \tag{8.51}$$

The P values from testing that these sums are zero are, respectively, 0.001, 0.011, 0.001, and 0.002. The distributions are nonstandard because not all the coefficients in any group can be expressed simultaneously as coefficients of zero mean, stationary variables. Nonetheless, the conventional P values are so small that rejection of the null hypotheses would seem reasonable. The implied long-run relationship is then

$$\bar{y} = -1.17 - 0.18\bar{x}_2 + 1.00\bar{x}_3 - 0.56\bar{x}_4 \qquad (8.52)$$

The elasticities in Eq. (8.52) correspond very closely to those in Table 8.6.

Next the relation in Table 8.7 is subjected to various tests. Figure 8.5 gives the actual and fitted series and the regression residuals, which are obviously an improvement on those from the cointegrating relation in Fig. 8.4. The Jarque-Bera statistic for testing the normality of the residuals is 3.33, with a *P* value of 0.19, so the normality assumption is not rejected. The Breusch-Godfrey asymptotic test for serial correlation up to the fourth order gives a *P* value of 0.40, so the hypothesis of zero autocorrelation in the residuals is not rejected. Tests for ARCH residuals with one up to five lags give *P* values between 0.64 and 0.91, so the assumption of homoscedastic residuals is not rejected in favor of ARCH residuals. The White heteroscedasticity test has a *P* value of 0.08, which points toward heteroscedasticity, but not decisively so. The Ramsey RESET test for specification error has a *P* value of 0.32, so there is no significant evidence of misspecification. The Chow forecast test for the four quarters of 1990 has a *P* value of 0.09, but extending the forecast to two years gives a significant rejection of constant parameters. Overall the relation has survived a formidable battery of tests with only the Chow test suggesting weakness. Reestimating the relation and omitting the dummy variable for 1974.1, whose inclusion sets the residual for that quarter to zero, give a somewhat larger standard error of regression and successful Chow tests over several years. However, we will stick with the relation in Table 8.7 as an acceptable first formulation of the ADL relation and search for acceptable simplifications.

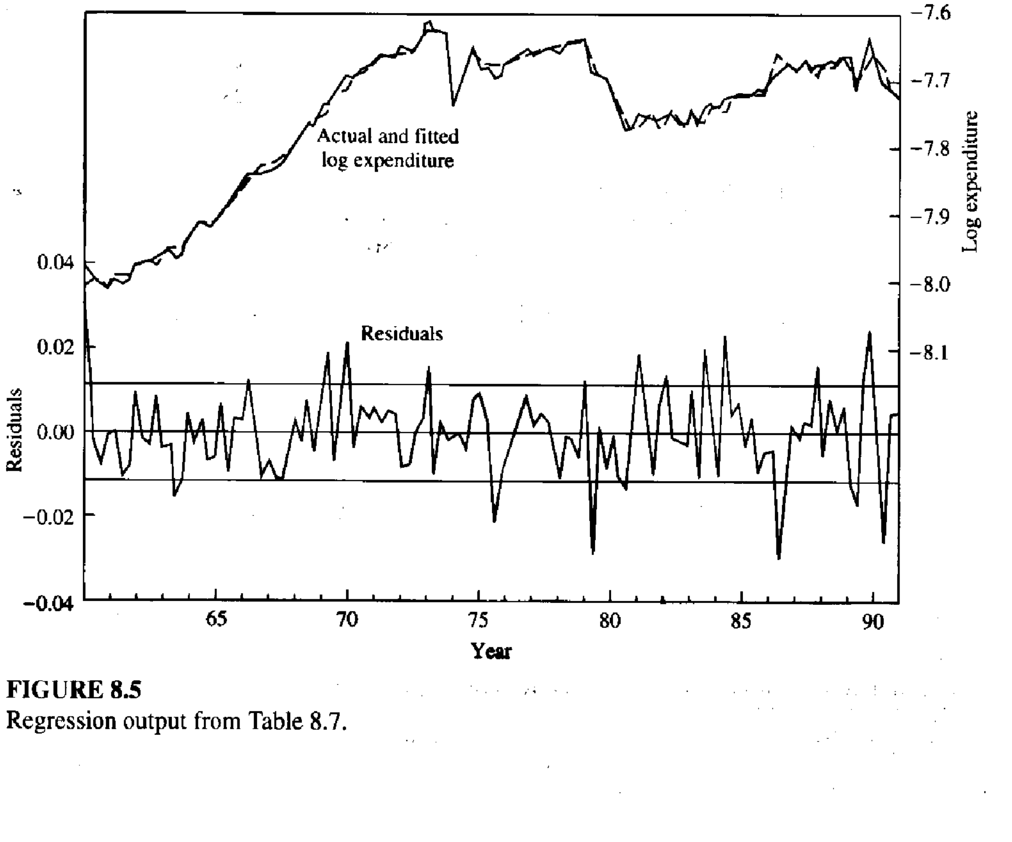


FIGURE 8.5
Regression output from Table 8.7.

8.4.5 A Reparameterization

The first step in the search is a reparameterization. If we ignore the dummy variable, the relation in Table 8.7 is

$$A(L)y_t = m + B_2(L)x_{2t} + B_3(L)x_{3t} + B_4(L)x_{4t} + u_t \tag{8.53}$$

where each polynomial in the lag operator is of the fifth order. The relation in Eq. (8.53) is solely in terms of the **levels** of the variables. For reasons to be explained, we wish to reparameterize in terms of both **levels** and **first differences.** Consider

$$\begin{aligned} B(L) &= \beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \beta_4 L^4 + \beta_5 L^5 \\ &= B(1)L + (1-L)(\delta_0 + \delta_1 L + \delta_2 L^2 + \delta_3 L^3 + \delta_4 L^4) \end{aligned} \tag{8.54}$$

In the second line of Eq. (8.54) the sum of the β coefficients is the coefficient of L, and the δ's are the coefficients of first difference terms from Δx_t to Δx_{t-4}. The process of multiplying out and equating coefficients of the powers of L gives the connections between the δ's and β's.[26] Applying Eq. (8.54) to a variable x_t gives

$$B(L)x_t = B(1)x_{t-1} + \delta_0\Delta x_t + \delta_1\Delta x_{t-1} + \delta_2\Delta x_{t-2} + \delta_3\Delta x_{t-3} + \delta_4\Delta x_{t-4}$$

This transformation is used for all the lag polynomials on the right-hand side of Eq. (8.53). A similar transformation is used on the regressand, namely,

$$\begin{aligned} A(L) &= 1 - \alpha_1 L - \alpha_2 L^2 - \alpha_3 L^3 - \alpha_4 L^4 - \alpha_5 L^5 \\ &= A(1)L + (1-L)(1 + \gamma_1 L + \gamma_2 L^2 + \gamma_3 L^3 + \gamma_4 L^4) \end{aligned}$$

which gives

$$A(L)y_t = \Delta y_t + [A(1)y_{t-1} + \gamma_1\Delta y_{t-1} + \gamma_2\Delta y_{t-2} + \gamma_3\Delta y_{t-3} + \gamma_4\Delta y_{t-4}]$$

These transformations give the model estimated in Table 8.8. There are several important points to notice about this table:

1. The standard error of regression is identical with its value in Table 8.7, as are the values of the log-likelihood, the Durbin-Watson statistic, and the information criteria. This result follows directly from the results on reparameterization derived in Appendix 8.1.
2. The coefficients of the lagged x values are the sums already given from Table 8.7 in Eq. (8.51). The sum of the coefficients on the lagged y values in Table 8.7 is 0.7507. Thus, $A(1) = 1 - 0.7507 = 0.2493$, which is the negative of the coefficient of the $Y(-1)$ regressor in Table 8.8. Furthermore, the P values attached to the lagged levels are exactly the P values for testing that these sums are zero, already given following Eq. (8.51). Thus, one advantage of the reparameterization is the direct estimation and testing of the sums that are relevant to the existence of a cointegrating relation.
3. Switching to first differences usually gives a substantial reduction in the collinearity of the regressors, thus reducing standard errors. It also facilitates the identification of possible simplifications of the relationship.

[26] See Problem 8.9.

TABLE 8.8
A reparameterized model

LS // Dependent Variable is DY
Sample: 1960:2–1990:4
Included observations: 123 after adjusting endpoints

Variable	Coefficient	Std. error	T-statistic	Prob.
X2(−1)	−0.045875	0.017792	−2.578445	0.0114
DX2	−0.198597	0.032181	−6.171346	0.0000
DX2(−1)	0.036571	0.037397	0.977915	0.3305
DX2(−2)	0.020029	0.037199	0.538422	0.5915
DX2(−3)	−0.095497	0.039033	−2.446545	0.0162
DX2(−4)	−0.029012	0.036749	−0.789468	0.4317
X3(−1)	0.248613	0.075099	3.310462	0.0013
DX3	0.164873	0.129176	1.276347	0.2048
DX3(−1)	0.061823	0.134400	0.459991	0.6465
DX3(−2)	−0.108124	0.131237	−0.823882	0.4120
DX3(−3)	−0.041326	0.125801	−0.328506	0.7432
DX3(−4)	0.004198	0.122806	0.034182	0.9728
X4(−1)	−0.139445	0.044512	−3.132745	0.0023
DX4	−1.557670	0.585739	−2.659324	0.0091
DX4(−1)	1.278829	0.782296	1.634712	0.1053
DX4(−2)	−0.999967	0.787947	−1.269080	0.2074
DX4(−3)	0.170998	0.775869	0.220395	0.8260
DX4(−4)	0.375389	0.621491	0.604013	0.5472
Y(−1)	−0.249322	0.074893	−3.329031	0.0012
DY(−1)	−0.169655	0.090504	−1.874549	0.0638
DY(−2)	−0.155025	0.084890	−1.826174	0.0709
DY(−3)	−0.321287	0.084289	−3.811749	0.0002
DY(−4)	−0.023884	0.083523	−0.285956	0.7755
DUM	−0.093403	0.012909	−7.235528	0.0000
C	−0.290653	0.129715	−2.240701	0.0273

R-squared	0.711630	Mean dependent var	0.002437
Adjusted R-squared	0.641009	S.D. dependent var	0.019072
S.E. of regression	0.011427	Akaike info criterion	−8.764246
Sum squared resid	0.012797	Schwarz criterion	−8.192664
Log likelihood	389.4717	F-statistic	10.07671
Durbin-Watson stat	1.872773	Prob(F-statistic)	0.000000

We look for sequential reductions of the equation in Table 8.8. There is no unique reduction path. One looks for groups of possibly redundant variables and/or restrictions that may be validated by the usual F tests. All fourth-quarter lags in Table 8.8 are insignificant. Testing the joint significance of the group gives the first reduction shown in Table 8.9. The F and P values show that the hypothesis of joint insignificance is not rejected. The conventional inference procedure is valid since the coefficients in question are all attached to mean-zero, stationary variables. Both the Schwarz criterion (SC) and $\bar{R}^2$ move in the right direction, and so the fourth-quarter lags are deleted from the relation. Looking at insignificant coefficients in this reduced relation suggests the reduction shown at the second step in Table 8.9.

TABLE 8.9

A sequential reduction

Step	Redundant variables	S.E. of regression	F	P	$\bar{R}^2$	SC
0		0.0114			0.6410	−8.19

1	DX2(−4), DX3(−4)
	DX4(−4), DY(−4)
		0.0112	0.24	0.91	0.6497	−8.34

2	DX2(−1,−2)
	DX3(−1 to −3)
	DX4(−1 to −3)
		0.112	0.96	0.47	0.6506	−8.58

3	DX3	0.0112	0.83	0.37	0.6511	−8.61

TABLE 8.10

A reduced equation

LS // Dependent Variable is DY
Sample: 1960:1–1990:4
Included observations: 124
Excluded observations: 0 after adjusting endpoints

Variable	Coefficient	Std. error	T-statistic	Prob.
X2(−1)	−0.040729	0.012311	−3.083273	0.0013
DX2	−0.200448	0.027990	−7.161404	0.0000
DX2(−3)	−0.089591	0.031497	−2.844409	0.0053
X3(−1)	0.237299	0.058366	4.065706	0.0001
X4(−1)	−0.129531	0.033296	−3.890273	0.0002
DX4	−1.001881	0.378209	−2.649010	0.0092
Y(−1)	−0.236415	0.057339	−4.123099	0.0001
DY(−1)	−0.216244	0.064030	−3.377200	0.0010
DY(−2)	−0.197030	0.058305	−3.379316	0.0010
DY(−3)	−0.318066	0.068948	−4.613115	0.0000
DUM	−0.098825	0.011943	−8.274917	0.0000
C	−0.287923	0.108354	−2.657236	0.0090

R-squared	0.682350	Mean dependent var	0.002383
Adjusted R-squared	0.651152	S.D. dependent var	0.019004
S.E. of regression	0.011224	Akaike info criterion	−8.887612
Sum squared resid	0.014110	Schwarz criterion	−8.614682
Log likelihood	387.0836	F-statistic	21.87173
Durbin-Watson stat	1.930793	Prob(F-statistic)	0.000000

That reduction is also accepted, as is the third step. The resultant equation is shown in Table 8.10. It is interesting to note that in Table 8.9 the humble, old-fashioned $\bar{R}^2$ moves in step with the more fashionable Schwarz criterion. A further simplification might be imposed on this equation by noting that the coefficients of lagged expenditure and lagged income are effectively equal and opposite, implying a unit income elasticity. This linear restriction would not be rejected by the data, but there is little to be gained by imposing it since for all practical purposes it is already there. There is no obvious explanation for the significant third-order lag on DX2. It implies that the price change a year ago has an effect on current demand, which could be a real effect or a consequence of the seasonal adjustment procedures used in producing these data.

Setting all first differences in Table 8.10 to zero gives

$$-0.2364y - 0.0407x_2 + 0.2373x_3 - 0.1295x_4 - 0.2879 = 0$$

which implies a long-run relationship

$$\bar{y} = -1.22 - 0.17\bar{x}_2 + 1.00\bar{x}_3 - 0.55\bar{x}_4 \tag{8.55}$$

These elasticities are effectively identical with those shown in Eq. (8.52).

The relationship in Table 8.10 may be recast in terms of **levels** and reestimated.[27] The resultant relationship survives the same battery of tests as the general ADL relation in Table 8.7. However, the Chow test still indicates failure to forecast more than a year ahead. As a final experiment we have reestimated the levels equation with the 1974.1 dummy omitted. The inevitable large residual at 1974.1 leads to rejection of the normality hypothesis for the residuals. Apart from that, the relation survives the same battery of tests as the levels equation with the dummy variable included and, in addition, survives Chow forecast tests for as much as four years ahead. Consequently the equation has been refitted to the period 1959.1 to 1987.4, leaving twelve quarters for forecasting. The result is shown in Table 8.11. The implied long-run relationship is

$$\bar{y} = -1.44 - 0.16\bar{x}_2 + 0.97\bar{x}_3 - 0.55\bar{x}_4 \tag{8.56}$$

which is in close agreement with previous estimates of the elasticities. The income elasticity is approximately unity, but notice that if income and mpg both increase by 1 percent the increase in demand is somewhat less than 0.5 percent.

The result of using the regression in Table 8.11 to forecast demand in the 12 quarters from 1988.1 to 1990.4 is shown in Fig. 8.6. The F value for the Chow forecast test is 1.10, with a P value of 0.37, so the hypothesis of constant parameters is not rejected. As the graph reveals, the demand fluctuations in 1989 and 1990 were substantially greater than in the earlier years of the decade. The second quarter of 1989 saw a drop of 4.2 percent, followed by two quarters each with a 3.8 percent increase, succeeded in turn by two quarterly falls of 3.6 percent and 3.4 percent, respectively. Thus the last years of the sample period provide a formidable test of any equation. The forecast shown in Fig. 8.6 is a **static forecast,** which uses the actual values of all regressors, including lags of the dependent variable. This conforms with the derivation of the Chow forecast test in Chapter 4.[28]

Space forbids any further analysis of this data set. We have merely made a beginning and illustrated some of the many diagnostic tests now available for the development of a model. The reader is challenged to develop superior models to those already presented.

[27] Since the residuals are the same in each parameterization most of the diagnostic test statistics have identical values whether the relationship is estimated in levels or first differences, but this result is not true for all statistics. See Problem 8.10.

[28] See Problem 8.11.

TABLE 8.11
A simplified ADL relationship

LS // Dependent Variable is Y
Sample: 1960:1–1987:4
Included observations: 112 after adjusting endpoints

Variable	Coefficient	Std. error	T-statistic	Prob.
X2	−0.257697	0.040958	−6.291714	0.0000
X2(−1)	0.220987	0.038821	5.692409	0.0000
DX2(−3)	−0.074531	0.043678	−1.706382	0.0910
X3(−1)	0.223670	0.076673	2.917175	0.0044
X4	−0.658079	0.516844	−1.273264	0.2058
X4(−1)	0.530448	0.540312	0.981745	0.3286
Y(−1)	0.549319	0.090818	6.048582	0.0000
Y(−2)	0.129267	0.103599	1.247766	0.2150
Y(−3)	−0.080684	0.108495	−0.743663	0.4588
Y(−4)	0.170835	0.091941	1.858099	0.0661
C	−0.332035	0.138318	−2.400520	0.0182

R-squared	0.986554	Mean dependent var	−7.762719
Adjusted R-squared	0.985223	S.D. dependent var	0.116079
S.E. of regression	0.014111	Akaike info criterion	−8.428585
Sum squared resid	0.020110	Schwarz criterion	−8.161590
Log likelihood	324.0797	F-statistic	741.0565
Durbin-Watson stat	1.931894	Prob(F-statistic)	0.000000

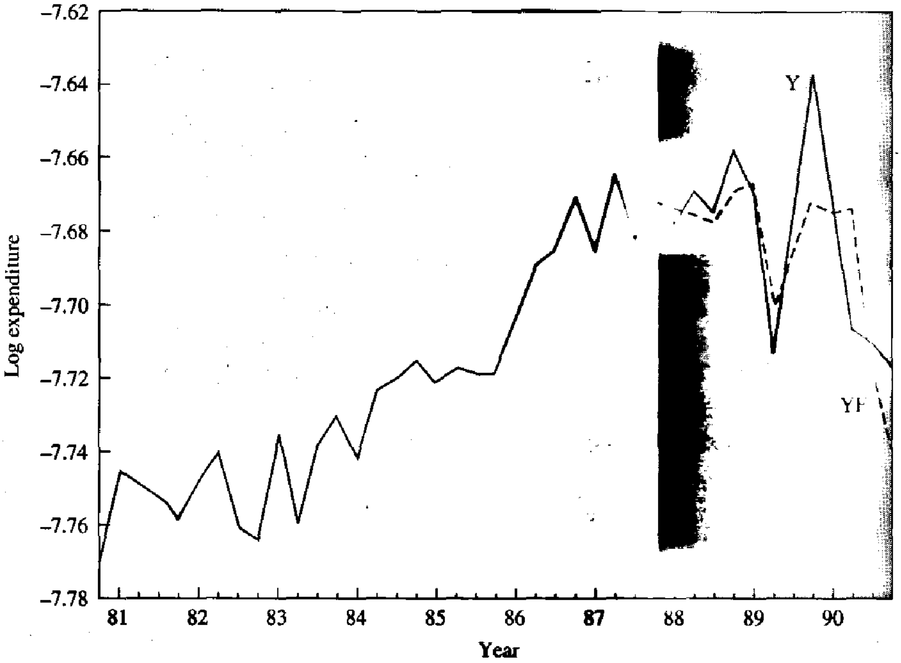


FIGURE 8.6
Actual (Y) and forecasts (YF) from Table 8.11.

8.5 NONNESTED MODELS

In the reduction sequence in Section 8.4 each model was nested within the previous model in the sense that each model was a special case of a more general model. Thus, at each stage the null hypothesis specified some restriction on the parameters of the maintained model. Rejection of the null implied acceptance of the maintained, alternative hypothesis. In many practical situations one may be faced with two models, where **neither nests** within the other. Consider

$$M_1: y = X\boldsymbol{\beta} + \boldsymbol{u}_1 \qquad \boldsymbol{u}_1 \sim N(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I}) \tag{8.57}$$

and

$$M_2: y = Z\boldsymbol{\gamma} + \boldsymbol{u}_2 \qquad \boldsymbol{u}_2 \sim N(\boldsymbol{0}, \sigma_2^2 \boldsymbol{I}) \tag{8.58}$$

where X is $n \times k$ and Z is $n \times l$. In general the two distinct models may have some explanatory variables in common, so we write

$$X = [X_1 \quad X_*] \qquad Z = [X_1 \quad Z_*]$$

If either X_* or Z_* were an empty set, one model would nest within the other and standard inference procedures would apply. In general, however, neither set of parameters can be expressed in terms of restrictions on the other set.

Testing is accomplished by setting up a **composite** or **artificial** model within which both models are nested. The composite model is

$$M_3: y = (1 - \alpha)X\boldsymbol{\beta} + \alpha(Z\boldsymbol{\gamma}) + \boldsymbol{u} \tag{8.59}$$

where α is a scalar parameter. When $\alpha = 0$, M_3 reduces to M_1. Conversely, when $\alpha = 1$, the composite model reduces to M_2. If the parameters of Eq. (8.59) could be estimated, tests on $\hat{\alpha}$ might point to one or the other model. Unfortunately $\hat{\alpha}$ cannot be recovered from the estimation of Eq. (8.59). The matrix of right-hand-side variables in that estimation is $[X_1 \; X_* \; Z_*]$, which contains fewer variables than there are structural parameters in α, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$. A solution to this problem was suggested by Davidson and MacKinnon.[29] If M_1 is being tested, the unknown $\boldsymbol{\gamma}$ vector in Eq. (8.59) is replaced by its OLS estimate from M_2. Thus $Z\boldsymbol{\gamma}$ is replaced by

$$Z\hat{\boldsymbol{\gamma}} = Z(Z'Z)^{-1}Z'y = P_z y = \hat{y}_2$$

where $\hat{y}_2$ denotes the vector of regression values from M_2. Regression (8.59) now contains $k + 1$ regressors, permitting the estimation of α and $\boldsymbol{\beta}$. If $H_0: \alpha = 0$ is not rejected, then M_1 is accepted; and conversely, rejection of H_0 implies rejection of M_1.

The same procedure may be applied to test M_2. Now the composite regression takes the form

$$y = (1 - \alpha)Z\boldsymbol{\gamma} + \alpha\hat{y}_1 + \boldsymbol{u}$$

where $\hat{y}_1 = X(X'X)^{-1}X'y = P_x y$ is the vector of regression values from M_1. The trouble with two possible tests is that there are four possible outcomes. One model

[29] Russell Davidson and James G. MacKinnon, "Several Tests for Model Specification in the Presence of Alternative Hypotheses," *Econometrica*, **49**, 1981, 781–793.

may be rejected and the other not, in which case the nonrejected model would be preferred. However, both models may be rejected or both models may not be rejected. If both are rejected our modelers obviously need to do some more work. If neither is rejected the data set does not contain sufficient information on the difference between the two specifications.

Encompassing

A related approach to the comparison of two (or more) models is based on the notion of encompassing.[30] If one model encompasses another, it is able to explain features of the rival model. For example, what can our modeler, who "believes" in Eq. (8.57), say about the γ vector in Eq. (8.58)? Our modeler might proceed in two equivalent ways. First he or she might use Eq. (8.57) to produce the $\hat{y}_1$ regression vector just defined, and then regress $\hat{y}_1$ on Z to produce his or her prediction of the γ vector, which we will denote by $\tilde{\gamma}$. Thus

$$\tilde{\gamma} = (Z'Z)^{-1}Z'P_xy = (Z'Z)^{-1}Z'X(X'X)^{-1}X'y \tag{8.60}$$

Alternatively, our modeler might recognize that the inevitable correlations between economic series will yield connections between X and Z, which may be described by the least-squares relations

$$X = Z\Pi + V \qquad \Pi = (Z'Z)^{-1}Z'X \tag{8.61}$$

Our modeler's view of the world then consists of Eqs. (8.57) and (8.61), which implies a relationship between y and Z, namely.

$$y = Z(\Pi\beta) + (u_1 - V\beta) \tag{8.62}$$

Thus our modeler expects $\gamma = \Pi\beta$, and his or her estimate will be

$$\tilde{\gamma} = \Pi\hat{\beta} = (Z'Z)^{-1}Z'X(X'X)^{-1}X'y$$

which is the estimate already defined in Eq. (8.60). The direct estimate of γ from Eq. (8.58) is

$$\hat{\gamma} = (Z'Z)^{-1}Z'y \tag{8.63}$$

The vector of contrasts is

$$\phi = \hat{\gamma} - \tilde{\gamma} = (Z'Z)^{-1}Z'M_xy \tag{8.64}$$

where $M_x = I - X(X'X)^{-1}X'$. Under M_1 we can replace y by $X\beta + u_1$.[31] Thus

$$\phi = (Z'Z)^{-1}Z'M_xu_1$$

It follows that

$$E(\phi) = 0 \qquad \text{and} \qquad \text{var}(\phi) = \sigma_1^2(Z'Z)^{-1}Z'M_xZ(Z'Z)^{-1} \tag{8.65}$$

By assuming normality, $\phi'[\text{var}(\phi)]^{-1}\phi \sim \chi^2(l)$. Substituting from Eqs. (8.64) and

[30] Grayham E. Mizon and Jean-Francois Richard, "The Encompassing Principle and Its Application to Testing Nonnested Hypotheses," *Econometrica*, **54**, 1986, 657–678.

[31] Do not confuse the use of boldface $\boldsymbol{M}$ for a matrix with M_i to denote model i.

(8.65) gives

$$\frac{1}{\sigma_1^2}y'M_xZ(Z'M_xZ)^{-1}Z'M_xy \sim \chi^2(l)$$

This expression needs a final modification to allow for the possibility of some variables being common to both models. We have

$$M_xZ = M_x[X_1 \quad Z_*] = [0 \quad M_xZ_*]$$

The relevant χ^2 quantity is now

$$\frac{1}{\sigma_1^2}y'M_xZ_*(Z_*'M_xZ_*)^{-1}Z_*'M_xy \sim \chi^2(l) \tag{8.66}$$

Implementation of the test requires an estimate of σ_1^2, but from the results on partitioned regression in Chapter 3 we see that the test is equivalent to the F test of $\boldsymbol{\gamma}_* = \mathbf{0}$ in the regression

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_*\boldsymbol{\gamma}_* + \boldsymbol{u} \tag{8.67}$$

Thus to test M_1, one supplements the variables in the M_1 model with those variables that appear in M_2 but not in M_1 and tests the joint significance of the latter group. This procedure may be contrasted with that of Davidson-MacKinnon, which supplements the M_1 variables with the single vector of regression values from M_2. If the $\boldsymbol{Z}_*$ variables are not significant, M_1 is said to **parameter encompass** M_2. It is also clear from Eq. (8.62) that the variance of the implied relationship between $\boldsymbol{y}$ and $\boldsymbol{Z}$ exceeds σ_1^2, the variance of M_1. Thus, if M_1 is true, it **variance encompasses** M_2, though sampling fluctuations may prevent the appearance of the correct variance inequality. Once again the models may be reversed and M_2 tested for parameter encompassing M_1 by running the regression of $\boldsymbol{y}$ on $\boldsymbol{Z}$ and $\boldsymbol{X}_*$ and testing the joint significance of the last group. Thus, ambiguous results may also be obtained from this procedure.

APPENDIX

APPENDIX 8.1
Nonsingular linear transformations of the variables in an equation

Suppose that we have an equation

$$y_t = m + \beta_0 x_t + \beta_1 x_{t-1} + u_t \tag{A8.1}$$

and we wish to reparamaterize as

$$y_t = m + \gamma_0 \Delta x_t + \gamma_1 x_{t-1} + u_t \tag{A8.2}$$

The data matrix for Eq. (A8.1) is

$$\boldsymbol{X} = \begin{bmatrix} \vdots & \vdots & \vdots \\ \boldsymbol{i} & \boldsymbol{x} & \boldsymbol{x}_{-1} \\ \vdots & \vdots & \vdots \end{bmatrix}$$

and the data matrix for Eq. (A8.2) is

$$Z = \begin{bmatrix} \vdots & \vdots & \vdots \\ i & (x - x_{-1}) & x_{-1} \\ \vdots & \vdots & \vdots \end{bmatrix}$$

The connection between the two matrices is $Z = XA$ where

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \quad \text{and} \quad A^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

In general, linear transformations imply a nonsingular A matrix, and we may write

$$y = X\beta + u = XAA^{-1}\beta + u = Z\gamma + u$$

where $$Z = XA \quad \text{and} \quad \gamma = A^{-1}\beta \tag{A8.3}$$

Inferences about β may be made either directly by fitting the regression on X in the usual way, or by fitting the regression on Z to estimate γ and using Eq. (A8.3) to test hypotheses about β. The direct regression gives the classic results:

$$b = (X'X)^{-1}X'y \qquad \text{var}(b) = s^2(X'X)^{-1} \qquad s^2 = e_x'e_x/(n-k)$$

$$e_x = M_x u \qquad M_x = I - X(X'X)^{-1}X'$$

The indirect estimate of β obtained from the regression on Z is

$$\begin{aligned} \hat{\beta} &= A\hat{\gamma} \\ &= A(Z'Z)^{-1}Z'y \\ &= A(A'X'XA)^{-1}A'X'y \\ &= (X'X)^{-1}X'y \\ &= b \end{aligned}$$

Thus, identical point estimates are obtained from the two methods. The two residual vectors are identical for $e_z = M_z u$ where $M_z = I - Z(Z'Z)^{-1}Z'$. Substitution from Eq. (A8.3) gives $M_z = M_x$. Thus the residual vectors are the same and each regression yields the same estimate of the residual variance. Finally,

$$\begin{aligned} \text{var}(\hat{\beta}) &= A \cdot \text{var}(\hat{\gamma}) \cdot A' \\ &= s^2A(Z'Z)^{-1}A' \\ &= s^2(X'X)^{-1} \\ &= \text{var}(b) \end{aligned}$$

Thus reparameterizations achieved by nonsingular linear transformations of the right-hand-side variables will yield identical inferences about the β vector irrespective of the reparameterization used for estimation.

First difference as regressand

Often one wishes to replace the regressand y_t by its first difference Δy_t. Consider the relation

$$y_t = \alpha y_{t-1} + \beta x_t + u_t \tag{A8.4}$$

where, for simplicity, the intercept has been suppressed. The reparameterization is

$$\Delta y_t = \gamma y_{t-1} + \beta x_t + u_t \qquad \gamma = \alpha - 1 \tag{A8.5}$$

The data matrix in each equation is

$$X = \begin{bmatrix} \vdots & \vdots \\ \boldsymbol{y}_{-1} & \boldsymbol{x} \\ \vdots & \vdots \end{bmatrix}$$

The regression in Eq. (A8.5) gives

$$\begin{aligned} \begin{bmatrix} \hat{\gamma} \\ \hat{\beta} \end{bmatrix} &= (X'X)^{-1}X'(y - y_{-1}) \\ &= (X'X)^{-1}X'y - (X'X)^{-1}X'y_{-1} \\ &= \begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} a-1 \\ b \end{bmatrix} \end{aligned}$$

where a and b are the estimated coefficients from regression (A8.4). The point estimates from the two parameterizations are identical, since $\hat{\beta} = b$ and $\hat{\alpha} = \hat{\gamma} + 1 = a$. The third step in the foregoing equation comes from

$$(X'X)^{-1}X'X = (X'X)^{-1}X'(y_{-1} \quad x) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The residuals from the two regressions are identical.[32] Thus inferences about α and β are independent of the parameterization used.

APPENDIX 8.2
To establish the equality of the test statistics in Eqs. (8.37) and (8.41)

Substituting Eqs. (8.39) and (8.40) in Eq. (8.41) gives

$$\frac{\hat{\delta}^2}{\mathrm{var}(\hat{\delta})} = \frac{1}{\sigma^2}(\hat{x}'M_x\hat{x})^{-1}(\hat{x}'M_xy)^2$$

From Eq. (8.38) $\hat{q} = (\hat{x}'\hat{x})^{-1}(\hat{x}'M_xy)$. The variance of $\hat{q}$ is

$$\mathrm{var}(\hat{q}) = \sigma^2\left[\frac{1}{(\hat{x}'\hat{x})} - \frac{1}{(x'x)}\right]$$

[32] See Problem 8.1.

Now
$$\begin{aligned}\frac{1}{\hat{x}'\hat{x}} - \frac{1}{x'x} &= \frac{x'x - \hat{x}'\hat{x}}{(\hat{x}'\hat{x})(x'x)}\\ &= \frac{1 - (x'x)^{-1}\hat{x}'\hat{x}}{\hat{x}'\hat{x}}\\ &= \frac{\hat{x}'\hat{x} - \hat{x}'\hat{x}(x'x)^{-1}\hat{x}'\hat{x}}{(\hat{x}'\hat{x})^2}\\ &= \frac{\hat{x}'M_x\hat{x}}{(\hat{x}'\hat{x})^2}\end{aligned}$$

Thus, $\text{var}(\hat{q}) = \sigma^2(\hat{x}'M_x\hat{x})(\hat{x}'\hat{x})^{-2}$. Substitution in Eq. (8.37) gives

$$\frac{\hat{q}^2}{\text{var}(\hat{q})} = \frac{1}{\sigma^2}(\hat{x}'M_x\hat{x})^{-1}(\hat{x}'M_xy)^2 = \frac{\hat{\delta}^2}{\text{var}(\hat{\delta})}$$

which completes the proof.

PROBLEMS

8.1. Show that the estimated residuals from regressions (A8.4) and (A8.5) are identical.

8.2. Derive a reparameterization of Eq. (8.12) in which the error correction term relates to period $t - 2$.

8.3. Derive a reparameterization of Eq. (8.12) that incorporates a unit elasticity assumption and that is suitable for direct estimation.

8.4. Prove that the residuals from the OLS fit of Eq. (8.16) yield an unbiased estimate of σ_u^2 when Eq. (8.15) is the correct specification for y_t.

8.5. The effects of erroneously excluding relevant variables or erroneously including irrelevant variables have been derived in the text for the simple specifications in Eqs. (8.15) and (8.16). Derive the *general* statement of these results in matrix terms, using the partitioning $X = [X_1 \; X_2]$, where X_2 can represent erroneously excluded or erroneously included variables.

8.6. The model in Eq. (8.29) may be rewritten in matrix form as

$$y = x\beta + \epsilon_1$$
$$x = x_{-1}\alpha_1 + y_{-1}\alpha_2 + \epsilon_2$$

Each equation is estimated by OLS to yield the residual vectors e_y and e_x. Two possible test regressions for weak exogeneity of x are

$$e_y \text{ on } x \text{ and } e_x$$
$$y \text{ on } x \text{ and } e_x$$

Prove that the coefficient on e_x and its estimated standard error are identical in the two regressions. Discuss the connection between the resultant t statistic and the LM test on nR^2 from the first test regression.

8.7. Examine possible reparameterizations of the equation

$$y_t = m + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \epsilon_t$$

where y and x are I(1) variables, to see which parameters may be shown as coefficients of mean-zero, I(0) variables and hence suitable for standard inference procedures.

8.8. Study the Perron article cited in Section 8.4 and apply his unit root test to the expenditure and price series in the numerical example of that section.

8.9. Develop the explicit relations between the δ and β parameters in Eq. (8.54).

8.10. Calculate a range of diagnostic test statistics for some of the equations in Section 8.4 that are formulated in both levels and first difference form. Which statistics have identical values for each form and which do not? Why?

8.11. Calculate the static forecast values shown in Fig. 8.6. Calculate also the dynamic forecast values for the same period and compare. In calculating the dynamic forecast the values of consumption beyond 1987.4 are assumed unknown and have to be replaced by forecast values.

CHAPTER 9

Multiple Equation Models

In Chapter 7 we analyzed *univariate, autoregressive* schemes, where a scalar variable is modeled in terms of its own past values. The AR(p) process, for example, is

$$y_t = m + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \epsilon_t$$

We now consider a **column vector** of k different variables, $\boldsymbol{y}_t = [y_{1t}\ y_{2t}\ \cdots\ y_{kt}]'$ and model this in terms of past values of the vector. The result is a **vector autoregression, or VAR.** The VAR(p) process is

$$\boldsymbol{y}_t = \boldsymbol{m} + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \boldsymbol{A}_2\boldsymbol{y}_{t-2} + \cdots + \boldsymbol{A}_p\boldsymbol{y}_{t-p} + \boldsymbol{\epsilon}_t \tag{9.1}$$

The $\boldsymbol{A}_i$ are $k \times k$ matrices of coefficients, $\boldsymbol{m}$ is a $k \times 1$ **vector of constants**, and $\boldsymbol{\epsilon}_t$ is a vector white noise process, with the properties

$$E(\boldsymbol{\epsilon}_t) = \boldsymbol{0} \qquad \text{for all } t \qquad E(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_s') = \begin{cases} \boldsymbol{\Omega} & s = t \\ \boldsymbol{0} & s \neq t \end{cases} \tag{9.2}$$

where the $\boldsymbol{\Omega}$ covariance matrix is assumed to be positive definite. Thus the $\boldsymbol{\epsilon}$'s are serially uncorrelated but may be contemporaneously correlated.

9.1 VECTOR AUTOREGRESSIONS (VARs)

9.1.1 A Simple VAR

To explain some of the basic features of VARs we will first consider the simple case where $k = 2$ and $p = 1$. This gives

$$\boldsymbol{y}_t = \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix} = \boldsymbol{m} + \boldsymbol{A}\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t \tag{9.3}$$

or, written out explicitly,

$$y_{1t} = m_1 + a_{11}y_{1,t-1} + a_{12}y_{2,t-1} + \epsilon_{1t}$$
$$y_{2t} = m_2 + a_{21}y_{1,t-1} + a_{22}y_{2,t-1} + \epsilon_{2t}$$

Thus, as in all VARs, each variable is expressed as a linear combination of lagged values of itself and lagged values of all other variables in the group. In practice the VAR equations may be expanded to include deterministic time trends and other exogenous variables, but we ignore these for simplicity of exposition. As may be expected from the univariate case, the behavior of the y's will depend on the properties of the A matrix.

Let the eigenvalues and eigenvectors of the A matrix be

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \qquad C = \begin{bmatrix} \vdots & \vdots \\ c_1 & c_2 \\ \vdots & \vdots \end{bmatrix}$$

Provided the eigenvalues are distinct, the eigenvectors will be linearly independent and C will be nonsingular. It then follows that

$$C^{-1}AC = \Lambda \qquad \text{and} \qquad A = C\Lambda C^{-1} \tag{9.4}$$

Define a new vector of variables z_t as

$$z_t = C^{-1}y_t \qquad \text{or} \qquad y_t = Cz_t \tag{9.5}$$

The process of premultiplying Eq. (9.3) by C^{-1} and simplifying gives

$$z_t = m^* + \Lambda z_{t-1} + \eta_t \tag{9.6}$$

where $m^* = C^{-1}m$ and $\eta_t = C^{-1}\epsilon_t$, which is a white noise vector. Thus

$$z_{1t} = m_1^* + \lambda_1 z_{1,t-1} + \eta_{1t}$$
$$z_{2t} = m_2^* + \lambda_2 z_{2,t-1} + \eta_{2t}$$

Each z variable follows a separate AR(1) scheme and is stationary, I(0), if the eigenvalue has modulus less than 1; is a random walk with drift, I(1), if the eigenvalue is 1; and is explosive if the eigenvalue exceeds 1 in numerical value. Explosive series may be ruled out as economically irrelevant. We will now consider various possible combinations of λ_1 and λ_2.

Case 1. $|\lambda_1| < 1$ and $|\lambda_2| < 1$. Each z is then I(0). Because Eq. (9.5) shows that each y is a linear combination of the z's, it follows that each y is I(0), which is written as $\boldsymbol{y}$ is I(0). Standard inference procedures apply to the VAR as formulated in Eq. (9.3) since all the variables are stationary. It also makes sense to investigate the **static equilibrium** of the system. The process of setting the disturbance vector in Eq. (9.3) to zero and assuming the existence of an equilibrium vector $\bar{y}$ gives

$$(I - A)\bar{y} = m \qquad \text{or} \qquad \Pi\bar{y} = m \tag{9.7}$$

where $\Pi = I - A$. This equation will solve for a unique, nonzero $\bar{y}$ if the Π

matrix is nonsingular. From the results on matrix algebra in Appendix A we have the following:

- The eigenvalues μ of $\boldsymbol{\Pi}$ are the complements of the eigenvalues λ of $\boldsymbol{A}$, that is, $\mu_i = 1 - \lambda_i$.
- The eigenvectors of $\boldsymbol{\Pi}$ are the same as those of $\boldsymbol{A}$.

Thus, $\boldsymbol{\Pi}$ is nonsingular in this case; and a unique static equilibrium, $\bar{y} = \boldsymbol{\Pi}^{-1}\boldsymbol{m}$, exists. The values of λ ensure that deviations from the equilibrium vector are transient and tend to die out with time.

Case 2. $\lambda_1 = 1$ and $|\lambda_2| < 1$. Now z_1 is I(1), being a random walk with drift, and z_2 is I(0). Each y is thus I(1) since it is a linear combination of an I(1) variable and an I(0) variable. We then write $\boldsymbol{y}$ is I(1). It does not now make sense to look for a static equilibrium relation between some $\bar{y}_1$ and some $\bar{y}_2$, but it is meaningful to ask if there is a cointegrating relation between y_{1t} and y_{2t}. Such a relation is readily found. The second (bottom) row in Eq. (9.5) gives

$$z_{2t} = \boldsymbol{c}^{(2)}\boldsymbol{y}_t \tag{9.8}$$

where $\boldsymbol{c}^{(2)}$ is the bottom row in $\boldsymbol{C}^{-1}$. Thus, z_2 is a linear combination of I(1) variables but is itself a stationary, I(0) variable. The cointegrating vector annihilates the I(1) component in $\boldsymbol{y}_t$. This result may be made explicit by writing Eq. (9.5) as

$$\boldsymbol{y}_t = \begin{bmatrix} \vdots \\ \boldsymbol{c}_1 \\ \vdots \end{bmatrix} z_{1t} + \begin{bmatrix} \vdots \\ \boldsymbol{c}_2 \\ \vdots \end{bmatrix} z_{2t}$$

Premultiplying across this equation by the row vector $\boldsymbol{c}^{(2)}$ then gives $\boldsymbol{c}^{(2)}\boldsymbol{y}_t = z_{2t}$ because the properties of nonsingular matrices give $\boldsymbol{c}^{(2)}\boldsymbol{c}_1 = 0$ and $\boldsymbol{c}^{(2)}\boldsymbol{c}_2 = 1$.

The cointegrating relation may also be shown in terms of the $\boldsymbol{\Pi}$ matrix defined in Eq. (9.7). Reparameterize Eq. (9.3) as

$$\Delta\boldsymbol{y}_t = \boldsymbol{m} - \boldsymbol{\Pi}\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t \tag{9.9}$$

The eigenvalues of $\boldsymbol{\Pi}$ are zero and $(1 - \lambda_2)$. Thus $\boldsymbol{\Pi}$ is a singular matrix with rank equal to one. Since it shares eigenvectors with $\boldsymbol{A}$ we have

$$\begin{aligned} \boldsymbol{\Pi} &= \boldsymbol{C}\begin{bmatrix} 0 & 0 \\ 0 & (1 - \lambda_2) \end{bmatrix}\boldsymbol{C}^{-1} \\ &= \begin{bmatrix} \vdots \\ \boldsymbol{c}_2(1 - \lambda_2) \\ \vdots \end{bmatrix} [\cdots \quad \boldsymbol{c}^{(2)} \quad \cdots] \end{aligned} \tag{9.10}$$

Thus $\boldsymbol{\Pi}$, which is of rank one, has been factorized into the product of a column vector and a row vector. This is termed an **outer product.** The row vector is the cointegrating vector already defined, and the column vector gives the weights with which the cointegrating relation enters into each equation of the VAR. This explanation may be seen more clearly by combining Eqs. (9.9) and (9.10) to get

$$\begin{aligned}\Delta y_{1t} &= m_1 - c_{12}(1-\lambda_2)z_{2,t-1} + \epsilon_{1t} \\ \Delta y_{2t} &= m_2 - c_{22}(1-\lambda_2)z_{2,t-1} + \epsilon_{2t}\end{aligned} \tag{9.11}$$

This reformulation of the VAR equations is expressed in terms of first differences and levels, all of which are I(0). It can be regarded as an **error correction** formulation of the VAR, since $z_{2,t-1}$ measures the extent to which $y_{1,t-1}$ and $y_{2,t-1}$ deviate from the long-run cointegrating relation.

NUMERICAL EXAMPLE. Consider the system

$$\begin{aligned}y_{1t} &= 1.2y_{1,t-1} - 0.2y_{2,t-1} + \epsilon_{1t} \\ y_{2t} &= 0.6y_{1,t-1} + 0.4y_{2,t-1} + \epsilon_{2t}\end{aligned} \tag{9.12}$$

where $\boldsymbol{m}$ has been set at zero. The eigenvalues of A come from the solution of

$$\begin{vmatrix}(a_{11}-\lambda) & a_{12} \\ a_{21} & (a_{22}-\lambda)\end{vmatrix} = 0$$

Thus the eigenvalues satisfy

$$\lambda_1 + \lambda_2 = \text{tr}A = 1.6 \qquad \lambda_1\lambda_2 = |A| = 0.6$$

giving $\lambda_1 = 1$ and $\lambda_2 = 0.6$. The eigenvector corresponding to the first root is obtained from

$$\begin{bmatrix}0.2 & -0.2 \\ 0.6 & -0.6\end{bmatrix}\begin{bmatrix}c_{11} \\ c_{21}\end{bmatrix} = \begin{bmatrix}0 \\ 0\end{bmatrix}$$

The eigenvector is determined only up to a scale factor. By letting $c_{21} = 1$, the first eigenvector is $\boldsymbol{c}_1 = [1 \quad 1]'$. Similarly the second eigenvector is $\boldsymbol{c}_2 = [1 \quad 3]'$. Thus,

$$\boldsymbol{C} = \begin{bmatrix}1 & 1 \\ 1 & 3\end{bmatrix} \quad \text{and} \quad \boldsymbol{C}^{-1} = \begin{bmatrix}1.5 & -0.5 \\ -0.5 & 0.5\end{bmatrix}$$

Equation (9.12) may be rewritten as

$$\begin{aligned}\Delta y_{1t} &= 0.2y_{1,t-1} - 0.2y_{2,t-1} + \epsilon_{1t} \\ \Delta y_{2t} &= 0.6y_{1,t-1} - 0.6y_{2,t-1} + \epsilon_{2t}\end{aligned}$$

or

$$\Delta \boldsymbol{y}_t = \begin{bmatrix}-0.4 \\ -1.2\end{bmatrix}[-0.5y_{1,t-1} + 0.5y_{2,t-1}] + \boldsymbol{\epsilon}_t$$

which is the numerical version of Eq. (9.11). The factorization of the $\boldsymbol{\Pi}$ matrix is not unique. Multiplication of the first vector by an arbitrary constant, followed by multiplication of the second by its reciprocal, leaves $\boldsymbol{\Pi}$ unchanged. Thus the cointegrating vector may be written as $z_t = (y_{1t} - y_{2t})$ with an appropriate adjustment to the weighting vector.

Case 3. $\lambda_1 = \lambda_2 = 1$. This case does not yield to the same analysis as the previous two cases, for there are not two linearly independent eigenvectors corresponding to the repeated eigenvalue, so that there is then no nonsingular matrix $\boldsymbol{C}$ to diagonalize A as in Eq. (9.4). As an illustration, consider the matrix

$$\boldsymbol{A} = \begin{bmatrix}0.8 & -0.4 \\ 0.1 & 1.2\end{bmatrix}$$

It is easily verified that there is a unit eigenvalue of multiplicity two; that is, $\lambda_1 = \lambda_2 = 1$. The equation $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{c} = \boldsymbol{0}$ gives

$$\begin{bmatrix} -0.2 & -0.4 \\ 0.1 & 0.2 \end{bmatrix}\begin{bmatrix} c_{11} \\ c_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Normalizing by setting $c_{21} = 1$ gives the eigenvector associated with the unit eigenvalue as $\boldsymbol{c}_1 = [-2 \quad 1]'$. But we are missing a second, linearly independent eigenvector. The source of the difficulty is that, in general, $\boldsymbol{A}$ is not symmetric. If it were symmetric, there would be two linearly independent eigenvectors associated with the repeated root.

Although $\boldsymbol{A}$ **cannot be diagonalized,** it is possible to find a nonsingular matrix $\boldsymbol{P}$ such that

$$\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} = \boldsymbol{J} \qquad \boldsymbol{A} = \boldsymbol{P}\boldsymbol{J}\boldsymbol{P}^{-1} \tag{9.13}$$

where

$$\boldsymbol{J} = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix} \tag{9.14}$$

is the **Jordan matrix** for an eigenvalue of λ with multiplicity two.[1] Now define

$$\boldsymbol{z}_t = \boldsymbol{P}^{-1}\boldsymbol{y}_t \qquad \boldsymbol{y}_t = \boldsymbol{P}\boldsymbol{z}_t \tag{9.15}$$

The process of substituting for $\boldsymbol{y}_t$ from Eq. (9.3) and simplifying gives

$$\boldsymbol{z}_t = \boldsymbol{J}\boldsymbol{z}_{t-1} + \boldsymbol{m}^* + \boldsymbol{\eta}_t \tag{9.16}$$

where $\boldsymbol{m}^* = \boldsymbol{P}^{-1}\boldsymbol{m}$ and $\boldsymbol{\eta}_t = \boldsymbol{P}^{-1}\boldsymbol{\epsilon}_t$. Spelling Eq. (9.16) out in detail, we write

$$\begin{aligned} z_{1t} &= \lambda z_{1,t-1} + z_{2,t-1} + m_1^* + \eta_{1t} \\ z_{2t} &= \lambda z_{2,t-1} + m_2^* + \eta_{2t} \end{aligned} \tag{9.17}$$

By substituting the unit eigenvalue, these equations become

$$\begin{aligned} (1 - L)z_{1t} &= z_{2,t-1} + m_1^* + \eta_{1t} \\ (1 - L)z_{2t} &= m_2^* + \eta_{2t} \end{aligned}$$

Multiplying through the first equation by $(1 - L)$ produces

$$(1 - L)^2 z_{1t} = m_2^* + (\eta_{1t} - \eta_{1,t-1} + \eta_{2,t-1})$$

Thus, z_{1t} is an I(2) series and z_{2t} is I(1). Consequently each y variable is I(2).

It is of interest to calculate the $\boldsymbol{P}$ matrix. From Eq. (9.13) it must, in general, satisfy

$$\boldsymbol{A}\begin{bmatrix} \vdots & \vdots \\ \boldsymbol{p}_1 & \boldsymbol{p}_2 \\ \vdots & \vdots \end{bmatrix} = \begin{bmatrix} \vdots & \vdots \\ \boldsymbol{p}_1 & \boldsymbol{p}_2 \\ \vdots & \vdots \end{bmatrix}\begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}$$

that is,

$$\boldsymbol{A}\boldsymbol{p}_1 = \lambda\boldsymbol{p}_1 \qquad \text{and} \qquad \boldsymbol{A}\boldsymbol{p}_2 = \boldsymbol{p}_1 + \lambda\boldsymbol{p}_2$$

[1] See Appendix A.

The first equation obviously gives the sole eigenvector already determined, namely $p_1 = c_1 = [-2 \quad 1]'$. The second equation becomes $(A - I)p_2 = p_1$. Solving produces $p_2 = [8 \quad 1]'$, and so

$$P = \begin{bmatrix} -2 & 8 \\ 1 & 1 \end{bmatrix} \qquad P^{-1} = \begin{bmatrix} -0.1 & 0.8 \\ 0.1 & 0.2 \end{bmatrix}$$

It can easily be verified that these matrices satisfy Eq. (9.13).

Finally we look for a possible cointegrating vector. Since

$$y_t = \begin{bmatrix} \vdots \\ p_1 \\ \vdots \end{bmatrix} z_{1t} + \begin{bmatrix} \vdots \\ p_2 \\ \vdots \end{bmatrix} z_{2t}$$

we need a vector orthogonal to p_1 in order to eliminate the z_1 component of y. Clearly the bottom row of P^{-1} does the trick. The cointegrating vector gives a linear combination of I(2) variables that is I(1). In this case the cointegrating variable is not stationary; but it satisfies the general definition of cointegration, which is that a vector of I(d) variables is said to be cointegrated of order (d, b), written CI(d, b), if a linear combination exists that is I$(d - b)$ for positive b. In this case y is CI(2, 1). It may also be seen that the Π matrix has rank one and the bottom row of Π also gives the cointegrating vector. All the variables in the VAR are nonstationary, as are all the first differences of these variables, so inference procedures are nonstandard in either case.

9.1.2 A Three-Variable VAR

We still retain the assumption of a first-order VAR but expand the system to three variables. Suppose the eigenvalues of the A matrix are $\lambda_1 = 1$, $|\lambda_2| < 1$, and $|\lambda_3| < 1$. Thus there exists a (3×3) nonsingular matrix, C, of eigenvectors of A. By defining a three-element z vector as in Eq. (9.5), it follows that z_{1t} is I(1) and z_{2t} and z_{3t} are each I(0). Thus all y variables are I(1). The y vector may be expressed as

$$y_t = \begin{bmatrix} \vdots \\ c_1 \\ \vdots \end{bmatrix} z_{1t} + \begin{bmatrix} \vdots \\ c_2 \\ \vdots \end{bmatrix} z_{2t} + \begin{bmatrix} \vdots \\ c_3 \\ \vdots \end{bmatrix} z_{3t}$$

To produce a linear combination of the y variables that is I(0), we need to annihilate the z_{1t} element. If we let $c^{(2)}$ and $c^{(3)}$ denote the second and third rows of C^{-1}, two cointegrating relations are available in

$$z_{2t} = c^{(2)}y_t \qquad \text{and} \qquad z_{3t} = c^{(3)}y_t \tag{9.18}$$

Each cointegrating vector is determined only up to a scale factor, but the move to three variables has introduced an entirely new consideration. A linear combination of I(0) variables is itself I(0). Thus any linear combination of the variables in Eq. (9.18) is also a cointegrating relation, with an associated cointegrating vector. **When two or more cointegrating vectors are found there is an infinity of cointegrating vectors.**

To look at the error correction formulation, we analyze the **Π** matrix. The eigenvalues are $\mu_1 = 0$, $\mu_2 = 1 - \lambda_2$, and $\mu_3 = 1 - \lambda_3$. Then

$$\begin{aligned}\Pi &= C(I - \Lambda)C^{-1} \\ &= \begin{bmatrix} \vdots & \vdots & \vdots \\ c_1 & c_2 & c_3 \\ \vdots & \vdots & \vdots \end{bmatrix}\begin{bmatrix} 0 & 0 & 0 \\ 0 & \mu_2 & 0 \\ 0 & 0 & \mu_3 \end{bmatrix}\begin{bmatrix} \cdots & c^{(1)} & \cdots \\ \cdots & c^{(2)} & \cdots \\ \cdots & c^{(3)} & \cdots \end{bmatrix} \\ &= \begin{bmatrix} \vdots & \vdots \\ \mu_2 c_2 & \mu_3 c_3 \\ \vdots & \vdots \end{bmatrix}\begin{bmatrix} \cdots & c^{(2)} & \cdots \\ \cdots & c^{(3)} & \cdots \end{bmatrix}\end{aligned} \tag{9.19}$$

Thus **Π** splits into the product of a 3 × 2 matrix of rank two and a 2 × 3 matrix, also of rank two. The latter matrix contains the two cointegrating vectors and the former gives the weights with which both cointegrating vectors enter into the error correction formulation for each Δy_i. The full set of equations, obtained by substitution in Eq. (9.9), is

$$\begin{aligned}\Delta y_{1t} &= m_1 - (\mu_2 c_{12})z_{2,t-1} - (\mu_3 c_{13})z_{3,t-1} + \epsilon_{1t} \\ \Delta y_{2t} &= m_2 - (\mu_2 c_{22})z_{2,t-1} - (\mu_3 c_{23})z_{3,t-1} + \epsilon_{2t} \\ \Delta y_{3t} &= m_3 - (\mu_2 c_{32})z_{2,t-1} - (\mu_3 c_{33})z_{3,t-1} + \epsilon_{3t}\end{aligned} \tag{9.20}$$

More compactly, repeating Eq. (9.9), we write

$$\Delta \boldsymbol{y}_t = \boldsymbol{m} - \Pi \boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t$$

The factorization of **Π** is written

$$\Pi = \boldsymbol{\alpha\beta}' \tag{9.21}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are 3 × 2 matrices of rank two.[2] The rank of **Π** is two, and there are two cointegrating vectors, shown as the rows of $\boldsymbol{\beta}'$. Substituting Eq. (9.21) in Eq. (9.9) gives

$$\Delta \boldsymbol{y}_t = \boldsymbol{m} - \boldsymbol{\alpha\beta}'\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t = \boldsymbol{m} - \boldsymbol{\alpha z}_{t-1} + \boldsymbol{\epsilon}_t \tag{9.22}$$

where $z_{t-1} = \boldsymbol{\beta}'\boldsymbol{y}_{t-1}$ contains the two cointegrating variables.

Before leaving the three-variable case, suppose that the eigenvalues are $\lambda_1 = 1$, $\lambda_2 = 1$, and $|\lambda_3| < 1$. If we follow the development in the foregoing Case 3, it is possible to find a nonsingular matrix $\boldsymbol{P}$ such that $\boldsymbol{P}^{-1}\boldsymbol{AP} = \boldsymbol{J}$, where the Jordan matrix is now

$$\boldsymbol{J} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

[2] This notation departs from our convention of using uppercase letters for matrices, but it has become embedded in the cointegration literature.

In defining a three-element vector $z_t = P^{-1}y_t$, it follows that z_1 is I(2), z_2 is I(1), and z_3 is I(0). In general all three y variables are then I(2), and we may write

$$y_t = \begin{bmatrix} \vdots \\ p_1 \\ \vdots \end{bmatrix} z_{1t} + \begin{bmatrix} \vdots \\ p_2 \\ \vdots \end{bmatrix} z_{2t} + \begin{bmatrix} \vdots \\ p_3 \\ \vdots \end{bmatrix} z_{3t}$$

Premultiplying by the second row of P^{-1}, namely, $p^{(2)}$, will annihilate both z_1 and z_3, giving $p^{(2)}y_t = z_{2t}$, which is I(1). Similarly, premultiplying by $p^{(3)}$ gives $p^{(3)}y_t = z_{3t}$, which is I(0). Thus, there are two cointegrating vectors, but only one produces a *stationary* linear combination of the y's. The reason is that y is I(2). The empirical data, however, suggest that most economic series are either I(1) or I(0).

Having a system of I(1) variables is possible even though there are multiple unit eigenvalues. Consider, for example, the matrix

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & a \end{bmatrix}$$

The first two elements in the last row are set at one for simplicity, for the only crucial element in this row is a. Clearly the eigenvalues are $\lambda_1 = 1$, $\lambda_2 = 1$, and $\lambda_3 = a$, where the last eigenvalue is assumed to have modulus less than one. The first two y variables are random walks with drift, and thus I(1), and the third equation in the VAR connects all three variables so that the third y is also I(1). The Π matrix is

$$\Pi = I - A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -1 & -1 & 1-a \end{bmatrix}$$

The rank of Π is one, and it may be factorized as

$$\Pi = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & a-1 \end{bmatrix}$$

where the row vector is the cointegrating vector. This result may be seen from

$$\begin{aligned} z_t &= y_{1t} + y_{2t} + (a-1)y_{3t} \\ &= y_{1t} + y_{2t} + (a-1)(y_{1,t-1} + y_{2,t-1} + ay_{3,t-1} + m_3 + \epsilon_{3t}) \\ &= \Delta y_{1t} + \Delta y_{2t} + az_{t-1} + (a-1)m_3 + (a-1)\epsilon_{3t} \\ &= \text{constant} + az_{t-1} + v_t \end{aligned}$$

where $v_t = \epsilon_{1t} + \epsilon_{2t} + (a-1)\epsilon_{3t}$ is a white noise series. Thus z_t follows a stable AR(1) process and is I(0).

9.1.3 Higher-Order Systems

So far we have only looked at first-order systems, but these have sufficed to illustrate the basic ideas. The extension to higher-order systems is fairly simple and may be illustrated with a second-order system,

$$\boldsymbol{y}_t = \boldsymbol{m} + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \boldsymbol{A}_2\boldsymbol{y}_{t-2} + \boldsymbol{\epsilon}_t \tag{9.23}$$

Subtracting $\boldsymbol{y}_{t-1}$ from each side gives

$$\Delta\boldsymbol{y}_t = \boldsymbol{m} + (\boldsymbol{A}_1 - \boldsymbol{I})\boldsymbol{y}_{t-1} + \boldsymbol{A}_2\boldsymbol{y}_{t-2} + \boldsymbol{\epsilon}_t$$

The process of adding and subtracting $(\boldsymbol{A}_1 - \boldsymbol{I})\boldsymbol{y}_{t-2}$ on the right side and simplifying results in

$$\Delta\boldsymbol{y}_t = \boldsymbol{m} + (\boldsymbol{A}_1 - \boldsymbol{I})\Delta\boldsymbol{y}_{t-1} - \boldsymbol{\Pi}\boldsymbol{y}_{t-2} + \boldsymbol{\epsilon}_t \tag{9.24}$$

where $\boldsymbol{\Pi} = \boldsymbol{I} - \boldsymbol{A}_1 - \boldsymbol{A}_2$. An alternative reparameterization is

$$\Delta\boldsymbol{y}_t = \boldsymbol{m} - \boldsymbol{A}_2\Delta\boldsymbol{y}_{t-1} - \boldsymbol{\Pi}\mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t \tag{9.25}$$

Thus, in the first difference reformulation of a second-order system, there will be *one* lagged first difference term on the right-hand side. The levels term may be lagged one period or two.

If we proceed in this way the VAR(p) system defined in Eq. (9.1) may be reparameterized as

$$\Delta\boldsymbol{y}_t = \boldsymbol{m} + \boldsymbol{B}_1\Delta\boldsymbol{y}_{t-1} + \cdots + \boldsymbol{B}_{p-1}\Delta\boldsymbol{y}_{t-p+1} - \boldsymbol{\Pi}\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t \tag{9.26}$$

where the $\boldsymbol{B}$s are functions of the $\boldsymbol{A}$s and $\boldsymbol{\Pi} = \boldsymbol{I} - \boldsymbol{A}_1 - \cdots - \boldsymbol{A}_p$. As shown in Appendix 9.2, the behavior of the $\boldsymbol{y}$ vector depends on the values of λ that solve $|\lambda^p\boldsymbol{I} - \lambda^{p-1}\boldsymbol{A}_1 - \cdots - \lambda\boldsymbol{A}_{p-1} - \boldsymbol{A}_p| = 0$. Ruling out explosive roots, we must consider three possibilities:

1. Rank ($\boldsymbol{\Pi}$) $= k$. If each root has modulus less than one. $\boldsymbol{\Pi}$ will have full rank and be nonsingular. All the y variables in Eq. (9.1) will be I(0), and unrestricted OLS estimates of Eq. (9.1) or Eq. (9.26) will yield identical inferences about the parameters.
2. Rank ($\boldsymbol{\Pi}$) $= r < k$. This situation will occur if there is a unit root with multiplicity $(k - r)$ and the remaining r roots are numerically less than one. The $\boldsymbol{y}$ vector will be I(1) or higher and $\boldsymbol{\Pi}$ may be expressed, following Eq. (9.21), as the outer product of two $(k \times r)$ matrices, each of rank r. The right-hand side of Eq. (9.26) then contains r cointegrating variables.
3. Rank ($\boldsymbol{\Pi}$) $= 0$. This case is rather special. It will only occur if $\boldsymbol{A}_1 + \cdots + \boldsymbol{A}_p = \boldsymbol{I}$, in which case $\boldsymbol{\Pi} = 0$ and Eq. (9.26) shows that the VAR should be specified solely in terms of first differences of the variables.

9.2 ESTIMATION OF VARs

There are two approaches to the estimation of VARs. One is the direct estimation of the system set out in Eq. (9.1) or in the alternative reparameterization of Eq. (9.26). From the argument of the previous section, direct estimation is appropriate if all the eigenvalues of $\boldsymbol{\Pi}$ are numerically less than one. The second approach, which is appropriate when the y variables are not stationary, is to determine the number r

of possible cointegrating vectors and then to estimate Eq. (9.26), with the Π matrix restricted to display r cointegrating variables. The latter approach will be discussed in the next section. In this section we will look at the *unrestricted* estimation of Eq. (9.1) or Eq. (9.26).

Since the right-hand-side variables are identical in each of the VAR equations, it follows from the discussion of seemingly unrelated regressions in Appendix 9.1 that efficient estimation of the VAR can be achieved by the separate application of OLS to each equation in the VAR. If, in addition, normally distributed disturbances may be assumed, then this procedure also provides ML estimates. This facilitates tests of various important hypotheses.

9.2.1 Testing the Order of the VAR

Suppose one fits a VAR of order p_1 and wishes to test the hypothesis that the order is $p_0 < p_1$. The null hypothesis is nested within the alternative hypothesis and may be tested by a likelihood ratio test. The maximized log-likelihood when a VAR with k variables is fitted to n observation points is[3]

$$l = \text{constant} + \frac{n}{2}\ln|\hat{\mathbf{\Omega}}^{-1}|$$

where $\hat{\mathbf{\Omega}}$ is the variance-covariance matrix of the residuals from the VAR equations, which most software packages routinely produce. When p_0 lags are used, the maximized log-likelihood is

$$l_0 = \text{constant} + \frac{n}{2}\ln|\hat{\mathbf{\Omega}}_0^{-1}|$$

and when p_1 lags are used the maximized log-likelihood is

$$l_1 = \text{constant} + \frac{n}{2}|\hat{\mathbf{\Omega}}_1^{-1}|$$

The likelihood ratio test statistic is then

$$\text{LR} = -2(l_0 - l_1) = n[\ln|\hat{\mathbf{\Omega}}_0| - \ln|\hat{\mathbf{\Omega}}_1|] \overset{a}{\sim} \chi^2(q)$$

It remains to determine the number of degrees of freedom, q. Its value is the number of restrictions imposed in determining the null hypothesis. For example, if one is testing for three lags instead of four in a two-variable VAR, two variables are excluded from each equation of the VAR, giving $q = 4$. In general, $q = k^2(p_1 - p_0)$.

9.2.2 Testing for Granger Causality

In the general VAR formulation such as Eq. (9.1), the lagged values of every variable appear in every equation of the VAR. Sometimes one may wish to test whether a

[3] James D. Hamilton, *Time Series Analysis*, Princeton University Press, 1994, 296.

specific variable or group of variables plays any role in the determination of other variables in the VAR. Suppose that a two-variable VAR is specified as

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

Here the lagged value of y_2 plays no role in the determination of y_1. Thus, y_2 is said to not Granger cause y_1. The hypothesis that y_2 does not Granger cause y_1 could be tested simply by running the regression of y_1 on lagged values of y_1 and y_2 and examining whether the coefficient of the latter variable is significantly different from zero. More generally, the $\boldsymbol{y}$ vector might be partitioned into two subvectors: $\boldsymbol{y}_1$ of order $k_1 \times 1$ and $\boldsymbol{y}_2$ of order $k_2 \times 1$. The hypothesis that the block $\boldsymbol{y}_2$ does not Granger cause $\boldsymbol{y}_1$ is tested by estimating the first k_1 equations of the VAR and testing whether the coefficients of the lagged $\boldsymbol{y}_2$ vectors differ significantly from zero. The simplest test is again a likelihood ratio test, based on the variance-covariance matrices of the residuals.

9.2.3 Forecasting, Impulse Response Functions, and Variance Decomposition

One of the principal uses of VAR systems is the production of forecasts, especially short-term forecasts. The approach is **atheoretical,** in the sense that there has been no use of economic theory to specify explicit structural equations between various sets of variables. The VAR system rests on the general proposition that economic variables tend to move together over time and also to be autocorrelated.

Suppose that we have observed the vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. Assuming a VAR(1), we will have used this data to estimate A and $\boldsymbol{\Omega}$. For the moment we will set these estimates aside and assume that we know the true values of these matrices. Suppose further that we now wish, at the end of period n, to make forecasts of the $\boldsymbol{y}$ vector one, two, three, or more periods ahead. No matter how far ahead the forecast period lies, we assume that no information is available beyond that known at the end of period n. The optimal (minimum mean squared error) forecast of $\boldsymbol{y}_{n+1}$ is the conditional expectation of $\boldsymbol{y}_{n+1}$, formed at time n, that is,

$$\hat{\boldsymbol{y}}_{n+1} = E(\boldsymbol{y}_{n+1} \mid \boldsymbol{y}_n, \ldots, \boldsymbol{y}_1) = A\boldsymbol{y}_n$$

where $\hat{\boldsymbol{y}}$ denotes a forecast vector. The usual vector of constants is omitted for simplicity. The optimal forecast two periods ahead is $E(\hat{\boldsymbol{y}}_{n+2} \mid \boldsymbol{y}_n, \ldots, \boldsymbol{y}_1)$. Evaluation of this requires an expression for $\boldsymbol{y}_{n+2}$. Repeated application of Eq. (9.3) with $m = 0$ gives

$$\boldsymbol{y}_{n+2} = A^2\boldsymbol{y}_n + A\boldsymbol{\epsilon}_{n+1} + \boldsymbol{\epsilon}_{n+2}$$

and so

$$\hat{\boldsymbol{y}}_{n+2} = A^2\boldsymbol{y}_n$$

In general

$$\boldsymbol{y}_{n+s} = A^s\boldsymbol{y}_n + A^{s-1}\boldsymbol{\epsilon}_{n+1} + \cdots + A\boldsymbol{\epsilon}_{n+s-1} + \boldsymbol{\epsilon}_{n+s}$$

and so

$$\hat{\boldsymbol{y}}_{n+s} = A^s\boldsymbol{y}_n \tag{9.27}$$

The vector of forecast errors in the forecast for s periods ahead is thus

$$e_s = y_{n+s} - \hat{y}_{n+s} = \epsilon_{n+s} + A\epsilon_{n+s-1} + \cdots + A^{s-1}\epsilon_{n+1}$$

and so the variance-covariance matrix for the forecast errors, s periods ahead, denoted by $\sum(s)$, is

$$\sum(s) = \Omega + A\Omega A' + A^2\Omega(A')^2 + \cdots + A^{s-1}\Omega(A')^{s-1} \tag{9.28}$$

Formulae (9.27) and (9.28) only apply to a first-order process (though the number of variables in the y vector is not restricted to two). Similar formulae can be developed for VAR(p) processes where $p > 1$. More importantly, the formulae are written in terms of the true matrices, so they only take account of innovation error and do not allow for coefficient uncertainty. In practice, point forecasts are derived by substituting the estimated A matrix in Eq. (9.27) or its generalization. Modifying the variance matrix for the forecast errors so as to allow for coefficient uncertainty is complicated. You should check whether your software indeed does so or merely substitutes estimated matrices in expressions such as Eq. (9.28).

9.2.4 Impulse Response Functions

Consider again the two-variable, first-order system

$$y_{1t} = a_{11}y_{1,t-1} + a_{12}y_{2,t-1} + \epsilon_{1t}$$
$$y_{2t} = a_{21}y_{1,t-1} + a_{22}y_{2,t-1} + \epsilon_{2t}$$

A perturbation in ϵ_{1t} has an immediate and one-for-one effect on y_{1t}, but no effect on y_{2t}. In period $t+1$, that perturbation in y_{1t} affects $y_{1,t+1}$ through the first equation and also affects $y_{2,t+1}$ through the second equation. These effects work through to period $t+2$, and so on. Thus a perturbation in one innovation in the VAR sets up a chain reaction over time in all variables in the VAR. Impulse response functions calculate these chain reactions.

EXAMPLE. Suppose we have a first-order system defined by

$$A = \begin{bmatrix} 0.4 & 0.1 \\ 0.2 & 0.5 \end{bmatrix} \qquad \Omega = \begin{bmatrix} 16 & 14 \\ 14 & 25 \end{bmatrix}$$

First, check that the eigenvalues of A satisfy the stationarity condition, because there is little point in studying impulse response functions for nonstationary systems. Set $y_0 = \mathbf{0}$ and postulate $\epsilon_1 = [4 \ 0]'$. This vector sets a one-standard-deviation innovation in the first equation and a zero innovation in the second equation in period one. Assume further that both innovations are zero in periods 2, 3, and so on. The first few y vectors are then given by

$$y_1 = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$$

$$y_2 = Ay_1 + \epsilon_2 = \begin{bmatrix} 0.4 & 0.1 \\ 0.2 & 0.5 \end{bmatrix}\begin{bmatrix} 4 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1.6 \\ 0.8 \end{bmatrix}$$

$$y_3 = Ay_2 = \begin{bmatrix} 0.72 \\ 0.72 \end{bmatrix}$$

TABLE 9.1
Impulse responses from $\epsilon_1 = [4\ 0]'$

Period	y_1	y_2
1	4	0
2	1.6	0.8
3	0.72	0.72
4	0.36	0.504
5	0.194	0.324

TABLE 9.2
Impulse responses from $\epsilon_1 = [0\ 5]'$

Period	y_1	y_2
1	0	5
2	0.5	2.5
3	0.45	1.35
4	0.315	0.765

The impulse responses in the first five periods for a perturbation of one standard deviation in ϵ_1 are given in Table 9.1. Similarly some impulse responses for a perturbation of one standard deviation in ϵ_2 are presented in Table 9.2.

9.2.5 Orthogonal Innovations

An objection to the procedure just illustrated for the computation of impulse response functions is that the innovations in the VAR are, in general, not contemporaneously independent of one another. That one innovation receives a perturbation and the other does not is implausible. A widely used "solution" to the problem is to transform the ϵ innovations to produce a new set of orthogonal innovations. These will be pairwise uncorrelated and have unit variances. We will denote the orthogonal innovations by u and illustrate for the two-variable case. Let $u_1 = b_{11}\epsilon_1$. The requirement of a unit sample variance gives $b_{11} = 1/s_1$, where s_1 is the sample standard deviation of ϵ_1. Next run the OLS regression of ϵ_2 on ϵ_1 to obtain the residual $u_2^* = \epsilon_2 - b_{21}\epsilon_1$. By construction this residual is uncorrelated with ϵ_1 and hence with u_1. If we denote the standard error of this regression by $s_{2.1}$, it follows that $u_2 = u_2^*/s_{2.1}$ will have unit variance and will be uncorrelated with u_1. The transformations may be summarized as

$$\boldsymbol{u}_t = \boldsymbol{P}\boldsymbol{\epsilon}_t \qquad \text{or} \qquad \boldsymbol{\epsilon}_t = \boldsymbol{P}^{-1}\boldsymbol{u}_t \tag{9.29}$$

where $\boldsymbol{u}_t = [u_{1t} \quad u_{2t}]'$, $\boldsymbol{\epsilon}_t = [\epsilon_{1t} \quad \epsilon_{2t}]'$, and

$$\boldsymbol{P} = \begin{bmatrix} \frac{1}{s_1} & 0 \\ \frac{-b_{21}}{s_{2.1}} & \frac{1}{s_{2.1}} \end{bmatrix} \qquad \text{giving} \qquad \boldsymbol{P}^{-1} = \begin{bmatrix} s_1 & 0 \\ b_{21}s_1 & s_{2.1} \end{bmatrix} \tag{9.30}$$

The sample covariance matrix for the u's is $\sum \boldsymbol{u}_t\boldsymbol{u}_t'/n$. From Eq. (9.29)

$$\frac{1}{n}\sum \boldsymbol{u}_t\boldsymbol{u}_t' = \boldsymbol{P}\left(\frac{1}{n}\sum \boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t'\right)\boldsymbol{P}' = \boldsymbol{P}\hat{\boldsymbol{\Omega}}\boldsymbol{P}'$$

But the sample covariance matrix for the u's is $\boldsymbol{I}$ by construction, and so

$$\hat{\boldsymbol{\Omega}} = \boldsymbol{P}^{-1}(\boldsymbol{P}^{-1})' \tag{9.31}$$

Equation (9.31) illustrates the **Choleski factorization** of the positive definite matrix $\hat{\boldsymbol{\Omega}}$. It is shown as the product of a lower triangular matrix, $\boldsymbol{P}^{-1}$, and its transpose, which is upper triangular.

EXAMPLE. Continuing the previous numerical example, we will suppose that the values in the A and Ω matrices have been estimated from sample data. The $\hat{\Omega}$ matrix then yields

$$s_1 = \text{standard deviation of } \epsilon_1 = \sqrt{16} = 4$$
$$b_{21} = 14/16 = 0.875$$
$$s_{2.1} = \sqrt{s_2^2(1 - r_{12}^2)} = \sqrt{25[1 - (14)^2/(16)(25)]} = 3.5707$$

giving
$$\boldsymbol{P}^{-1} = \begin{bmatrix} 4 & 0 \\ 3.5 & 3.5707 \end{bmatrix}$$

Suppose that we postulate a $\boldsymbol{u}_1 = [1 \quad 0]'$ vector and set all subsequent $\boldsymbol{u}$ vectors to zero. This vector gives a one standard deviation perturbation in the first orthogonal disturbance. From Eq. (9.29) this implies

$$\boldsymbol{\epsilon}_1 = \boldsymbol{P}^{-1}\boldsymbol{u}_1 = \begin{bmatrix} 4 & 0 \\ 3.5 & 3.5707 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 \\ 3.5 \end{bmatrix}$$

The second element in $\boldsymbol{\epsilon}_1$ is now nonzero. It is, in fact, the expected value of ϵ_{21}, given that $\epsilon_{11} = 4$. The values of the $\boldsymbol{y}$ vector may then be calculated as before. The first few values are presented in Table 9.3. Compared with the earlier assumption of a one standard deviation perturbation in just ϵ_{11}, there is now an important impact on y_2 in the first period, followed by noticeably greater impacts in subsequent periods. If a perturbation of one standard deviation in the second orthogonalized innovation is assumed, the $\boldsymbol{\epsilon}_1$ vector is given by

$$\boldsymbol{\epsilon}_1 = \boldsymbol{P}^{-1}\boldsymbol{u}_1 = \begin{bmatrix} 4 & 0 \\ 3.5 & 3.5707 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 3.5707 \end{bmatrix}$$

and the successive $\boldsymbol{y}$ vectors may be computed in the usual way.

Orthogonalized innovations were developed to deal with the problem of nonzero correlations between the original innovations. However, the solution of one problem creates another. The new problem is that the order in which the $\boldsymbol{\epsilon}$ variables are orthogonalized can have dramatic effects on the numerical results.[4] The interpretation of impulse response functions is thus a somewhat hazardous operation, and there has been intense debate on their possible economic significance.[5]

TABLE 9.3
Impulse responses from $u_1 = [1\ 0]'$

Period	y_1	y_2
1	4	3.5
2	1.95	2.55
3	1.035	1.665
4	0.580	1.039

[4] See Problem 9.4.

[5] For a very useful summary of the issues see James D. Hamilton, *Time Series Analysis*, Princeton University Press, 1994, 324–336.

9.2.6 Variance Decomposition

The variance-covariance matrix for the forecast errors was given in Eq. (9.28). For forecasts one period ahead, the relevant matrix is simply var($\boldsymbol{\epsilon}$) = $\boldsymbol{\Omega}$. Thus var($\hat{y}_{11}$) is given by the top left element in $\boldsymbol{\Omega}$, and var($\hat{y}_{21}$) is given by the bottom right element. We wish to express these forecast variances in terms of the variances of the orthogonal innovations. From Eq. (9.29) it follows that

$$\begin{aligned}\boldsymbol{\Omega} &= \boldsymbol{P}^{-1}\cdot \text{var}(\boldsymbol{u})\cdot(\boldsymbol{P}^{-1})' \\ &= \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22}\end{bmatrix}\begin{bmatrix} v_1 & 0 \\ 0 & v_2\end{bmatrix}\begin{bmatrix} c_{11} & c_{21} \\ c_{12} & c_{22}\end{bmatrix}\end{aligned} \tag{9.32}$$

where the c's denote elements of $\boldsymbol{P}^{-1}$ and v_i = var(u_i) for $i = 1, 2$. By construction, of course, each u has unit variance, and we will make that substitution in a moment. Multiplying out Eq. (9.32) gives

$$\text{var}(\hat{y}_{11}) = c_{11}^2 v_1 + c_{12}^2 v_2 \qquad \text{and} \qquad \text{var}(\hat{y}_{21}) = c_{21}^2 v_1 + c_{22}^2 v_2$$

From Eq. (9.30), $c_{12} = 0$; and so all the variance of $\hat{y}_{11}$ is attributed to the first orthogonal innovation and is equal to c_{11}^2. The variance of $\hat{y}_{21}$ is the sum of two components, namely, a proportion, $c_{21}^2/(c_{21}^2 + c_{22}^2)$, attributed to the first orthogonal innovation and the remaining proportion, $c_{22}^2/(c_{21}^2 + c_{22}^2)$, associated with the second innovation. This result is the decomposition of the forecast variance. For forecasts two or more periods ahead we return to the formula for the variance-covariance matrix of forecast errors given in Eq. (9.28). Substituting Eq. (9.31) in Eq. (9.28) it may be rewritten as

$$\Sigma(s) = \boldsymbol{P}^{-1}(\boldsymbol{P}^{-1})' + (\boldsymbol{A}\boldsymbol{P}^{-1})(\boldsymbol{A}\boldsymbol{P}^{-1})' + \cdots + (\boldsymbol{A}^{s-1}\boldsymbol{P}^{-1})(\boldsymbol{A}^{s-1}\boldsymbol{P}^{-1})' \tag{9.33}$$

Similar calculations to those already described are made for the relevant number of matrix products in Eq. (9.33) and variance decompositions are obtained. As with the impulse response functions, the numerical variance decompositions are often very sensitive to the order in which the original innovations are orthogonalized. The cautions already stated for response functions apply with equal force to variance decompositions.

9.3 VECTOR ERROR CORRECTION MODELS

When the variables in the VAR are integrated of order one or more, unrestricted estimation, as described in the previous section, is subject to the hazards of regressions involving nonstationary variables. However, the presence of nonstationary variables raises the possibility of cointegrating relations. The relevant procedure then consists of three steps:

1. Determine the **cointegrating rank,** that is, the number of cointegrating relations.
2. Estimate the matrix of cointegrating vectors, $\boldsymbol{\beta}$, and the associated weighting matrix $\boldsymbol{\alpha}$. This step amounts to determining the factorization $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$.
3. Estimate the VAR, incorporating the cointegrating relations from the previous step.

There are several methods of tackling these problems; but the maximum likelihood approach, laid out in a series of papers by Johansen, seems to have attracted the most attention from applied researchers and software developers.[6]

9.3.1 Testing for Cointegration Rank

There are two statistics for testing the hypothesis that the cointegrating rank is **at most** r ($< k$). In one case the alternative hypothesis is that the rank is k, and the test statistic is known as the **trace statistic.** In the second case, the alternative hypothesis is that the rank is $r + 1$. The test statistic is known as the **max statistic.** Some cases may be unresolved if the two statistics give conflicting indications. Distributions of the test statistics are nonstandard, and approximate asymptotic critical values have to be obtained by simulation. The paper by M. Osterwald-Lenum gives the most comprehensive set of critical values for VARs with up to 11 variables.[7] There are five tables of critical values, and it is important to select the correct one in any practical application. The tables differ according to various possible specifications of the VAR with respect to the inclusion of intercepts and time trends in both the VAR equations and the cointegrating equations. The specific range of options for the Johansen cointegration test in EViews is shown in Table 9.4. To carry out the cointegration rank test, one needs to choose from the five possible specifications the one that seems most plausible for the data in hand, and one must also specify the number of lags to include in the VAR. The default option in EViews is the third, namely, that there is an intercept in both the cointegrating equation and the differenced form of the VAR. The presence of both intercepts implies a linear trend in the levels of the series.

TABLE 9.4
Johansen cointegration test

Cointegrating equation (CE) and VAR specification	Information
Test assumes no deterministic trend in data	The test VAR is estimated in differenced form.
No intercept or trend in CE or test VAR	
Intercept (no trend) in CE—no intercept in VAR	
Test allows for linear deterministic trend in data	CE and data trend assumptions apply to levels.
Intercept (no trend) in CE and test VAR	
Intercept and trend in CE—no trend in VAR	
Test allows for quadratic deterministic trend in data	
Intercept and trend in CE—linear trend in VAR	

[6] S. Johansen, "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control*, **12**, 1988, 231–254; ———, "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models," *Econometrica*, **59**, 1991, 1551–1580; ——— and K. Juselius, "Maximum Likelihood Estimation and Inference on Cointegration—With Applications to the Demand for Money," *Oxford Bulletin of Economics and Statistics*, **52**, 1990, 169–210.

[7] M. Osterwald-Lenum, "A Note with Quantiles of the Asymptotic Distribution of the Maximum Likelihood Cointegration Rank Test Statistics," *Oxford Bulletin of Economics and Statistics*, **54**, 1992, 461–471.

As an illustration, we will apply the cointegration rank test to the gasoline data used in the numerical analysis in Section 8.4. Referring to Fig. 8.1, we see a linear deterministic trend is plausible for the income series X3, but not for consumption Y or price X2. Miles per gallon X4 (not shown) is a fairly smooth trend series. Thus, if we wish to allow for an intercept in the cointegrating equation, we should use the second or third option in Table 9.4, but there is no clear indication from the data on the choice between them. Table 9.5 gives the result of choosing the second option. The test is carried out sequentially. The first line tests the hypothesis that $r \leq 0$, that is, that there are no cointegrating relations. This is rejected at the 1 percent level. The next line tests the hypothesis of at most one cointegrating vector, and this is not rejected, so the analysis proceeds no further. The finding of one cointegrating vector does not conflict with the analysis in Section 8.4, where a plausible cointegrating relationship with stationary disturbances was established. The EViews critical values are for the trace statistic.

9.3.2 Estimation of Cointegrating Vectors

Johansen develops maximum likelihood estimators of cointegrating vectors. In the gasoline usage example there is only one such vector. The estimate is given in Table 9.6. The signs are reversed compared with the equations given in Section 8.4, since the latter had consumption as the regressand and price, income, and mpg as regressors, whereas in the cointegrating equation all variables are on the same side of the equation. The explanatory variables have the expected signs, but the numerical elasticities, especially for price and mpg, are quite different from those given by a straightforward, unlagged regression, as in Table 8.6, or from the values in Eqs. (8.55) and (8.56) yielded by the ADL analysis. The ambiguity that arises in the case of two or more cointegrating relations is not present here; nonetheless, the estimated cointegrating coefficients do not appear to be meaningful estimates of the long-run elasticities.

TABLE 9.5

Cointegration rank test of the gasoline data

Sample: 1959:1–1990:4
Included observations: 123
Test assumption: No deterministic trend in the data
Series: Y X2 X3 X4
Lags interval: 1 to 4

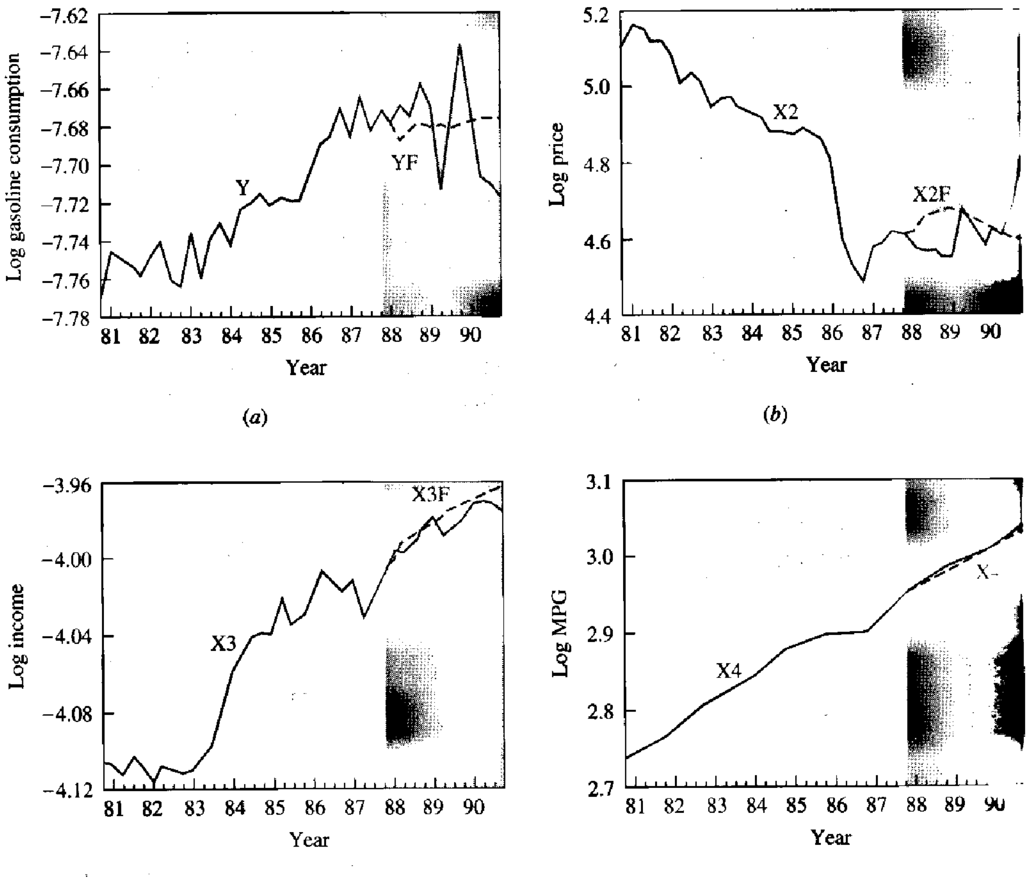
Eigenvalue	Likelihood Ratio	5 Percent Critical Value	1 Percent Critical Value	Hypothesized No. of CE(s)
0.207438	62.62120	53.12	60.16	None **
0.158998	34.02559	34.91	41.07	At most 1
0.071479	12.72675	19.96	24.60	At most 2
0.028882	3.604783	9.24	12.97	At most 3

*(**) denotes rejection of the hypothesis at 5% (1%) significance level
L.R. test indicates 1 cointegrating equation(s) at 5% significance level

TABLE 9.6
Johnasen cointegrating vector for the gasoline data

Normalized cointegrating coeffiecients: 1 cointegrating equation(s)

Y Consumption	X2 Price	X3 Income	X4 Miles per gallon	C Constant
1.000000	0.615293 (0.28260)	−0.817377 (0.13559)	0.910106 (0.24200)	−0.972113 (1.52998)
Log-likelihood	1117.203			



(*a*)

(*b*)

(*c*)

(*d*)

FIGURE 9.1
Vector error correction model with gasoline consumption, price, income, and miles per gallon: (*a*) log gasoline consumption (Y) and its forecast (YF); (*b*) log gasoline price (X2) and its forecast (X2F); (*c*) log income (X3) and its forecast (X3F); (*d*) log MPG (X4) and its forecast (X4F).

9.3.3 Estimation of a Vector Error Correction Model

The output from a VAR model is voluminous and is often best illustrated graphically. Figure 9.1 shows the result of fitting a vector error correction model to the data from 1959.1 to 1987.4, using the second option in Table 9.4 and incorporating just one cointegrating vector. The model is then used to forecast all four series for the 12 quarters from 1988.1 to 1990.4, in all cases using only data prior to 1988.1. The "forecasts" fail to capture the substantial swings in price and consumption in this period. The income forecast is reasonable, and miles per gallon presented no difficulty. Figure 9.1*a* may be compared with the 12-quarter forecast for gasoline consumption shown in Fig. 8.6, based on the ADL model of that chapter. The crucial difference is that the static forecast in Chapter 8 used the actual values of all regressors, including lags of the dependent variable, in making the forecast, whereas the VAR forecast uses no actual data beyond 1987.4.

9.4 SIMULTANEOUS STRUCTURAL EQUATION MODELS

VARs have serious limitations as a tool for the analysis of economic systems. The first problem concerns the number of variables to include in the VAR. If we are studying the macroeconomy, should we have a 10-variable system, a 20-variable system, or do we need 100 or more variables? As we will see, this question in not unique to VARs, but increasing the size of the VAR causes serious estimation problems. For example, a system of 20 variables with 4 lags would require the estimation of at least 80 coefficients in each equation of the VAR. The phenomenon might be described as the **vanishing degrees of freedom** problem, since the number of unknown coefficients can rapidly approach the available sample size. As more variables are added to the VAR, problems also arise in testing the cointegration rank. The test statistics have nonstandard distributions, which require simulations; and the currently available tables only handle up to 11 variables. With more than one cointegrating relation there is ambiguity in the interpretation of the estimated cointegrating vectors. Sometimes in applied studies one finds an author claiming economic significance for an estimated cointegrating relation on the grounds that the coefficients are close to those predicted by some economic theory. This procedure seems somewhat strange since skepticism about such theoretical specifications was a major stimulus for the development of VARs.

Letting the data "speak" with a minimum of theoretical restrictions is a laudable objective. In practice, however, the message may not be very clear and progress can only be made by imposing some more structure on the problem. Economics has a rich store of theoretical models and in **simultaneous structural equation models** we attempt to confront such theories with relevant data. Perhaps the simplest example of a structural model is the partial equilibrium, demand/supply model for a single market. On each side of the market is a set of economic agents whose behavior is described by a stochastic structural relation. *Demanders* regulate their purchases in accordance with the price that they face, and theory predicts that the partial derivative of quantity demanded with respect to price is negative. Similarly, *suppliers* adjust the quantity supplied positively with respect to price. Some mechanism clears the market each period. The linear model describing these phenomena is

$$\begin{aligned} y_{1t} + \beta_{12}y_{2t} + \gamma_{11} &= u_{1t} \\ \beta_{21}y_{1t} + y_{2t} + \gamma_{21} &= u_{2t} \end{aligned} \tag{9.34}$$

where y_1 indicates price and y_2 quantity. The model is structural, because each equation pictures the behavior of a set of economic agents, and simultaneous, because the current values of the variables appear in the equations. If the first equation depicts the demand relation, the restriction $\beta_{12} > 0$ ensures a negative slope; and $\beta_{21} < 0$ ensures a positively sloped supply function. We would also want to impose an additional restriction $\gamma_{11} < 0$ to ensure a positive intercept for the demand function. The disturbance terms u_1 and u_2 represent shifts in the functions that are the net effects of variables that, so far, have not been explicitly modeled. If both disturbances in period t were zero, the model would be represented by the D, S lines in Fig. 9.2; in this figure the equilibrium price and quantity are indicated by y_1^*, y_2^*. Nonzero disturbances shift the demand and supply curves up or down from the position shown

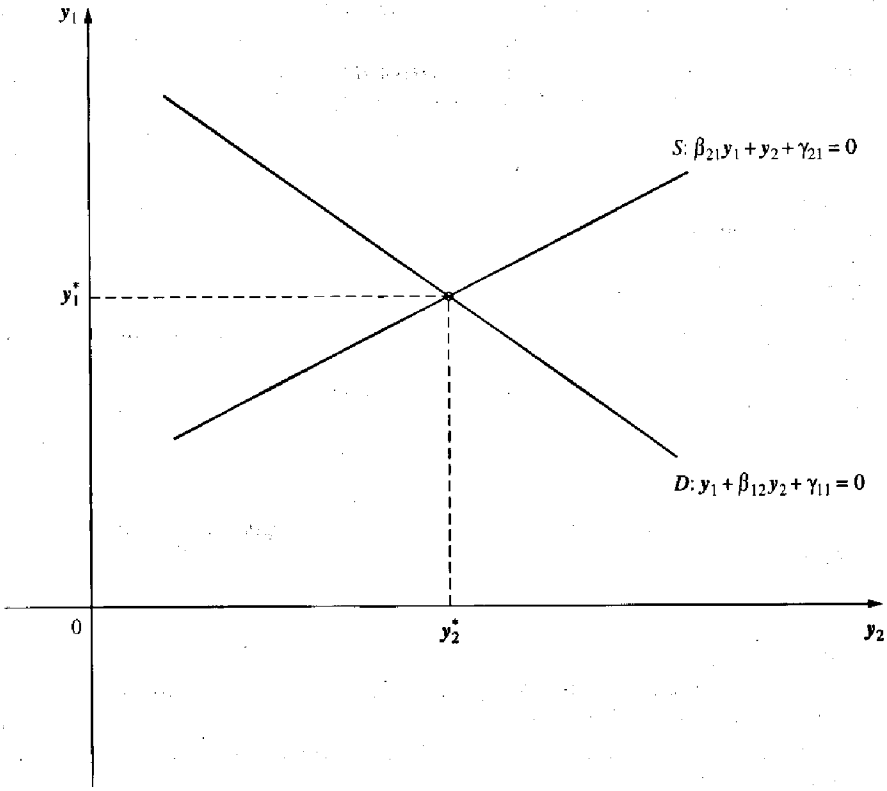


FIGURE 9.2
Partial equilibrium, demand/supply model for a single market.

in Fig. 9.2. Thus a set of random disturbances would generate a two-dimensional scatter of observations clustered around the y_1^*, y_2^* point.

A fundamentally new problem now arises. Given this two-dimensional scatter in price-quantity space, a demand analyst might fit a regression and think he was estimating a demand function. A supply analyst might fit a regression to the same data and think she was estimating a supply equation. A "general equilibrium" economist, wishing to estimate both functions, would presumably be halted on her way to the computer by the thought, "How can I estimate two separate functions from one two-dimensional scatter?" The new problem is labeled the **identification problem.** The question is, can in fact the parameters of any specific equation in a model be estimated? It is not a question of the *method of estimation,* nor of sample size, but of whether meaningful estimates of structural parameters can be obtained.[8]

We will explain the basics of the identification problem in a general matrix framework. Express Eq. (9.34) as

$$\boldsymbol{B}\boldsymbol{y}_t + \boldsymbol{C}\boldsymbol{x}_t = \boldsymbol{u}_t \tag{9.35}$$

where[9]

$$\boldsymbol{B} = \begin{bmatrix} 1 & \beta_{12} \\ \beta_{21} & 1 \end{bmatrix} \qquad \boldsymbol{y}_t = \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} \qquad \boldsymbol{C} = \begin{bmatrix} \gamma_{11} \\ \gamma_{21} \end{bmatrix} \qquad \boldsymbol{x}_t = 1 \qquad \boldsymbol{u}_t = \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \tag{9.36}$$

The model is completed with appropriate assumptions about the disturbance vector. We will assume

$$\boldsymbol{u}_t \sim \text{iid}\, N(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{9.37}$$

where $\boldsymbol{\Sigma}$ is a positive definite, variance-covariance matrix. In words, the disturbances are assumed to be normally distributed, homoscedastic, and serially uncorrelated, though possibly contemporaneously correlated. The variables in the model are classified into **endogenous** and **exogenous** categories. The endogenous variables are y_1 and y_2, and in this case the only exogenous variable is the dummy variable 1, to allow for the intercept terms. The two-equation model determines the two **current endogenous** variables y_{1t} and y_{2t} in terms of the exogenous variable and the disturbances. This dependence is shown more explicitly by premultiplying through Eq. (9.35) by $\boldsymbol{B}^{-1}$ to obtain

$$\boldsymbol{y}_t = \boldsymbol{\Pi}\boldsymbol{x}_t + \boldsymbol{v}_t \tag{9.38}$$

where[10]

$$\boldsymbol{\Pi} = -\boldsymbol{B}^{-1}\boldsymbol{C} \qquad \boldsymbol{v}_t = \boldsymbol{B}^{-1}\boldsymbol{u}_t \tag{9.39}$$

[8] In univariate time series analysis identification is used to refer to the determination of the order of the ARIMA scheme to be fitted to the data. This procedure is entirely different from the present use in the context of structural equation models.

[9] In this example C denotes a vector, and we are departing from the usual convention of denoting a vector by a lowercase symbol, for in most applications of Eq. (9.35) C will be a matrix.

[10] This use of the $\boldsymbol{\Pi}$ matrix should not be confused with the use of the same symbol in the cointegration literature in earlier sections of this chapter. Both uses are firmly embedded in the relevant literatures, so we will not attempt any change. The correct interpretation should always be evident from the context.

Equation (9.38) is known as **the reduced form** of the model. From Eqs. (9.37) and (9.39) it also follows that

$$v_t \sim \text{iid } N(\mathbf{0}, \boldsymbol{\Omega}) \qquad \boldsymbol{\Omega} = \boldsymbol{B}^{-1}\boldsymbol{\Sigma}(\boldsymbol{B}^{-1})' \tag{9.40}$$

From the reduced form Eq. (9.38) and the assumption about the reduced form disturbances in Eq. (9.40), the distribution of $\boldsymbol{y}_t$ conditional on $\boldsymbol{x}_t$ is

$$p(\boldsymbol{y}_t \mid \boldsymbol{x}_t) = (2\pi)^{-1}|\boldsymbol{\Omega}|^{-1/2}\exp(-\frac{1}{2}\boldsymbol{v}_t'\boldsymbol{\Omega}^{-1}\boldsymbol{v}_t)$$

The likelihood of the sample y's conditional on the x's is

$$\begin{aligned} L &= p(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n \mid \boldsymbol{X}) \\ &= (2\pi)^{-n}|\boldsymbol{\Omega}|^{-n/2}\exp\left(-\frac{1}{2}\sum_{t=1}^{n}\boldsymbol{v}_t'\boldsymbol{\Omega}^{-1}\boldsymbol{v}_t\right) \\ &= (2\pi)^{-n}|\boldsymbol{\Omega}|^{-n/2}\exp\left[-\frac{1}{2}\sum_{t=1}^{n}(\boldsymbol{y}_t - \boldsymbol{\Pi}\boldsymbol{x}_t)'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}_t - \boldsymbol{\Pi}\boldsymbol{x}_t)\right] \end{aligned} \tag{9.41}$$

The likelihood is seen to be completely determined by the $\boldsymbol{\Omega}$ and $\boldsymbol{\Pi}$ matrices, defined in Eqs. (9.39) and (9.40).

Now suppose another theorist constructs his market model by taking linear combinations of the structural equations in Eq. (9.35). The resultant model can be written

$$\boldsymbol{GBy}_t + \boldsymbol{GCx}_t = \boldsymbol{Gu}_t \tag{9.42}$$

where, by assumption, $\boldsymbol{G}$ is a nonsingular matrix. This second model will look the same as Eq. (9.35), in that there are two linear equations in y_1 and y_2; but the coefficients on these variables will be linear combinations of the first set of structural coefficients. The reduced form of this model is

$$\begin{aligned} \boldsymbol{y}_t &= -\boldsymbol{B}^{-1}\boldsymbol{G}^{-1}\boldsymbol{GCx}_t + \boldsymbol{B}^{-1}\boldsymbol{G}^{-1}\boldsymbol{Gu}_t \\ &= \boldsymbol{\Pi x}_t + \boldsymbol{v}_t \end{aligned}$$

which is exactly the reduced form already derived in Eq. (9.38). The two structural models have the same reduced form and the same likelihood. The structural parameters in Eq. (9.34) are then said to be unidentified. Even perfect knowledge of $\boldsymbol{\Pi}$ cannot yield the values of the β's and γ's. Another way to see this is to note that in the demand/supply model the $\boldsymbol{\Pi}$ matrix is of order 2 × 1. The $\boldsymbol{B}$ and $\boldsymbol{C}$ matrices contain four unknown parameters, namely, two β's and two γ's. There is thus an infinity of $\boldsymbol{B}$ and $\boldsymbol{C}$ matrices that satisfy $\boldsymbol{\Pi} = -\boldsymbol{B}^{-1}\boldsymbol{C}$ for any given $\boldsymbol{\Pi}$. Finally we note that the two structural equations in Eq. (9.34) are **statistically indistinguishable,** in that each is a linear combination of the same variables.

This simple demand/supply model is **unidentified.** Consider a somewhat more realistic demand/supply model,

$$\begin{aligned} y_{1t} + \beta_{12}y_{2t} + \gamma_{11}x_{1t} + \gamma_{12}x_{2t} \qquad\qquad\qquad &= u_{1t} \\ \beta_{21}y_{1t} + y_{2t} + \gamma_{21}x_{1t} \qquad\quad + \gamma_{23}x_{3t} + \gamma_{24}x_{4t} &= u_{2t} \end{aligned} \tag{9.43}$$

The variable x_1 could be taken as a dummy with the value of one in all periods to cater for the intercept term; x_2 might represent income, which economic theory suggests affects demand; and x_3 and x_4 could be variables influencing supply. It is also possible that some x variables are *lagged* y values. Lagged price, for instance, may affect current supply. It is also possible to have lagged values of income or other exogenous variables in the specification. The category of **lagged endogenous** variables and **current and lagged exogenous** variables constitutes the set of **predetermined variables.** The crucial characteristic of the predetermined variables is that they are independent of the current and future disturbances. This property holds for the exogenous variables by assumption and it holds for the lagged endogenous variables because of the assumed serial independence of the disturbance terms. This model can be cast in matrix form, as in Eq. (9.35), with

$$\boldsymbol{B} = \begin{bmatrix} 1 & \beta_{12} \\ \beta_{21} & 1 \end{bmatrix} \qquad \boldsymbol{C} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & 0 & 0 \\ \gamma_{21} & 0 & \gamma_{23} & \gamma_{24} \end{bmatrix} \tag{9.44}$$

The matrix of reduced form coefficients is then

$$-\boldsymbol{B}^{-1}\boldsymbol{C} = \frac{1}{\Delta}\begin{bmatrix} (-\gamma_{11} + \beta_{12}\gamma_{21}) & -\gamma_{12} & \beta_{12}\gamma_{23} & \beta_{12}\gamma_{24} \\ (\beta_{21}\gamma_{11} - \gamma_{21}) & \beta_{21}\gamma_{12} & -\gamma_{23} & -\gamma_{24} \end{bmatrix}$$

where $\Delta = 1 - \beta_{12}\beta_{21}$. This matrix may be contrasted with the unrestricted specification

$$\Pi = \begin{bmatrix} \pi_{11} & \pi_{12} & \pi_{13} & \pi_{14} \\ \pi_{21} & \pi_{22} & \pi_{23} & \pi_{24} \end{bmatrix} \qquad 8$$

and the question is which structural coefficients can be recovered from the π_{ij}'s. Inspection shows

$$\beta_{21} = -\pi_{22}/\pi_{12}$$

$$\beta_{12} = -\pi_{13}/\pi_{23} = -\pi_{14}/\pi_{24}$$

There are two alternative but equivalent ways of obtaining β_{12}, a reflection ~~of the fact that~~ there are eight reduced form parameters and just seven structural parameters. Having obtained the β's, we can calculate Δ and go on to determine all five γ coefficients in an obvious fashion. Thus the demand and supply equations in Eq. (9.43) are identified.

9.5 IDENTIFICATION CONDITIONS

We need some general rules for establishing whether a structural equation is identified or not. The general linear simultaneous equation model is written as

$$\boldsymbol{B}\boldsymbol{y}_t + \boldsymbol{C}\boldsymbol{x}_t = \boldsymbol{u}_t \qquad t = 1, \ldots, n \tag{9.45}$$

where $\boldsymbol{B}$ is a $G \times G$ matrix of coefficients of current endogenous variables, $\boldsymbol{C}$ is a $G \times K$ matrix of coefficients of predetermined variables, and $\boldsymbol{y}_t, \boldsymbol{x}_t$, and $\boldsymbol{u}_t$ are column vectors of G, K, and G elements, respectively, or

$$\boldsymbol{B} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1G} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2G} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{G1} & \beta_{G2} & \cdots & \beta_{GG} \end{bmatrix} \qquad \boldsymbol{C} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1K} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{G1} & \gamma_{G2} & \cdots & \gamma_{GK} \end{bmatrix}$$

$$\boldsymbol{y}_t = \begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{Gt} \end{bmatrix} \qquad \boldsymbol{x}_t = \begin{bmatrix} x_{1t} \\ x_{2t} \\ \vdots \\ x_{Kt} \end{bmatrix} \qquad \boldsymbol{u}_t = \begin{bmatrix} u_{1t} \\ u_{2t} \\ \vdots \\ u_{Gt} \end{bmatrix}$$

By contrast with the previous examples, the β coefficients have not yet been normalized. There are many normalization rules from which to choose. If appropriate, one might set the coefficient of the first endogenous variable at unity in each equation, that is, the first column in $\boldsymbol{B}$ is replaced by the unit vector. More commonly the principal diagonal of $\boldsymbol{B}$ is replaced by the unit vector. The set of equations in Eq. (9.45) may be written more compactly as

$$\boldsymbol{A}\boldsymbol{z}_t = [\boldsymbol{B} \quad \boldsymbol{C}]\begin{bmatrix} \boldsymbol{y}_t \\ \boldsymbol{x}_t \end{bmatrix} = \boldsymbol{u}_t \tag{9.46}$$

where $\boldsymbol{A}$ is the $G \times (G+K)$ matrix of all structural coefficients and $\boldsymbol{z}_t$ is the $(G+K) \times 1$ vector of observations on all variables at time t. We will consider the identification of the *first* equation of the system. The methods derived can then be applied to any structural equation. The first structural equation may be written as

$$\boldsymbol{\alpha}_1 \boldsymbol{z}_t = u_{1t}$$

where $\boldsymbol{\alpha}_1$ denotes the first row of $\boldsymbol{A}$.

Economic theory typically places restrictions on the elements of $\boldsymbol{\alpha}_1$. The most common restrictions are *exclusion* restrictions, which specify that certain variables do not appear in certain equations. Suppose, for example, that y_3 does not appear in the first equation. The appropriate restriction is then $\beta_{13} = 0$, which may be expressed as a homogeneous linear restriction on the elements of $\boldsymbol{\alpha}_1$, namely,

$$[\beta_{11} \quad \beta_{12} \quad \beta_{13} \quad \cdots \quad \gamma_{11} \quad \cdots \quad \gamma_{1K}]\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0$$

There may also be linear homogeneous restrictions involving two or more elements of $\boldsymbol{\alpha}_1$. The specification that, say, the coefficients of y_1 and y_2 are equal would be expressed as

$$[\beta_{11} \quad \beta_{12} \quad \cdots \quad \gamma_{1K}]\begin{bmatrix} 1 \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0$$

If these be the only a priori restrictions on the first equation, they may be expressed in the form

$$\boldsymbol{\alpha}_1\boldsymbol{\Phi} = \mathbf{0} \tag{9.47}$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} 0 & 1 \\ 0 & -1 \\ 1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}\begin{matrix} G \\ + \\ K \end{matrix}$$

$$\underbrace{\qquad\qquad}_{R}$$

The $\boldsymbol{\Phi}$ matrix has $G + K$ rows and a column for each a priori restriction on the first equation.

In addition to the restrictions embodied in Eq. (9.47) there will also be restrictions on $\boldsymbol{\alpha}_1$ arising from the relations between structural and reduced form coefficients. From Eq. (9.39) we may write

$$A = [B\ C] \qquad \boldsymbol{B\Pi} + \boldsymbol{C} = \mathbf{0} \qquad [B, C]\begin{bmatrix} \Pi \\ I_K \end{bmatrix} \Rightarrow B\Pi + CI_K$$

or

$$AW = [B\ C]\begin{bmatrix} \Pi \\ I_K \end{bmatrix} = 0 \rightarrow \boldsymbol{AW} = \mathbf{0} \qquad G+K$$

where

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{\Pi} \\ \boldsymbol{I_K} \end{bmatrix}\begin{matrix} G \\ K \end{matrix} \qquad K+R$$

The restrictions on the coefficients of the first structural equation are thus

$$\text{1 转换} \rightarrow \boldsymbol{\alpha}_1\boldsymbol{W} = \mathbf{0} \tag{9.48}$$

Combining Eqs. (9.47) and (9.48) gives the complete set of restrictions as

$$\boldsymbol{\alpha}_1[\boldsymbol{W} \quad \boldsymbol{\Phi}] = \mathbf{0} \tag{9.49}$$

There are $G + K$ unknowns in $\boldsymbol{\alpha}_1$. The matrix $[\boldsymbol{W} \quad \boldsymbol{\Phi}]$ is of order $(G + K) \times (K + R)$, where R is the number of columns in $\boldsymbol{\Phi}$. On the assumption that $\boldsymbol{\Pi}$ is known, all the elements in $[\boldsymbol{W} \quad \boldsymbol{\Phi}]$ are known. Thus, Eq. (9.49) constitutes a set of $K + R$ equations in $G + K$ unknowns. Identification of the first equation requires that the rank of $[\boldsymbol{W} \quad \boldsymbol{\Phi}]$ be $G + K - 1$, for then all solutions to Eq. (9.49) will lie on a single ray through the origin. Normalizing the first equation by setting one coefficient to unity (say, $\beta_{11} = 1$) gives a single point on the solution ray and thus determines $\boldsymbol{\alpha}_1$ uniquely. To summarize, identification of the first structural equation requires

$$\rho[\boldsymbol{W} \quad \boldsymbol{\Phi}] = G + K - 1 \quad \Longleftrightarrow \rho(A\Phi) = G-1 \tag{9.50}$$

This condition may be used to examine the identifiability of any structural equation in the model by determining the $\boldsymbol{\Phi}$ matrix implied by the a priori restrictions on that equation.

Implementation of this rank condition is not usually feasible for anything other than very small systems. However, a **necessary condition for identifiability** is easy to derive and apply. The rank condition cannot hold if $[\boldsymbol{W} \quad \boldsymbol{\Phi}]$ does not have at least $G + K - 1$ columns. Thus, a necessary condition for Eq. (9.50) to hold is

$$K + R \geq G + K - 1$$

which gives
$$R \geq G - 1 \tag{9.51}$$

In words,

> The number of a priori restrictions should be at least as great as the number of equations in the model less 1.

When the restrictions are solely exclusion restrictions, the necessary condition is restated as

> The number of variables excluded from the structural equation should be at least as great as the number of equations in the model less 1.

Finally, an alternative form of this last condition may be derived by letting

g = number of current endogenous variables included in the equation

k = number of predetermined variables included in the equation

Then
$$R = (G - g) + (K - k)$$

and the necessary condition becomes
$$(G - g) + (K - k) \geq G - 1$$

or
$$K - k \geq g - 1$$

that is,

> The number of predetermined variables excluded from the equation must be at least as great as the number of endogenous variables included less 1.

The necessary condition is commonly referred to as the **order condition** for identifiability. In large models this is often the only condition that can be applied since application of the rank condition becomes difficult. An alternative form of the rank condition affords an easier application, though it is still not likely to be feasible outside small-scale models. The alternative form may be stated as[11]

$$\rho[W \quad \boldsymbol{\Phi}] = G + K - 1 \qquad \text{if and only if} \qquad \rho(A\boldsymbol{\Phi}) = G - 1 \tag{9.52}$$

Note that $[W \quad \boldsymbol{\Phi}]$ is a matrix consisting of the two indicated submatrices, whereas $A\boldsymbol{\Phi}$ is the *product* of two matrices. The second form of the condition only involves the structural coefficients and thus affords an easier application. When the restrictions are all exclusion restrictions, the first row of $A\boldsymbol{\Phi}$ is a zero vector, and the remaining $G - 1$ rows consist of the coefficients in the other structural equations of the variables that do not appear in the first equation.

If equality holds in the order condition, that is, $R = G - 1$, the matrix $A\boldsymbol{\Phi}$ will be of order $G \times (G - 1)$. However, the first row of this matrix is zero by virtue of $\boldsymbol{\alpha}_1\boldsymbol{\Phi} = \mathbf{0}$. This leaves a square matrix of order $G - 1$, which, barring some freakish conjunction of coefficients, will be nonsingular. The first equation is then said to be **exactly identified** or **just identified**. If $R > G - 1$, then $A\boldsymbol{\Phi}$ has G or more columns.

[11] See F. M. Fisher, *The Identification Problem in Econometrics*, McGraw-Hill, 1966, Chapter 2; or for a shorter proof, R. W. Farebrother, "A Short Proof of the Basic Lemma of the Linear Identification Problem," *International Economic Review*, **12**, 1971, 515–516.

There are now more restrictions than the minimum required for identification, and, in general, there will be more than one square submatrix of order $G - 1$ satisfying the rank condition. The equation is then said to be **overidentified.**

EXAMPLE. Consider the system

$$\beta_{11}y_{1t} + \beta_{12}y_{2t} + \gamma_{11}x_{1t} + \gamma_{12}x_{2t} = u_{1t}$$
$$\beta_{21}y_{1t} + \beta_{22}y_{2t} + \gamma_{21}x_{1t} + \gamma_{22}x_{2t} = u_{2t}$$

As it stands, neither equation is identified, since no a priori restrictions have yet been imposed. Suppose the restrictions are

$$\gamma_{11} = 0 \qquad \gamma_{12} = 0$$

For the first equation

$$\Phi = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and

$$A\Phi = \begin{bmatrix} 0 & 0 \\ \gamma_{21} & \gamma_{22} \end{bmatrix}$$

$$\begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ \gamma_{21} & \gamma_{22} \end{bmatrix}$$

4×2

$$\begin{bmatrix} \beta_{12}\gamma_{12} & \beta_{12}\gamma_{22} \\ \beta_{22}\gamma_{21} & \beta_{22}\gamma_{22} \end{bmatrix}$$

The rank of $A\Phi$ is seen to be one and so the first equation is identified, since G equals 2. The second equation is not identified since there are no restrictions imposed on it.

Alternatively, looking at Eq. (9.49) for this model gives

$$\alpha_1[W \quad \Phi] = 0$$

or

$$[\beta_{11} \quad \beta_{12} \quad \gamma_{11} \quad \gamma_{12}] \begin{bmatrix} \pi_{11} & \pi_{12} & 0 & 0 \\ \pi_{21} & \pi_{22} & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} = [0 \quad 0 \quad 0 \quad 0]$$

Writing the equations out explicitly gives

$$\beta_{11}\pi_{11} + \beta_{12}\pi_{21} + \gamma_{11} = 0$$
$$\beta_{11}\pi_{12} + \beta_{12}\pi_{22} + \gamma_{12} = 0$$
$$\gamma_{11} = 0$$
$$\gamma_{12} = 0$$

Setting $\beta_{11} = 1$ then gives

$$\beta_{12} = -\frac{\pi_{11}}{\pi_{21}} = -\frac{\pi_{12}}{\pi_{22}}$$

This statement does not imply a contradiction, for both expressions for β_{12} will yield an identical value. The prior specifications and the normalization rule in this example give the model

$$y_{1t} + \beta_{12}y_{2t} = u_{1t}$$
$$\beta_{21}y_{1t} + y_{2t} + \gamma_{21}x_{1t} + \gamma_{22}x_{2t} = u_{2t}$$

The matrix of reduced form coefficients is

$$\Pi = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} = \frac{1}{\Delta}\begin{bmatrix} \beta_{12}\gamma_{21} & \beta_{12}\gamma_{22} \\ -\gamma_{21} & -\gamma_{22} \end{bmatrix}$$

where $\Delta = 1-\beta_{12}\beta_{21}$. Although Π is a 2×2 matrix, its rank is only 1. This is an example of overidentification. Only one prior restriction is needed to identify the first equation, but we have two. The consequence is a restriction on the reduced form coefficients. Notice also that even in the overidentified case $\rho(A\Phi)$ cannot exceed $G-1$. The matrix has G rows and at least G columns, but the first row is always zero for homogeneous restrictions and so the rank cannot exceed $G-1$. Finally we note that if Π were to be replaced in an actual two-equation problem by $\hat{\Pi}$, the matrix of *estimated* reduced form coefficients, the rank of $\hat{\Pi}$ would almost certainly be 2 and not 1, so that estimating β_{12} by $-\hat{\pi}_{11}/\hat{\pi}_{21}$ or by $-\hat{\pi}_{12}/\hat{\pi}_{22}$ would yield two different values. Such a method of estimating structural parameters is called **indirect least squares.** It only yields unique estimates of the parameters of exactly identified equations. In the more general case of overidentified equations other estimators are required.

9.6 ESTIMATION OF STRUCTURAL EQUATIONS

Consider the first equation in Eq. (9.45), which we write out explicitly as

$$y_{1t} = -\beta_{12}y_{2t} - \cdots - \beta_{1g}y_{gt} - \gamma_{11}x_{1t} - \cdots - \gamma_{1k}x_{kt} + u_{1t} \qquad t = 1,\ldots,n \tag{9.53}$$

There are several points to notice about this equation. First, the normalization condition $\beta_{11} = 1$ has been imposed. Second, it has been assumed that $g-1$ current endogenous variables appear as explanatory variables, and the variables have been suitably renumbered if necessary so that the indices run sequentially. Similarly it has been assumed that the first k predetermined variables also appear in this equation. In other words, $G-g$ current endogenous variables and $K-k$ predetermined variables have been excluded from this equation. The reduced form Eqs. (9.38) and (9.39) show that each current endogenous variable is a function of all the structural disturbances. Thus the explanatory variables $y_{2t},\ldots,y_{gt}$ in Eq. (9.53) are correlated with the disturbance u_{1t} in that equation. It therefore follows from the discussion in Chapter 5 that **the application of OLS to Eq. (9.53) will give biased and inconsistent estimates.**

The discussion in Chapter 5 also suggests that consistent estimates may be obtained by the use of **instrumental variables.** Collecting all observations in Eq. (9.53), we may write the structural equation in matrix form as

$$\boldsymbol{y} = Y_1\boldsymbol{\beta} + X_1\boldsymbol{\gamma} + \boldsymbol{u} \tag{9.54}$$

where $\boldsymbol{y}$ is the $n\times 1$ vector of observations on y_1, Y_1 is the $n\times(g-1)$ matrix of observations on the current endogenous variables on the right-hand side of the equation, X_1 is the $n\times k$ matrix of observations on the included predetermined variables, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ collect the coefficients in Eq. (9.53). This equation may be written more compactly as

$$\boldsymbol{y} = Z_1\boldsymbol{\alpha} + \boldsymbol{u} \tag{9.55}$$

where $Z_1 = [Y_1 \quad X_1]$ and $\boldsymbol{\alpha}' = [\boldsymbol{\beta}' \quad \boldsymbol{\gamma}']$. The data matrices for all variables in the model may be written

$$Y = [\boldsymbol{y} \quad Y_1 \quad Y_2] \qquad X = [X_1 \quad X_2] \tag{9.56}$$

where Y_2 is the $n \times (G - g)$ matrix of observations on the current endogenous variables that do not appear in this equation, and X_2 is the $n \times (K - k)$ matrix of observations on the excluded predetermined variables. Since by assumption all the predetermined variables are in the limit uncorrelated with the disturbances, X is an obvious set of instruments for Z_1. There are $k + g - 1$ variables in Z_1 and K variables in X. The requirement that we should have at least as many instruments as coefficients to be estimated gives the condition

$$K \geq k + g - 1$$

which is the order condition for identification of the equation.

As seen in Chapter 5 the IV estimator may be obtained by the application of two-stage least squares (2SLS). First regress Z_1 on X to obtain the matrix of predicted values,

$$\hat{Z}_1 = X(X'X)^{-1}X'Z_1 = P_X Z_1 \tag{9.57}$$

Then regress y on $\hat{Z}_1$ to obtain the IV (2SLS) estimator,

$$\hat{\alpha} = (Z_1'P_X Z_1)^{-1}Z_1'P_X y \tag{9.58}$$

with variance-covariance matrix

$$\text{var}(\hat{\alpha}) = s^2(Z_1'P_X Z_1)^{-1} \qquad s^2 = (y - Z_1\hat{\alpha})'(y - Z_1\hat{\alpha})/n \tag{9.59}$$

Tests of linear restrictions on Eq. (9.55) may be carried out as described in the last section of Chapter 5.

The matrix Z_1 in Eq. (9.55) and the matrix X of instruments have a submatrix X_1 in common. This leads to an alternative way of expressing the IV (2SLS) estimator. We have

$$Z_1'P_X Z_1 = \begin{bmatrix} Y_1'P_X Y_1 & Y_1'P_X X_1 \\ X_1'P_X Y_1 & X_1'P_X X_1 \end{bmatrix} \quad \text{and} \quad Z_1'P_X y = \begin{bmatrix} Y_1'P_X y \\ X_1'P_X y \end{bmatrix}$$

Also

$$P_X X_1 = X(X'X)^{-1}X'X_1 = [X_1 \quad X_2]\begin{bmatrix} I_k \\ 0 \end{bmatrix} = X_1$$

That is, regressing X_1 on X simply gives X_1. The estimator in Eq. (9.58) may then be written

$$\begin{bmatrix} Y_1'X(X'X)^{-1}X'Y_1 & Y_1'X_1 \\ X_1'Y_1 & X_1'X_1 \end{bmatrix}\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} Y_1'X(X'X)^{-1}X'y \\ X_1'y \end{bmatrix} \tag{9.60}$$

EXAMPLE. The first structural equation in a three-equation model is

$$y_{1t} = \beta_{12}y_{2t} + \gamma_{11}x_{1t} + \gamma_{12}x_{2t} + u_t$$

There are four predetermined variables in the complete model, and the $X'X$ matrix is

$$X'X = \begin{bmatrix} 10 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

In addition we are given

$$Y'X = \begin{bmatrix} 2 & 3 & 4 & 1 \\ 1 & 0 & 2 & 1 \end{bmatrix}$$

The necessary condition for identification is satisfied because $K - k = 2$ and $g - 1 = 1$, so the equation is overidentified. To estimate the parameters by IV (2SLS) we need to establish a correspondence between the data in this problem and the vectors and matrices in Eq. (9.60). Thus,

$$y = \begin{bmatrix} \vdots \\ y_1 \\ \vdots \end{bmatrix} \quad Y_1 = \begin{bmatrix} \vdots \\ y_2 \\ \vdots \end{bmatrix} \quad X_1 = \begin{bmatrix} \vdots & \vdots \\ x_1 & x_2 \\ \vdots & \vdots \end{bmatrix} \quad X_2 = \begin{bmatrix} \vdots & \vdots \\ x_3 & x_4 \\ \vdots & \vdots \end{bmatrix}$$

$$Y_1'X = [1 \quad 0 \quad 2 \quad 1] \qquad Y_1'X_1 = [1 \quad 0]$$

$$X_1'X_1 = \begin{bmatrix} 10 & 0 \\ 0 & 5 \end{bmatrix} \qquad X'y = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 1 \end{bmatrix} \qquad X_1'y = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

and so $$Y_1'X(X'X)^{-1}X'Y_1 = [1 \quad 0 \quad 2 \quad 1]\begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ 2 \\ 1 \end{bmatrix} = 1.6$$

$$Y_1'X(X'X)^{-1}X'y = [0.1 \quad 0 \quad 0.5 \quad 0.5]\begin{bmatrix} 2 \\ 3 \\ 4 \\ 1 \end{bmatrix} = 2.7$$

The IV (2SLS) equations are then

$$\begin{bmatrix} 1.6 & 1 & 0 \\ 1 & 10 & 0 \\ 0 & 0 & 5 \end{bmatrix}\begin{bmatrix} \hat{\beta}_{12} \\ \hat{\gamma}_{11} \\ \hat{\gamma}_{12} \end{bmatrix} = \begin{bmatrix} 2.7 \\ 2 \\ 3 \end{bmatrix}$$

with solution $$\begin{bmatrix} \hat{\beta}_{12} \\ \hat{\gamma}_{11} \\ \hat{\gamma}_{12} \end{bmatrix} = \begin{bmatrix} 1.6667 \\ 0.0333 \\ 0.6000 \end{bmatrix}$$

When dealing with medium- to large-scale econometric models, the foregoing suggestion that the matrix of all predetermined variables in the model X constitutes a suitable set of instruments for any specific structural equation may not be very helpful. The reason is that the number of variables in X may be close to, or even exceed, the number of sample observations. One possibility is to narrow the choice of instruments for each structural equation to the predetermined variables appearing in the structural equations for the variables in the Y_1 matrix relevant to that structural equation.[12]

In practice OLS is still widely used in the estimation of structural equations in spite of its acknowledged inconsistency. A possible rationalization lies in the contrast

[12] F. M. Fisher, "Dynamic Structure and Estimation in Economy-Wide Econometric Models," eds. J. Duesenberry, G. Fromm, L. R. Klein, and E. Kuh, *The Brookings Quarterly Econometric Model of the United States*, Rand-McNally, Skokie, IL, 1965, 589–636.

between small-sample and large-sample properties. Consistency is a large-sample, or asymptotic, property. Consistent estimators are not necessarily unbiased in finite samples: in fact, they usually display finite sample bias. Moreover, the sampling variance of consistent estimators, especially for a poor choice of instruments, can exceed that of OLS estimators. Thus, in finite samples OLS may show a smaller **mean squared error** than consistent estimators.

9.6.1 Nonstationary Variables

The essence of a structural equation model is an explanation of the movement of the endogenous variables in terms of the exogenous variables. The generating mechanism of the exogenous variables is not specified, but it is implicitly assumed that the endogenous variables play no role in it: if they did, the endogenous/exogenous classification would have to be respecified and the size of the structural model expanded. If the exogenous variables are integrated, say, of order one, then the endogenous variables will also be integrated of order one, and the structural equations are essentially cointegrating relations.

We saw in Chapter 8 that nonstationary variables posed special problems for conventional inference procedures from OLS regressions. A crucial question arises whether similar problems arise in the context of 2SLS regressions. This problem has been investigated by Cheng Hsiao.[13] The perhaps surprising conclusion is that the conventional 2SLS inference procedures are still valid:

> Nothing needs to be changed in applying conventional 2SLS estimator formula to estimate the unknown parameters and formulate Wald type test statistics. One gets the same point estimates and asymptotic covariance matrix. The resulting Wald type test statistic remains asymptotically chi-square distributed. In other words, nonstationarity and cointegration do not call for new estimation methods or statistical inference procedures. One can just follow the advice of Cowles Commission in constructing and testing structural equation models....
>
> For empirical structural model builders, the message is clear—one still needs to worry about the issue of identification and simultaneity bias, but one needs not to worry about the issues of nonstationarity and cointegration. All one needs to do in structural model building is to follow the conventional wisdom.

9.6.2 System Methods of Estimation

The IV (2SLS) approach is a **single equation estimator.** It may be used to estimate any identified structural equation that is the focus of interest. It may also be used seriatim to estimate each identified equation of a complete structural model. A **system estimator** estimates all (identified) parameters of a model jointly. The system

[13] Cheng Hsiao, "Statistical Properties of the Two Stage Least Squares Estimator under Cointegration," Working Paper, University of Southern California, Los Angeles, 1994.

version of two-stage least squares is **three-stage least squares** (3SLS).[14] This allows for the possibility of contemporaneous correlation between the disturbances in different structural equations. It is essentially an application of the **seemingly unrelated regression** procedure to a structural model. The identified structural equations are first estimated by 2SLS, and the resultant residuals used to estimate the disturbance covariance matrix, which is then used to estimate all identified structural parameters jointly. If the estimation process is iterated rather than stopped at the third stage, the estimates converge on the **full information maximum likelihood** (FIML) estimates of the structural model. System methods of estimation are, in principle, more efficient than single equation methods, **provided the system specification is correct.** Therein lies the rub: misspecification of a single equation can contaminate all estimates in the system.

APPENDIX

APPENDIX 9.1
Seemingly Unrelated Regressions (SUR)[15]

Suppose that the ith equation in a set of m equations is

$$y_i = X_i\boldsymbol{\beta}_i + \boldsymbol{u}_i \qquad i = 1, \ldots, m \tag{A9.1}$$

where $\boldsymbol{y}_i$ is an $n \times 1$ vector of observations on the ith variable; $\boldsymbol{X}_i$ an $n \times k_i$ matrix of observations on explanatory variables; $\boldsymbol{\beta}_i$ a $k_i \times 1$ vector of coefficients; and $\boldsymbol{u}_i$ an $n \times 1$ vector of disturbances.[16] The disturbance and explanatory variables in each equation are assumed to be uncorrelated. The y variables might be a set of consumption goods, unemployment rates in different states, or whatever. The crucial question is whether the equations should be treated separately or as a set. One possible reason for the latter is that there might be some common factors influencing the disturbances in the different equations that have not been specified explicitly in the matrices of explanatory variables. The set of equations may be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} X_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & X_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & X_m \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_m \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix} \tag{A9.2}$$

[14] A. Zellner and H. Theil, "Three Stage Least Squares: Simultaneous Estimation of Simultaneous Equations," *Econometrica*, **30**, 1962, 54–78.

[15] The basic idea comes from A. Zellner, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, **57**, 1962, 348–368.

[16] It is important in dealing with multiple equation models to avoid confusion between the use of $\boldsymbol{y}_t$ to denote observations on a number of variables at sample point t and $\boldsymbol{y}_i$ to indicate n sample observations on the ith variable.

or
$$y = X\boldsymbol{\beta} + u$$

By definition the variance-covariance matrix for u is (A9.3)

$$\Sigma = E(uu') = \begin{bmatrix} E(u_1u_1') & E(u_1u_2') & \cdots & E(u_1u_m') \\ E(u_2u_1') & E(u_2u_2') & \cdots & E(u_2u_m') \\ \vdots & \vdots & \ddots & \vdots \\ E(u_mu_1') & E(u_mu_2') & \cdots & E(u_mu_m') \end{bmatrix} \qquad \text{(A9.4)}$$

Each term in the principal diagonal of Σ is an $n \times n$ variance-covariance matrix. Thus, $E(u_iu_i')$ is the variance-covariance matrix for the disturbance in the ith equation. Each off-diagonal term in Σ represents an $n \times n$ matrix whose elements are the contemporaneous and lagged covariances between disturbances from a pair of equations. By assumption,

$$E(u_iu_j') = \sigma_{ij}I \qquad i, j = 1, \ldots, m \qquad \text{(A9.5)}$$

Setting $i = j$ gives the disturbance in any single equation as homoscedastic and nonautocorrelated. The value of the constant variance may, of course, be different in different equations. When $i \neq j$ the assumption gives a nonzero correlation between contemporaneous disturbances in the ith and jth equations but zero correlations between all lagged disturbances. Substituting Eq. (A9.5) in Eq. (A9.4) gives

$$\Sigma = \begin{bmatrix} \sigma_{11}I & \sigma_{12}I & \cdots & \sigma_{1m}I \\ \sigma_{21}I & \sigma_{22}I & \cdots & \sigma_{2m}I \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1}I & \sigma_{m2}I & \cdots & \sigma_{mm}I \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{bmatrix} \otimes I = \Sigma_c \otimes I \qquad \text{(A9.6)}$$

where I is the identity matrix of order $n \times n$ and the $\otimes$ symbol denotes Kronecker multiplication, that is, each element in Σ_c is multiplied by I.

In view of Eq. (A9.6), generalized least squares (GLS) will give a best linear unbiased estimator of the $\boldsymbol{\beta}$ vector in Eq. (A9.3); that is, the set of equations should be estimated as a group and not seriatim. The GLS estimator is

$$b_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \qquad \text{(A9.7)}$$

where
$$\Sigma^{-1} = \Sigma_c^{-1} \otimes I = \begin{bmatrix} \sigma^{11}I & \cdots & \sigma^{1m}I \\ \vdots & \ddots & \vdots \\ \sigma^{m1}I & \cdots & \sigma^{mm}I \end{bmatrix} \qquad \text{(A9.8)}$$

The variance-covariance matrix for the GLS estimator is

$$\text{var}(b_{GLS}) = (X'\Sigma^{-1}X)^{-1} \qquad \text{(A9.9)}$$

The obvious operational difficulty with this estimator is that the elements of Σ_c are unknown. Zellner's suggestion is to construct a **feasible GLS** estimator by estimating each of the m equations separately by OLS and using the residuals to estimate the σ_{ij}. Inference procedures on the resultant model now have only asymptotic validity.

There are two important special cases of the SUR estimator. If

$$\sigma_{ij} = 0 \qquad i \neq j$$

or $$X_1 = X_2 = \cdots = X_m$$

the GLS estimator reduces to the application of OLS to each equation separately.[17] If the disturbances are also normally distributed, the OLS estimators are also ML estimators.

APPENDIX 9.2
Higher-order VARs

The general VAR(p) process was defined in Eq. (9.1) as

$$\boldsymbol{y}_t = \boldsymbol{m} + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \boldsymbol{A}_2\boldsymbol{y}_{t-2} + \cdots + \boldsymbol{A}_p\boldsymbol{y}_{t-p} + \boldsymbol{\epsilon}_t$$

where the $\boldsymbol{y}$ vectors contain k variables and there are p lags. Most of Section 9.1 was devoted to an examination of just VAR(1) processes, namely,

$$\boldsymbol{y}_t = \boldsymbol{m} + \boldsymbol{A}\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t \tag{A9.10}$$

There we saw that the stationarity or otherwise of the y variables was determined by the eigenvalues of $\boldsymbol{A}$. This appendix examines the extension to higher-order VARs. Before doing so, it will be helpful to look at the first-order case in a slightly different way.

A9.2.1 A VAR(1) Process

Omitting the disturbance vector reduces a VAR to a set of **simultaneous linear difference equations,**

$$\boldsymbol{y}_t = \boldsymbol{m} + \boldsymbol{A}\boldsymbol{y}_{t-1} \tag{A9.11}$$

A solution to Eq. (A9.11) expresses $\boldsymbol{y}_t$ as a function of time and will consist of the sum of a **particular integral** and the **complementary function.** The particular integral is any solution to Eq. (A9.11). The simplest such solution is obtained by setting $\boldsymbol{y}_t = \boldsymbol{y}_{t-1} = \bar{\boldsymbol{y}}$, which gives

$$(\boldsymbol{I} - \boldsymbol{A})\bar{\boldsymbol{y}} = \boldsymbol{\Pi}\bar{\boldsymbol{y}} = \boldsymbol{m} \tag{A9.12}$$

where $\boldsymbol{\Pi} = \boldsymbol{I} - \boldsymbol{A}$. If we assume for the moment that $\boldsymbol{\Pi}$ is nonsingular, the particular integral is $\bar{\boldsymbol{y}} = \boldsymbol{\Pi}^{-1}\boldsymbol{m}$. The complementary function is the solution to the **homogeneous equation**

$$\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{y}_{t-1} \tag{A9.13}$$

As a possible solution try $\boldsymbol{y}_t = \boldsymbol{c}\lambda^t$, where $\boldsymbol{c}$ is a column vector of k constants and λ is a scalar. The process of substituting the trial solution in Eq. (A9.13) and dividing through by λ^{t-1} gives $\lambda\boldsymbol{c} = \boldsymbol{A}\boldsymbol{c}$, which may be rewritten as

$$(\lambda\boldsymbol{I} - \boldsymbol{A})\boldsymbol{c} = \boldsymbol{0} \tag{A9.14}$$

[17] See Problem 9.3.

A nontrivial solution for c requires the determinant of the coefficient matrix to be zero; that is,

$$|\lambda I - A| = 0 \tag{A9.15}$$

The λ's that solve Eq. (A9.15) are the *eigenvalues* of A. Substituting each value of λ in Eq. (A9.14) gives a corresponding c vector, which is the *eigenvector* associated with that λ. The complete solution to Eq. (A9.11) for the two-variable case is then

$$y_t = c_1\lambda_1^t + c_2\lambda_2^t + \bar{y} \tag{A9.16}$$

If each eigenvalue has modulus less than one, the terms in λ^t tend to zero with increasing t, and y_t converges on the constant vector $\bar{y}$. In this case the latter vector may be interpreted as a static equilibrium, since $\Pi = I - A$ is nonsingular. The reason is that if A had a unit root, then substitution in Eq. (A9.15) would give $|I - A| = 0$, giving a singular Π matrix. However, in the present case there is no unit root and $|I - A| \neq 0$, signifying a nonsingular Π matrix. Finally, we note that Eq. (A9.11) may be written, using a polynomial in the lag operator, as

$$A(L)y_t = m \qquad \text{where} \qquad A(L) = I - AL$$

Writing $A(L) = I - AL = (1 - \lambda_1 L)(1 - \lambda_2 L)$ shows that the condition that the λ's have modulus less than one is the same as the roots of $A(L)$ lying outside the unit circle. If there are one or more unit roots and one or more roots with modulus less than one, Π is singular and we have the cointegration error correction formulation discussed in Section 9.1. Should there be no unit roots but one or more λ's with modulus greater than one, Π is nonsingular and a vector $\bar{y}$ exists; but it has no equilibrium connotations because Eq. (A9.16) shows one or more terms in λ^t increasing without limit as t increases.

9.2.2 A VAR(2) Process

Setting the disturbance to zero, this process may be written as

$$y_t = m + A_1 y_{t-1} + A_2 y_{t-2} \tag{A9.17}$$

The particular integral is

$$(I - A_1 - A_2)\bar{y} = \Pi\bar{y} = m \tag{A9.18}$$

where $\Pi = I - A_1 - A_2$. The homogeneous equation may be solved by trying $y_t = c\lambda^t$ as before. Substitution in Eq. (A9.17) gives the determinantal equation

$$\left|\lambda^2 I - \lambda A_1 - A_2\right| = 0 \tag{A9.19}$$

This equation has $2k$ roots, where k is the number of variables in the VAR. If all roots have modulus less than one, the vector $y = \Pi^{-1}m$ exists since

$$|\Pi| = |I - A_1 - A_2| \neq 0$$

The solution is then

$$y_t = \sum_{i=1}^{2k} c_i\lambda^t + \bar{y} \tag{A9.20}$$

showing y_t converging to $\bar{y}$ with increasing t. Unit or explosive roots have the same interpretation as in the VAR(1) case.

An alternative approach is based on the fact that a VAR of order 2 or more can always be transformed into a VAR(1) in transformed variables. As an illustration, Eq. (A9.17) can be rewritten as

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} A_1 & A_2 \\ I & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} m \\ 0 \end{bmatrix} \tag{A9.21}$$

If we denote the matrix of coefficients in Eq. (A9.21) by A, the characteristic equation is

$$|\lambda I - A| = \begin{vmatrix} \lambda I - A_1 & -A_2 \\ -I & \lambda I \end{vmatrix} = 0 \tag{A9.22}$$

where the identity matrix on the left is of order $2k \times 2k$, and the one in the partitioned form is of order $k \times k$. Multiplying the first k rows in Eq. (A9.22) by λ and dividing the last k columns by λ will leave the value of the determinant unchanged, giving

$$|\lambda I - A| = \begin{vmatrix} \lambda^2 I - \lambda A_1 & -A_2 \\ -I & I \end{vmatrix} = 0 \tag{A9.23}$$

One formula for the determinant of a partitioned matrix is

$$\begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = |A_{22}| \cdot |A_{11} - A_{12}A_{22}^{-1}A_{21}|$$

Applying this result to Eq. (A9.23) gives the characteristic equation as

$$|\lambda^2 I - \lambda A_1 - A_2| = 0$$

which is the same as Eq. (A9.19).

By a similar argument the characteristic equation for the general VAR(p) process defined in Eq. (9.1) is

$$\left|\lambda^p I - \lambda^{p-1} A_1 - \cdots - \lambda A_{p-1} - A_p\right| = 0$$

PROBLEMS

9.1. In a two-equation, first-order VAR system choose A matrices to illustrate Cases 1, 2, and 3 of Section 9.1. In each case carry out a small Monte Carlo experiment on your PC by generating some appropriate innovations and calculating the resultant y series. Graph the various y series and any cointegrating series you find.

9.2. In a three-equation, first-order VAR suppose the A matrix is

$$A = \begin{bmatrix} 1.75 & -0.25 & -0.25 \\ 1.75 & 0.75 & -1.25 \\ 1 & 0 & 0 \end{bmatrix}$$

Find the eigenvalues of A. Generate some y series and determine the order of integration. What is the rank of the Π matrix? What cointegrating vectors can you find?

Repeat the exercise for

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0.5 \end{bmatrix}$$

9.3. Prove that the SUR estimator reduces to the application of OLS separately to each equation in the set if the disturbances in the equations are pairwise uncorrelated or if the matrix of explanatory variables is the same in each equation.

9.4. Reverse the order of the ϵ variables in the orthogonal innovations example in Section 9.2, compute some terms in the impulse response functions, and compare with the results in Section 9.2.

9.5. The structure of the Klein macro model is

$$\begin{aligned}
C &= \alpha_0 + \alpha_1(W_P + W_G) + \alpha_2\Pi + \alpha_3\Pi_{-1} + u_1 \\
I &= \beta_0 + \beta_1\Pi + \beta_2\Pi_{-1} + \beta_3K_{-1} + u_2 \\
W_P &= \gamma_0 + \gamma_1(Y + T - W_G) + \gamma_2(Y + T - W_G)_{-1} + \gamma_3 t + u_3 \\
Y &= C + I + G \\
\Pi &= Y - W_P - T \\
K &= K_{-1} + I
\end{aligned}$$

The six endogenous variables are Y (output), C (consumption), I (net investment), W_P (private wages), Π (profits), and K (capital stock at year-end). In addition to the constant, the exogenous variables are G (government nonwage expenditure), W_G (public wages), T (business taxes), and t (time). Examine the rank condition for the identifiability of the consumption function.

9.6. In the model

$$\begin{aligned}
y_{1t} + \beta_{12}y_{2t} + \gamma_{11}x_{1t} &= u_{1t} \\
y_{2t} + \beta_{21}y_{1t} + \gamma_{22}x_{2t} + \gamma_{23}x_{3t} &= u_{2t}
\end{aligned}$$

the y's are endogenous, the x's are exogenous, and $u_t' = [u_{1t} \quad u_{2t}]$ is a vector of serially independent normal random disturbances with zero mean vector and the same nonsingular covariance matrix for each t. Given the following sample second moment matrix, calculate the 2SLS estimates of β_{12} and γ_{11}.

	y_1	y_2	x_1	x_2	x_3
y_1	14	6	2	3	0
y_2	6	10	2	1	0
x_1	2	2	1	0	0
x_2	3	1	0	1	0
x_3	0	0	0	0	1

(University of Michigan, 1981)

9.7. An investigator has specified the following two models and proposes to use them in some empirical work with macroeconomic time series data.

Model 1.

$$\begin{aligned} c_t &= \alpha_1 y_t + \alpha_2 m_{t-1} + u_{1t} \\ i_t &= \beta_1 y_t + \beta_2 r_t + u_{2t} \\ y_t &= c_t + i_t \end{aligned}$$

Jointly dependent variables: c_t, i_t, y_t

Predetermined variables: r_t, m_{t-1}

Model 2.

$$\begin{aligned} m_t &= \gamma_1 r_t + \gamma_2 m_{t-1} + v_{1t} \\ r_t &= \delta_1 m_t + \delta_2 m_{t-1} + \delta_3 y_t + v_{2t} \end{aligned}$$

Jointly dependent variables: m_t, r_t

Predetermined variables: m_{t-1}, y_t

(*a*) Assess the identifiability of the parameters that appear as coefficients in the foregoing two models (treating the two models separately).

(*b*) Obtain the reduced form equation for y_t in model 1 and the reduced form equation for r_t in model 2.

(*c*) Assess the identifiability of the two-equation model comprising the reduced form equation for y_t in model 1 (an IS curve) and the reduced form equation for r_t in model 2 (an LM curve).

(Yale University, 1980)

9.8. (*a*) Assess the identifiability of the parameters of the following five-equation system:

$$\begin{aligned} y_{1t} + \beta_{12} y_{2t} + \beta_{14} y_{4t} + \gamma_{11} z_{1t} + \gamma_{14} z_{4t} &= u_{1t} \\ y_{2t} + \beta_{23} y_{3t} + \beta_{25} y_{5t} + \gamma_{22} z_{2t} &= u_{2t} \\ y_{3t} + \gamma_{31} z_{1t} + \gamma_{33} z_{3t} &= u_{3t} \\ \beta_{41} y_{1t} + \beta_{43} y_{3t} + y_{4t} + \gamma_{42} z_{2t} + \gamma_{44} z_{4t} &= u_{4t} \\ 2y_{3t} + y_{5t} - z_{2t} &= 0 \end{aligned}$$

(*b*) How are your conclusions altered if $\gamma_{33} = 0$? Comment.

(*c*) Explain briefly how you would estimate the parameters of each equation in this model. What can be said about the parameters of the second equation?

(University of London, 1979)

9.9. The model given by

$$\begin{aligned} y_{1t} &= \beta_{12} y_{2t} + \gamma_{11} z_{1t} + \gamma_{12} z_{2t} + \epsilon_{1t} \\ y_{2t} &= \beta_{21} y_{1t} + \gamma_{23} z_{3t} + \epsilon_{2t} \end{aligned}$$

generates the following matrix of second moments:

	y_1	y_2	z_1	z_2	z_3
y_1	3.5	3	1	1	0
y_2		11.5	1	3	4
z_1			1	0	0
z_2				1	1
z_3					2

Calculate the following:

(*a*) OLS estimates of the unrestricted reduced form parameters

(*b*) Indirect least squares (ILS) estimates of the parameters of the first equation

(*c*) 2SLS estimates of the parameters of the second equation

(*d*) The restricted reduced form derived from parts (*b*) and (*c*)

(*e*) A consistent estimate of $E(\epsilon_{1t}\epsilon_{2t}) = \sigma_{12}$

(University of London, 1973)

9.10. Let the model be

$$y_{1t} + \beta_{12}y_{2t} + \gamma_{12}x_{2t} + \gamma_{13}x_{3t} = u_{1t}$$
$$\beta_{21}y_{1t} + y_{2t} + \gamma_{21}x_{1t} + \gamma_{24}x_{4t} = u_{2t}$$

If the second-moment matrices of a sample of 100 observations are

$$Y'Y = \begin{bmatrix} 80.0 & -4.0 \\ -4.0 & 5.0 \end{bmatrix} \qquad Y'X = \begin{bmatrix} 2.0 & 1.0 & -3.0 & -5.0 \\ -0.5 & 1.5 & 0.5 & -1.0 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 3.0 & 0 & 0 & 0 \\ 0 & 2.0 & 0 & 0 \\ 0 & 0 & 1.0 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}$$

find the 2SLS estimates of the coefficients of the first equation and their standard errors.

(University of London, 1979)

9.11. The $X'X$ matrix for all the exogenous variables in a model is

$$X'X = \begin{bmatrix} 7 & 0 & 3 & 1 \\ 0 & 2 & -2 & 0 \\ 3 & -2 & 5 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

Only the first of these exogenous variables has a nonzero coefficient in a structural equation to be estimated by 2SLS. This equation includes two endogenous variables, and the OLS estimates of the reduced form coefficients for these two variables are

$$\begin{bmatrix} 0 & 1 & 3 & 2 \\ 1 & -1 & 1 & -1 \end{bmatrix}$$

Taking the first endogenous variable as the dependent variable, state and solve the equation for the 2SLS estimates.

9.12. For the model

$$y_{1t} = \beta_{12}y_{2t} + \gamma_{11}x_{1t} + u_{1t}$$
$$y_{2t} = \beta_{21}y_{1t} + \gamma_{22}x_{2t} + \gamma_{23}x_{3t} + u_{2t}$$

you are given the following information:

1. The OLS estimates of the reduced form coefficients are

$$\begin{bmatrix} 5 & 10 & 2 \\ 10 & 10 & 5 \end{bmatrix}$$

2. The estimates of variance of the errors of the coefficients in the first reduced form equation are 1, 0.5, 0.1.
3. The corresponding covariances are all estimated to be zero.
4. The estimate of the variance of the error in the first reduced form equation is 2.0.

Use this information to reconstruct the 2SLS equations for the estimates of the coefficients of the first structural equation, and compute these estimates.

9.13.

$$y_{1t} = \beta_{12}y_{2t} + \beta_{13}y_{3t} + \gamma_{11}x_{1t} + u_{1t}$$

is one equation in a three-equation model that contains three other exogenous variables. Observations give the following matrices:

$$\boldsymbol{Y'Y} = \begin{bmatrix} 20 & 15 & -5 \\ 15 & 60 & -45 \\ -5 & -45 & 70 \end{bmatrix} \qquad \boldsymbol{Y'X} = \begin{bmatrix} 2 & 2 & 4 & 5 \\ 0 & 4 & 12 & -5 \\ 0 & -2 & -12 & 10 \end{bmatrix} \qquad \boldsymbol{X'X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

Obtain 2SLS estimates of the parameters of the equation and estimate their standard errors (on the assumption that the sample consisted of 30 observation points).

CHAPTER 10

Generalized Method of Moments

We now turn our attention to a class of estimators with desirable asymptotic or *large-sample* properties: generalized method of moments (GMM) estimators. Most of the estimators discussed in this text are special cases of GMM. As much of the literature is beyond the technical level of this book, we focus on presenting the ideas as simply as possible and leave the interested reader to the cited references for details on the asymptotic theory involved.

There has been an explosion of macroeconomic and microeconomic research using GMM estimators in the past decade, especially since Hansen's seminal paper in 1982.[1] There are two reasons for its current popularity:

1. GMM nests many common estimators and provides a useful framework for their comparison and evaluation.
2. GMM provides a "simple" alternative to other estimators, especially when it is difficult to write down the maximum likelihood estimator.

However, in econometrics (as in life) there is no free lunch, and these features come at a price.

First, GMM is a *large-sample* estimator. That is, its desirable properties are likely to be achieved only in very large samples. Typically, GMM estimators are asymptotically efficient in a large class but are rarely efficient in finite samples. Second, these estimators often require some small amount of programming to implement, although with a little cleverness they can sometimes be coaxed out of standard statistical software packages that do not have an explicit GMM estimation program.

[1] L. Hansen, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, **50**, 1982, 646–660.

10.1 THE METHOD OF MOMENTS

A useful starting point is the so-called method of moments (MOM), which we will generalize later. Although you may not have been aware of this fact, much of this book has focused on the problem of moments. We will define a population moment γ rather simply as the expectation of some continuous function g of a random variable x:

$$\gamma = E[g(x)]$$

The most commonly discussed moment is the mean $\mu_{\prime}$, where $g(\cdot)$ is merely the identity function:

$$\mu_{\prime} = E(x)$$

Traditionally, MOM considers powers of x. The mean $\mu_{\prime}$ is sometimes called the first moment, and

$$\mu_{\prime\prime} = E(x^2)$$

is sometimes called the uncentered second moment.

It is a small step from these moments to other more interesting characteristics of populations. As shown in Appendix B, the variance can be expressed as a function of the two moments we have just defined:

$$\text{var}(x) = E(x^2) - (E[x])^2 \tag{10.1}$$

$$= \mu_{\prime\prime} - \mu_{\prime}^2 \tag{10.2}$$

Following conventional usage, we also call functions of moments, such as var(x), moments.

So far we have been talking about characteristics of populations; hence we have been concerned with *population moments*. To see how this discussion might be useful in estimation we need to define another kind of moment—a *sample* moment. The sample moment is merely the *sample* version of the population moment in a particular random sample:

$$\hat{\gamma} = \frac{1}{n}\sum g(x) \tag{10.3}$$

We can easily construct sample analogs to the populations. In a particular sample, the analog to the mean is merely

$$\hat{\mu}_{\prime} = \frac{1}{n}\sum x \tag{10.4}$$

Likewise, the sample analog to the population second moment is

$$\hat{\mu}_{\prime\prime} = \frac{1}{n}\sum x^2 \tag{10.5}$$

Now that we have defined population and sample moments, what is MOM? MOM is merely the following proposal:

> To estimate a population moment (or a function of population moments) merely use the corresponding sample moment (or functions of sample moments).

Before we explain why this approach might be a good idea, let us illustrate MOM estimation with a simple example. Suppose we are interested in estimating the variance of x. The MOM proposal is to replace the *population* moments in Eq. (10.1) with the corresponding *sample* moments. Thus, the MOM estimator of the variance is merely

$$\widehat{\text{var}(x)} = \frac{1}{n}\sum x^2 - \left[\frac{1}{n}\sum x\right]^2 \tag{10.6}$$

A little rearranging shows that this estimator is similar to our usual estimator of the variance,

$$\widehat{\text{var}(x)} = \frac{1}{n}\sum x^2 - \left[\frac{1}{n}\sum x\right]^2 \tag{10.7}$$

$$= \frac{1}{n}\sum (x - \bar{x})^2 \tag{10.8}$$

$$\approx \frac{1}{n-1}\sum (x - \bar{x})^2 \tag{10.9}$$

where Eq. (10.9) is our usual (unbiased) estimate of the variance. Note that the MOM estimator is biased, as it divides the sum of squared deviations from the mean by n instead of by $n - 1$ as in the unbiased estimator in Eq. (10.9). On the other hand, the difference between the two estimators nears zero as the sample grows large: the MOM estimator is consistent.

Alternatively, we could have begun with the conventional definition of the population variance and substituted sample analogs directly:

$$\text{var}(x) = E[x - E(x)]^2 \tag{10.10}$$

The sample analog would be merely

$$\widehat{\text{var}(x)} = \frac{1}{n}\sum [x - \bar{x}]^2 \tag{10.11}$$

where we replaced $E[x]$ in brackets with its sample analog $\bar{x}$. As we will show later, the MOM principle, apart from being intuitive, also produces estimators with desirable large-sample properties. Hence, it is not a surprise that in this simple example our MOM estimator looks similar to the conventional estimator.

10.2 OLS AS A MOMENT PROBLEM

The usefulness of GMM comes from the fact that the object of interest in many estimation exercises is simply a function of moments. To illustrate, let us begin with the simple linear regression:

$$y = X\beta + \epsilon \tag{10.12}$$

where ϵ is distributed as $Q(0, \sigma^2)$. Q is some distribution (not necessarily normal), ϵ has mean zero and variance σ^2, and X is $(n \times k)$. Let us further assume that the model is correctly specified, and as a consequence

$$E(X'\epsilon) = \mathbf{0} \tag{10.13}$$

The condition given by Eq. (10.13) is so important that we will discuss what it means in more detail shortly. For the moment, it will suffice to recall that the condition holds as a consequence of our model being correctly specified.

It may not be obvious how to proceed to obtain an estimate of β. Note, however, that in the population

$$E[X'(y - X\beta)] = \mathbf{0} \tag{10.14}$$

where we have simply used the fact that $\epsilon = y - X\beta$. We now have an interesting problem. By assumption $E[X'(y - X\beta)] = \mathbf{0}$. However, we do not know what β is. The MOM principle suggests that we replace the left-hand side of Eq. (10.14), known as a moment or orthogonality condition, with its sample analog

$$\frac{1}{n}X'(y - X\beta) \tag{10.15}$$

Furthermore, since we know that the true β sets the *population* moment equal to zero in expectation, it seems reasonable to assume that a good choice of $\hat{\beta}$ would be one that sets the *sample* moment to zero. That turns out to be correct. The MOM procedure suggests an estimate of β that solves

$$\frac{1}{n}X'(y - X\hat{\beta}) = \mathbf{0} \tag{10.16}$$

Note that in writing the problem we have generalized the MOM by allowing the moment to depend on unknown parameters, in this case β. The solution turns out to be easy in this case; this is merely a set of k simultaneous equations with k unknown parameters. Hence we can find a unique solution for $\hat{\beta}$ that satisfies Eq. (10.16) exactly (provided that X has full column rank). Rewriting Eq. (10.16), we see that the MOM estimate is

$$\hat{\beta}_{\text{MOM}} = (X'X)^{-1}X'y \tag{10.17}$$

But this is just the OLS estimator for β!

10.3 INSTRUMENTAL VARIABLES AS A MOMENT PROBLEM

Let us now consider a slightly more difficult problem. Consider the following model:

$$y = \alpha + x_1\beta_1 + \epsilon \tag{10.18}$$

where we suspect that $E(x_1'\epsilon) \neq 0$; and, to keep things clear, x_1 is a $n \times 1$ vector. If this is the case, OLS will be inconsistent. As we learned from our discussion of instrumental variables, one proposal is to find instrumental variables Z that are correlated with x_1 but uncorrelated with ϵ; that is, $E(Z'\epsilon) = \mathbf{0}$. To illustrate, here are some examples:

- Suppose y is hourly wages, x_1 is veteran status, and the instrumental variables Z are month and day of birth. The hypothesis to be tested is that veterans experience positive discrimination. That is, given a set of characteristics related to productivity, a veteran receives *higher* wages than a nonveteran. A potential problem that arises in using OLS is that veterans differ from nonveterans in ways that are unobserved by the econometrician. Thus $E(x_1'\epsilon) \neq 0$. In the Vietnam War, the military drafted men based on randomly chosen dates of birth (this procedure was called a draft lottery). Hence, for people who were of draft age during the war, one's date of birth was directly related to the probability one became a veteran (and is presumably unrelated to wages). In this case, month and date of birth may be appropriate instrumental variables.[2]
- Suppose y is the log of firm employment, and x_1 is contract wages. Ideally, one would like to estimate a labor demand curve, but the problem is that employment and wages are the product of both supply and demand changes. Hence, $E(x_1'\epsilon) \neq 0$. Since contract wages are negotiated in advance, a possible instrumental variable is *unexpected inflation*. Since by definition unexpected inflation is not known to either the union or the firm at the time the contract is signed, it shifts the real wage. If inflation was unexpectedly high, for example, this would lower the real wage and employers would move down their labor demand curve.[3]

Assume that we have found two instrumental variables, which we denote by z_1, z_2; we also include the constant 1 as an instrument for itself. We can put these into matrix form and get

$$X = \begin{bmatrix} \mathbf{1} & x_1 \end{bmatrix}$$

$$Z = \begin{bmatrix} \mathbf{1} & z_1 & z_2 \end{bmatrix}$$

It will also be convenient to partition the parameters of the model as well, that is,

$$\boldsymbol{\beta} = \begin{bmatrix} \alpha & \beta_1 \end{bmatrix}$$

The orthogonality condition for this problem is $E(Z'\epsilon) = 0$, so the procedure we developed before suggests a good estimate would be one that sets the sample moment to zero, namely,

$$\frac{1}{n}Z'(y - X\hat{\boldsymbol{\beta}}) = \mathbf{0} \tag{10.19}$$

Given our OLS example, one might be tempted to try to estimate $\hat{\boldsymbol{\beta}}$ with

[2] J. Angrist, "Lifetime Earnings and the Vietnam Era Draft Lottery—Evidence from Social Security Administrative Records," *American Economic Review*, **80**, 1990, 313–336.

[3] D. Card, "Unexpected Inflation, Real Wages, and Employment Determination in Union Contracts," *American Economic Review*, **80**, 1990, 669–688.

10.4 GMM AND THE ORTHOGONALITY CONDITION

We are now ready to turn to a more careful analysis of GMM estimators. The elements are these:

1. "Theory" or a priori information yields an assertion about a population *orthogonality condition*, which is usually of the form $E[g(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta})] = \boldsymbol{0}$, where $g(\cdot)$ is some continuous function of data $(\boldsymbol{y}, \boldsymbol{X})$ and parameters $\boldsymbol{\theta}$.
2. We construct the sample analog $m(\boldsymbol{\theta})$ to the population orthogonality condition and minimize the following with respect to $\widehat{\boldsymbol{\theta}}$:
$$m(\boldsymbol{y}, \boldsymbol{X}, \hat{\boldsymbol{\theta}})' \cdot \boldsymbol{W} \cdot m(\boldsymbol{y}, \boldsymbol{X}, \hat{\boldsymbol{\theta}}) \tag{10.25}$$
where $\boldsymbol{W}$ is best chosen to be a consistent estimate of $\text{var}[m(\cdot)]^{-1}$ as in the White covariance matrix discussed in Chapter 6 or, in the time series context, the appropriate Newey-West covariance matrix.[4]
3. If the optimal $\boldsymbol{W}$ is chosen, the minimized value of the quadratic form in Eq. (10.25) is asymptotically distributed as χ^2 with degrees of freedom equal to the excess of moment conditions R over parameters k under the null hypothesis that the moment conditions are satisfied. This turns out to be extremely useful, especially for problems similar to (linear or nonlinear) 2SLS or 3SLS.

Point (1), the orthogonality condition, is particularly important. Consider again the simple OLS model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{10.26}$$

Recall that the canonical OLS model was introduced with several stringent conditions: The model had to include all the relevant variables, the error terms were homoscedastic and distributed normally, etc. Unfortunately, these conditions are rarely met in practice. Fortunately, not all of them are required. If we limit our attention to the consistency of $\boldsymbol{\beta}$, we have already learned in Chapter 5, for example, that we can dispense with homoscedastic errors. The requirement that the model include *all* of the relevant variables is quite stringent and is unlikely to be satisfied in practice. A reasonable question is, how many variables are *enough* to get reliable estimates? This question is not easy to answer, but the GMM approach makes it clear what conditions need to be satisfied for large samples.

In particular, we can dispense with normality provided the error term has a zero mean. More important, however, is the requirement imposed by the moment restriction $E(\boldsymbol{X}'\boldsymbol{\epsilon}) = \boldsymbol{0}$. To see what this implies, consider the most classic estimation design: the controlled experiment. Suppose we have discovered a new treatment to aid in quitting the smoking habit. We get a sample of $m = 2n$ smokers and *randomly* assign the treatment T to half of the sample. The other half of the sample receives a placebo treatment—something that looks like the treatment but is actually inert.

[4] The Newey-West estimator provides a way to calculate consistent covariance matrices in the presence of both serial correlation *and* heteroscedasticity. See the lucid discussion in Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics*, Chapter 17.5. See also William H. Greene, *Econometric Analysis*, 2nd edition, p. 377–378.

Classical experimental design suggests that a good estimate of the efficacy of the treatment is to compare the proportion of smokers in the two groups at the end of the program or

$$\text{Treatment effect} = \bar{y}^t - \bar{y}^c$$

where c and t refer to the placebo (control) group and the treatment group, respectively, and $\bar{y}^t = (1/n)\sum_{j=1}^{n} y_j^t$ and $\bar{y}^c = (1/n)\sum_{j=n+1}^{m} y_j^c$. It is easy to see that we can recast this as a simple OLS regression:

$$y = \alpha + \beta x + \epsilon \tag{10.27}$$

where x is a dummy variable that equals 1 if the subject gets the treatment and 0 if the subject gets the placebo. Likewise, y is an indicator variable that equals 1 if the subject is cured, and 0 otherwise. It is a good exercise to show that the OLS estimate of β in this case is given by

$$\hat{\beta}_{\text{OLS}} = \bar{y}^t - \bar{y}^c$$

This problem surely does not meet the rather stringent conditions of the canonical OLS model that we introduced in previous chapters. The errors are certainly not normal (the student should verify that the error can take on only four values: $-\alpha$, $1-\alpha$, $-\alpha-\beta$, or $1-\alpha-\beta$), and certainly there are other determinants of smoking besides the treatment! Is our confidence in classical experimental design misplaced? The answer is no, for classical experimental design works by ensuring that even if all the relevant variables are not included, these relevant variables are uncorrelated with our right-hand side variable, x. In other words,

$$E(x'\epsilon) = 0$$

where ϵ is understood to include the potentially relevant but unincluded variables. The reason so much care is given to *random assignment,* that is, dispersing the treatment to subjects *randomly,* is to ensure this orthogonality condition holds.

To see this point more clearly, suppose the true model was

$$y = \alpha + \beta x + \gamma z + \epsilon \tag{10.28}$$

where all is as before, and z is some variable like "years of smoking," so $\gamma < 0$ presumably. If we let $\Phi = \gamma z + \epsilon$, classical experimental design amounts to running the following regression:

$$y = \alpha + \beta x + \Phi \tag{10.29}$$

The salient question is whether $E(x'\Phi) = 0$.[5] It is sufficient to evaluate whether $E(\Phi \mid x = 1) = E(\Phi \mid x = 0)$, in which case the orthogonality condition is still satisfied. If we assume that ϵ is random noise, this is equivalent to asking whether $E(z \mid x = 1) = E(z \mid x = 0)$. That is, on average have the people who receive the treatment been smoking as long as those who get the placebo? Because the essence of randomized design is that there is no systematic difference between the groups

[5] We assume that α is not a parameter of interest, so that if, for example, the mean of Φ is not zero, it gets "absorbed" into the constant.

receiving or not receiving the treatment, we see that for a well-designed experiment the orthogonality condition is satisfied, and there is no bias.

Suppose we were suspicious that the experiment had been done incorrectly. Is there any way we could *test* whether x was truly uncorrelated with the errors? In this simple example, there is no test because the sample moment condition that produces our estimates of the treatment effect, letting $X = [\mathbf{1}\ \mathbf{x}]$,

$$\frac{1}{n}X'(y - X\hat{\beta}) = 0$$

has only one answer, and it exactly sets the sample moment to zero. In other words, there are no *over*identifying restrictions to check.[6] As we will see shortly, one of the advantages of 2SLS is that it allows us to test some of these restrictions.

10.5 DISTRIBUTION OF THE GMM ESTIMATOR

Before we turn to some applications, let us derive the distribution of the GMM estimator. This discussion will be very heuristic. The interested reader may wish to consult Hansen or Hall for a formal treatment, although these articles are quite demanding.[7]

Let us suppose that we have a well-defined moment condition or orthogonality condition of the form

$$E[g(y, X, \boldsymbol{\theta}_0)] = 0 \tag{10.30}$$

where y, X refers to the data and $\boldsymbol{\theta}_0$ to the unique value of a set of parameters that makes the expectation equal to zero. We then collect a random sample. For a given sample, the GMM estimator minimizes the following with respect to the parameters $\boldsymbol{\theta}$. The estimator $\hat{\boldsymbol{\theta}}$ is merely the solution to

$$\min_{\hat{\theta}}\left(m(y, X, \hat{\boldsymbol{\theta}})' \cdot W_n \cdot m(y, X, \hat{\boldsymbol{\theta}})\right) \tag{10.31}$$

where $m(y, X, \boldsymbol{\theta}) = (1/n)\sum_1^n g(y_i, X_i, \boldsymbol{\theta})$, and the subscript on W indicates that it can be a function of the data. We also assume that W_n is positive definite and symmetric and converges in probability to some matrix W that is also symmetric and positive definite. Provided that, in the limit, the true value of the parameters $\boldsymbol{\theta}_0$ minimize Eq. (10.31) and suitable regularity conditions[8] hold, the estimator produced by Eq. (10.31) is consistent.

We find $\hat{\boldsymbol{\theta}}$ by solving the following first-order condition:

$$\frac{\partial m(y, X, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'} \cdot W_n \cdot m(y, X, \hat{\boldsymbol{\theta}}) = 0 \tag{10.32}$$

[6] Note that in many classical randomized trials, the researchers will compare the characteristics of the treatment and control groups. Often this comparison is used to verify that the randomization was done properly. If designed properly, the mean characteristics of the two groups should be the same on average.

[7] A. Hall,"Some Aspects of Generalized Method of Moments Estimation," Chapter 15, *Handbook of Statistics,* Vol. 11, 1993, Elsevier.

[8] These are technical conditions that allow one to establish asymptotic results.

Denote the first derivative matrix as $G = \partial m/\partial \boldsymbol{\theta}$. Because the estimator produced by minimizing Eq. (10.31) is consistent, $G(\hat{\boldsymbol{\theta}})$ converges in probability to $G(\boldsymbol{\theta}_0)$. We have already assumed that W_n converges in probability to W, so that

$$\operatorname{plim} G(\hat{\boldsymbol{\theta}}) \cdot W_n = G(\boldsymbol{\theta}_0) \cdot W \tag{10.33}$$

The distribution of $\hat{\boldsymbol{\theta}}$ is derived by taking a Taylor series approximation of $g(\hat{\boldsymbol{\theta}})$ around the truth, $\boldsymbol{\theta}_0$, from the first-order condition in Eq. (10.32). Given certain regularity conditions, the distribution of $\hat{\boldsymbol{\theta}}$ can be shown to be

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} \mathrm{N}(\boldsymbol{\theta}_0, (G'WG)^{-1}G'W\boldsymbol{\Omega}WG(G'WG)^{-1}) \tag{10.34}$$

where $\boldsymbol{\Omega} = E[g(\boldsymbol{\theta}_0)g(\boldsymbol{\theta}_0)']$, or, since $E[g(y, X, \boldsymbol{\theta}_0)] = 0$, this is merely the variance of the moment condition. Hansen (1982) showed that an optimal choice for W is merely a heteroscedasticity (and autocorrelation) consistent estimate of $E[g(\boldsymbol{\theta}_0)g(\boldsymbol{\theta}_0)']^{-1} = \boldsymbol{\Omega}^{-1}$. Given a consistent estimate $\hat{\boldsymbol{\theta}}$, an estimate $\boldsymbol{\Omega}^{-1}$ is obtainable. In this special case with an optimal W_n, Eq. (10.34) becomes:

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} \mathrm{N}(\boldsymbol{\theta}_0, (G'\boldsymbol{\Omega}^{-1}G)^{-1}) \tag{10.35}$$

The student can verify that any other choice of W leads to a covariance matrix that exceeds the optimal choice by a positive definite matrix. Regardless of the weighting matrix that is used, GMM is always consistent and asymptotically unbiased. When the correct weighting matrix is used, GMM is also asymptotically efficient in the class of estimators defined by the orthogonality conditions.

10.6 APPLICATIONS

Moment conditions can be very general. In this section we go through some simple examples of estimation and testing with GMM.

10.6.1 Two-Stage Least Squares, and Tests of Overidentifying Restrictions

One of the reasons for the popularity of GMM is that it allows for a clear procedure to *test* restrictions that come out of well-specified econometric models. The leading case is 2SLS.

Recall from Eq. (10.22) that our moment condition $E(Z'\boldsymbol{\epsilon}) = \mathbf{0}$ led us to a GMM estimator that solved the following:

$$\min_{\hat{\boldsymbol{\beta}}} \left(\frac{1}{n}[Z'(y - X\hat{\boldsymbol{\beta}})]' \cdot W_n \cdot \frac{1}{n}[Z'(y - X\hat{\boldsymbol{\beta}})] \right) \tag{10.36}$$

where we will generalize the example so that Z is $(n \times L)$, X is $(n \times k)$, W_n is an $(L \times L)$ weighting matrix, and $L > k$. Note that Z and X may have columns in common. So far we have left the issue of choosing W_n in the background; we now turn to this issue. Recall that a good choice of W_n should be an estimate of the inverse of the asymptotic variance matrix of our moment condition, $[\text{var}(1/n)(Z'\epsilon)]^{-1}$, which we can denote $[(1/n^2)\widehat{Z'\Omega Z}]^{-1}$. How might we get a consistent estimate of this? It would seem that we require consistent estimates of the ϵ_i's. But these are impossible since the number of ϵ_i's to estimate increases at the same rate as the sample. Fortunately, although the dimension of ϵ does increase with n, the dimension of $(1/n)Z'\epsilon$ does not.

The procedure works in two steps:

1. First, generate a consistent estimate $\hat{\beta}_c$ of β. This can be done in several ways. One way is first to do ordinary 2SLS, which amounts to using $(Z'Z)^{-1}$ for W_n in the first step. Fortunately, GMM produces consistent estimates with *any* positive definite weighting matrix. For example, another choice (often used when the problem is nonlinear) is just the identity matrix.
2. With an estimate $\hat{\beta}_c$ in hand, compute the residuals, which in this case are $r \equiv y - X\hat{\beta}_c$. Provided that observations are independent (which is typically assumed in cross-section data) a White estimate of $[(1/n^2)Z'\Omega Z]^{-1}$ is simply

$$W_n = \left(\frac{1}{n^2}\sum\nolimits_i z_i z_i' r_i^2\right)^{-1} \tag{10.37}$$

where z_i are the columns of Z.

With our estimate W_n in hand, we then return to our original minimization problem:

$$\min_{\hat{\beta}_{\text{GMM}}} \left(\frac{1}{n}[Z'(y - X\hat{\beta}_{\text{GMM}})]' \cdot W_n \cdot \frac{1}{n}[Z'(y - X\hat{\beta}_{\text{GMM}})]\right) \tag{10.38}$$

Following the same logic for Eq. (10.24) and letting $\widehat{Z'\Omega Z} \equiv \sum_i z_i z_i' r_i^2$, we get

$$\hat{\beta}_{\text{GMM}} = [X'Z(\widehat{Z'\Omega Z})^{-1}Z'X]^{-1}X'Z(\widehat{Z'\Omega Z})^{-1}Z'y \tag{10.39}$$

Recall from Eq. (10.24) that with homoscedastic errors GMM and 2SLS are the same. In the presence of heteroscedastic errors, however, the GMM estimator differs from the 2SLS estimator in general, and 2SLS is asymptotically less efficient than GMM. The estimator in Eq. (10.39) is often referred to as the ***generalized 2SLS estimator.*** (As an exercise, the student should show that the GMM estimator and the 2SLS estimator are equivalent when the model is *exactly identified,* that is, when the column rank of Z equals the column rank of X.)

The GMM approach yields some additional benefits. If $L > k$, $\hat{\beta}_{\text{GMM}}$ is overidentified. That is, the number of moment restrictions L is greater than the number of parameters. In this case the minimand is also a test statistic for the validity of these restrictions. Under the null that these restrictions are valid,

$$\text{Test}_{\text{GMM}} \equiv \left[Z'(y - X\hat{\beta}_{\text{GMM}})\right]' \cdot \left(\sum_i z_i z_i' r_i^2\right)^{-1} \cdot \left[Z'(y - X\hat{\beta}_{\text{GMM}})\right] \overset{a}{\sim} \chi^2(L-k) \tag{10.40}$$

It has been noted that when the errors are homoscedastic and serially independent, the test defined in Eq. (10.40) has a particularly simple form:

$$\text{Test}_{\text{GMM}} \equiv nR^2 \tag{10.41}$$

where the test statistic is merely the number of observations times the uncentered R^2 from the regression of $\hat{r}$ on Z:

$$\hat{r} = Z\pi + \text{error}$$

where $$\hat{r} \equiv y - X\hat{\beta}_{\text{GMM}}$$

The intuition is clear. If $E(Z'\epsilon) = 0$, then it seems reasonable to suspect that the instrumental variables should be orthogonal to the residuals. (Note that this does not work for OLS, for example, because the "instruments"—the Xs—are orthogonal to the residual by construction.) If they are, the R^2 from the regression will be low, and we will accept the hypothesis that the overidentifying restrictions are valid.[9] The test in Eq. (10.40) is often misunderstood. It is *not* a test for whether all the instrumental variables are "valid." Instead the test answers the question: given that a subset of the instrumental variables is valid and exactly identifies the coefficients, are the "extra" instrumental variables valid?

We note in passing that the GMM estimator and the corresponding test would be of the same form even if the model were nonlinear. For a general nonlinear regression

$$y = f(X, \beta) + \epsilon \tag{10.42}$$

$[Z'(y - X\hat{\beta}_{\text{GMM}})]$ would be replaced with $[Z'(y - f(X, \hat{\beta}_{\text{GMM}}))]$, and the proper weighting matrix would be computed.

10.6.2 Wu-Hausman Tests Revisited

An interesting class of specification tests that are closely related to GMM tests are called ***Hausman***, or ***Wu-Hausman***, or ***Durbin-Wu-Hausman*** tests. Three standard citations for this literature are Durbin, Wu, and the very influential paper by Hausman.[10] The relationship between Hausman tests and GMM tests is explored in Newey.[11]

These tests appear in the literature in a variety of situations. We will discuss them solely in the context of 2SLS and OLS, although they are actually quite general. The standard case involves evaluating the moment conditions that define an estimator.

Consider the usual model,

$$y_1 = \beta y_2 + \epsilon_1 \tag{10.43}$$

[9] An asymptotically equivalent version of this test can also be found in R. Basmann, "On Finite Sample Distributions of Generalized Classical Linear Identifiability Test Statistics," *Journal of the American Statistical Association, 55*, 1960, 650–659. It is often referred to as the "Basmann" test.

[10] J. Durbin, "Errors in Variables," *Review of the International Statistical Institute, 22*, 1954, 23–32; D. Wu, "Alternative Tests of Independence between Stochastic Regressors and Disturbances," *Econometrica, 41*, 1973, 733–750; J. Hausman, "Specification Tests in Econometrics," *Econometrica, 46*, 1978, 1251–1271. See also Chapter 8.2.5.

[11] W. Newey, "Generalized Method of Moments Specification Testing,"*Journal of Econometrics, 29*, 1985, 229–256.

If $E(y_2'\epsilon_1) = 0$, we have seen that the GMM estimator (which is merely OLS) produces consistent estimates of β. Now we may have reason to suspect that y_2 is endogenous, or "contaminated," so that the orthogonality condition does not hold. If we have a set of instrumental variables that are orthogonal to ϵ_1, we can construct a 2SLS estimator of β that is consistent whether or not y_2 is correlated with ϵ_1. If $E(y_2'\epsilon_1) = 0$, however, the 2SLS estimator remains consistent but is less efficient than OLS. It would be useful then to devise a test to assess whether OLS is adequate.

Hausman (1978) suggests the following test:

$$h \equiv (\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{2SLS}})'(\text{var}(\hat{\beta}_{\text{2SLS}}) - \text{var}(\hat{\beta}_{\text{OLS}}))^{-1}(\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{2SLS}}) \overset{a}{\sim} \chi^2(g) \tag{10.44}$$

where g, the number of potentially endogenous regressors, is 1 in our example. Hausman (1978) showed that the term in the middle of Eq. (10.44)—the difference between the covariance matrix of the coefficients estimated under 2SLS and the covariance matrix of the coefficients estimated under OLS—takes this particularly convenient form, where one does not have to compute the covariance between the two estimators.

If the difference between our two estimates is large, we would *reject* the adequacy of OLS. It is for this reason that the test is sometimes discussed as if it were a test of the "endogeneity" of y_2. As we will see, that is not quite right. The test evaluates whether the endogeneity has any effect on the consistency of β.

To give some insight into this test, let us consider Eq. (10.43) as part of a two-equation system that includes the following, where Z is a matrix of instrumental variables:

$$y_2 = Z\delta + \epsilon_2 \tag{10.45}$$

Given the assumptions we have made so far, Eq. (10.45) amounts to partitioning the variance of y_2 into two parts. One part, $Z\delta$, is uncorrelated with ϵ_1, the error in y_1. The other part, ϵ_2, is *possibly* correlated with ϵ_1. The proposed test therefore can be thought of as a test of whether $\text{cov}(\epsilon_1, \epsilon_2) = 0$.

Some additional insight can be acquired by considering an alternative development of the Hausman test. The paper by Davidson and MacKinnon, on which this discussion is based, is very helpful.[12] Two other papers that explore Hausman tests are those by Ruud and by Davidson and MacKinnon and the references cited therein.[13] In many cases, it is straightforward to compute Hausman tests from simple artificial regressions.

It is useful to consider the Hausman test as a ***vector of contrasts.*** Consider the canonical linear model

$$y = X\beta + \epsilon \tag{10.46}$$

where ϵ has mean zero and variance σ^2, and X is $(n \times k)$ and y and ϵ are both $(n \times 1)$.

[12] R. Davidson and J. MacKinnon, "Testing for Consistency Using Artificial Regressions," *Econometric Theory*, **5**, 1989, 363–384.

[13] P. Ruud, "Tests of Specification in Econometrics," *Econometric Reviews*, **3**, 1984, 211–242; R. Davidson and J. MacKinnon, "Specification Tests Based on Artificial Regressions," *Journal of the American Statistical Association*, **85**, 1990, 220–227.

We wish to compare one estimator of $\boldsymbol{\beta}$ for this model, say OLS,

$$\boldsymbol{\beta}_{\text{OLS}} = (X'X)^{-1}X'y \tag{10.47}$$

to another estimator, say $\boldsymbol{\beta}_A$,

$$\boldsymbol{\beta}_A = (X'AX)^{-1}X'Ay \tag{10.48}$$

where $\boldsymbol{A}$ is a symmetric $(n \times n)$ matrix with rank no less than k. We shall describe $\boldsymbol{A}$ in a moment.

We make this comparison by computing the vector of contrasts:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}_{\text{OLS}} &= (X'AX)^{-1}X'Ay - (X'X)^{-1}X'y \\ &= (X'AX)^{-1}\left[X'Ay - X'AX(X'X)^{-1}X'y\right] \\ &= (X'AX)^{-1}X'A\left[I - X(X'X)^{-1}X'\right]y \\ &= (X'AX)^{-1}X'AM_Xy\end{aligned} \tag{10.49}$$

where $\boldsymbol{M}_X = \boldsymbol{I} - \boldsymbol{P}_X = \boldsymbol{I} - X(X'X)^{-1}X'$ is the familiar $(n \times n)$ symmetric, idempotent "residual-maker" matrix, and $\boldsymbol{P}_X$ is the familiar "predicted value-maker" matrix. That is, for some $(n \times l)$ matrix $\boldsymbol{Z}, \boldsymbol{M}_Z\boldsymbol{X}$ is the $(n \times k)$ matrix of residuals from a regression of each column of $\boldsymbol{X}$ on the $\boldsymbol{Z}$s; $\boldsymbol{P}_Z\boldsymbol{X}$ is the $(n \times k)$ matrix of predicted values from a regression of each column of $\boldsymbol{X}$ on $\boldsymbol{Z}$.

The choice of A will depend on the problem that the researcher is facing. When $A = \boldsymbol{P}_Z$, $\hat{\boldsymbol{\beta}}_A$ is the two-stage least squares estimator of $\boldsymbol{\beta}$ using $\boldsymbol{Z}$ as the instrumental variables. For the fixed effects estimator (see Chapter 12), $\boldsymbol{A} = \boldsymbol{M}_D$, where $\boldsymbol{D}$ is the set of dummy variables for the cross-section units, and $\boldsymbol{M}_D$ is the matrix that therefore transforms the data into deviations from individual-specific means.

If the model in Eq. (10.46) is correct then the probability limit of the difference in Eq. (10.49) will be zero. More importantly, since $(X'AX)^{-1}$ is just a $(k \times k)$ matrix of full rank, the vector of contrasts will have a probability limit of zero whenever

$$\text{plim}\,\frac{1}{n}(X'AM_Xy) = \mathbf{0} \tag{10.50}$$

Consider comparing OLS and 2SLS. In this case we can partition the matrix $\boldsymbol{X}$ as $[X_1\ X_2]$ where X_1 is an $(n \times g)$ submatrix of the potentially endogenous regressors and X_2 is the $[n \times (k-g)]$ submatrix of exogenous right-hand side regressors. We have a set of instruments $\boldsymbol{Z} = [Z^*\ X_2]$, where $\boldsymbol{Z}$ is our $(n \times l)$ matrix of identifying instruments, where $(l \geq k)$, so that our $\boldsymbol{A}$ matrix is merely $\boldsymbol{A} = \boldsymbol{P}_Z$.

We are interested in whether

$$\text{plim}\,\frac{1}{n}(X'P_ZM_Xy) = \mathbf{0} \tag{10.51}$$

It is evident that several columns of $\boldsymbol{X'P_ZM_X}$ will be identical to zero. This fact can be seen by noting that

$$X'P_Z = \begin{bmatrix}\hat{X}'_1 \\ \hat{X}'_2\end{bmatrix} \tag{10.52}$$

where the $\hat{X}_1$ signifies the predicted value matrix from a regression of the columns X_1 on Z. Since X and Z have X_2 in common, it is clear that $\hat{X}_2 = X_2$. In that case, note that

$$\begin{bmatrix} \hat{X}_1' \\ X_2' \end{bmatrix} M_X \tag{10.53}$$

will be identical to zero for those rows corresponding to x_2, because the residuals from a regression of x_2 on X will be identical to zero.

We can therefore restrict our attention to determining whether

$$\operatorname{plim} \frac{1}{n} X_1' P_Z M_X y = \operatorname{plim} \frac{1}{n} \hat{X}_1' M_X y = \mathbf{0} \tag{10.54}$$

We can perform this test by performing an F test on $\boldsymbol{\delta}$ in the artificial regression

$$y = X\boldsymbol{\beta} + \hat{X}_1\boldsymbol{\delta} + \text{residuals} \tag{10.55}$$

Denoting the model $\boldsymbol{\delta} = \mathbf{0}$ as the restricted model, the familiar F test is merely

$$H = \frac{(\text{RSS}_r - \text{RSS}_u)/g}{\text{RSS}_u/(n - k - g)} \tag{10.56}$$

We noted previously that the Hausman test is often interpreted as a test of whether the columns of x_1 are endogenous, whereas it may be more appropriate to interpret it as a test of whether the "endogeneity" has any significant effect on estimates of $\boldsymbol{\beta}$. The latter interpretation can be seen most easily by considering an omitted variables version of the Hausman test. We begin with the same problem as before. The difference is that instead of evaluating the difference between OLS and 2SLS, we compare one OLS estimator to another OLS estimator that includes the Z^* as additional regressors. In this case, our vector of contrasts compares the OLS estimate from Eq. (10.46) with the OLS estimate of $\boldsymbol{\beta}$ from

$$y = X\boldsymbol{\beta} + Z^*\boldsymbol{\gamma} + \boldsymbol{\nu} \tag{10.57}$$

where Z^* is the instrumental variable matrix from our 2SLS example, less those instruments that are also in X.

We are interested in whether the coefficients on the set of X not included in Z are affected by the inclusion of the additional variables. Recall that the Frisch-Waugh-Lovell theorem[14] allows us to get the right estimates of $\boldsymbol{\beta}$ for this model by first regressing X on Z^*, taking the residuals, and regressing y against these residuals.

In that case, we run an OLS regression of y on X after x has had its predicted value $[P_{Z^*}X = Z^*(Z^{*\prime}Z^*)^{-1}Z^{*\prime}X]$ "removed." The matrix that does this is $M_{Z^*} = I - P_{Z^*}$:

$$y = M_{Z^*}X\boldsymbol{\beta} + \boldsymbol{\nu} \tag{10.58}$$

Since M_{Z^*} is idempotent, the OLS estimate from this model is merely

$$\hat{\boldsymbol{\beta}}_{\text{augmented}} = (X'M_{Z^*}X)^{-1}X'M_{Z^*}y \tag{10.59}$$

[14] See the Appendix Chapter 3.2.

It is now evident that the A matrix of Eq. (10.48) is merely M_Z, so the correct artificial regression becomes

$$\begin{aligned} y &= X\beta + M_{Z^*}X\delta + \text{residuals} \\ y &= X\beta + e_{X_1,Z^*}\delta + \text{residuals} \end{aligned} \qquad (10.60)$$

where e_{X_1,Z^*} are the residuals from running a regression of the g columns of X_1 on Z. The F test of the hypothesis that $\delta = \mathbf{0}$ in this artificial regression is numerically equivalent to the test we previously derived using an artificial regression, but that was based on a comparison of OLS and 2SLS! (You are asked to show this in an exercise.) That is, a comparison of β from our original OLS specification to 2SLS with Z as the instrumental variables yields the same test statistic as the comparison of β from our original OLS specification to the OLS specification augmented with Z^*, though there is no problem of "endogeneity" in this latter case.

In sum, there are three ways to do the Hausman test for a comparison of OLS and 2SLS:

1. Directly compute the vector of contrasts, as in Eq. (10.44).
2. Regress the potentially endogenous regressors on the instruments and compute the predicted value from these regressions. Run OLS on the system including these created variables and test whether they are significant, as in Eq. (10.55).
3. Regress the potentially endogenous regressors on the instruments and compute the residual from these regressions. Run OLS on the system including these created variables and test whether they are significant as in Eq. (10.60).

See Davidson and MacKinnon for a discussion of the last two methods and of the extension to other comparisons.[13]

10.6.3 Maximum Likelihood

Maximum likelihood estimators also have a GMM interpretation. Recall from Chapter 5 that in order to maximize the likelihood, we set the first derivative of the log-likelihood $\partial \ln[L(X, \theta)]/\partial\theta$ (the score) to zero:

$$m(y, X, \theta) \equiv \frac{\partial \ln L}{\partial \theta} = \mathbf{0} \qquad (10.61)$$

This condition is simply a moment condition. If we again take the simplest case, the "GMM way" to write this problem is as a solution to

$$\min_{\hat{\theta}} \left(m(y, X, \theta)' \cdot H^{-1} \cdot m(y, X, \theta)\right)$$

where the so-called *weighting matrix* H is merely the variance of the moment condition. That is, $H = -E(\partial^2 \ln L/\partial\theta\partial\theta')$.

In this simplest case, $\hat{\theta}$ solves the first-order condition to the minimization problem just defined. Thus $\hat{\theta}$ must satisfy

$$\frac{\partial^2 \ln L}{\partial\theta\partial\theta'} H^{-1} \frac{\partial \ln L}{\partial\theta} = \mathbf{0} \qquad (10.62)$$

But this is the equation that defines MLE, hence MLE can be viewed as a GMM estimator.

If this is the case, then why do some researchers prefer GMM to MLE? One reason is tractability. Sometimes when the maximum likelihood estimator is difficult to compute, there is a GMM estimator that, although less asymptotically efficient than the ML estimator, is still consistent and easier to compute. A second reason is that sometimes, although not enough is known about the data generation process to specify the likelihood function completely, enough is known to specify moment conditions for a GMM estimator.

10.6.4 Euler Equations

Another example that is unique to GMM is the so-called Euler equation approach. Euler equations are the first-order conditions of dynamic optimization problems. GMM treats these first-order conditions as moment conditions. An example will illustrate. Suppose the representative consumer has a utility function over consumption each period and tries to maximize

$$E_t\left[\sum_{\tau=0}^{T}(1+\delta)^{-\tau}u(c_{t+\tau})\right] \tag{10.63}$$

subject to

$$A_t = \sum_{\tau=0}^{T}(1+r)^{-\tau}(c_{t+\tau} - w_{t+\tau}) \tag{10.64}$$

where E_t = expectation given information at time t
δ = rate of subjective time preference
r = fixed real rate of interest
T = length of economic life
c_t = consumption at time t
w_t = earnings at time t
A_t = assets at time t

Hall (1978) notes that this model implies an Euler equation (a first-order condition) of the form

$$E_t u'(c_{t+1}) = \gamma u'(c_t) \tag{10.65}$$

where $u'(\cdot)$ is the marginal utility of consumption, and $\gamma = (1+\delta)/(1+r)$.[15] An equivalent way to write Eq. (10.65) is

$$u'(c_{t+1}) = \gamma u'(c_t) + \epsilon_{t+1} \tag{10.66}$$

where ϵ_{t+1} represents the divergence of discounted lagged marginal utility of consumption from its value today. This error term has several properties apart from being mean zero and serially uncorrelated. If $\delta = r$, marginal utility would be a constant except for *new* information arriving between time t and $t + 1$. Hence the error or "innovation" to marginal utility is uncorrelated with information arriving on or before

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

time t. This condition sounds just like an orthogonality condition,

$$E_t Z_t' [u'(c_{t+1}) - \gamma u'(c_t)] = 0 \tag{10.67}$$

where Z_t is *any* information dated time t or earlier. If we assume that the utility function is quadratic in consumption (which implies marginal utility is linear in consumption), we can write the following:

$$c_{t+1} = \beta_0 + \gamma c_t - \epsilon_{t+1} \tag{10.68}$$

A difficult question is, what do we include in Z_t? In principle, we can include any information from time t or earlier, essentially providing us with unlimited numbers of instrumental variables. In practice, however, the test is not very persuasive if we test for whether aggregate sales of water balloons 10 periods ago help predict c_{t+1}. We might try something suggested by other theories of consumption. Duesenberry (1948) suggested that past levels of income or consumption (more than one period ago) might matter, since once people have reached some local peak in their life cycle profile of consumption, they are much more reluctant to consume less and will draw down their savings to maintain consumption.[16] In that case, income may be helpful in predicting consumption.

Staying with the case where the utility function is quadratic, we might construct a $\mathbf{Z}$ matrix as follows:

$$Z_t = [1 \quad c_t \quad y_t]$$

where c_t, y_t are consumption and income. In this example, the GMM test statistic of overidentifying restrictions is

$$\frac{\text{RSS}_R - \text{RSS}_A}{\text{RSS}_R/\text{n}} \overset{d}{\sim} \chi^2(1) \tag{10.69}$$

where RSS_R is the sum of squared residuals from a regression of c_{t+1} on a constant and c_t (the "restricted" model), and RSS_A is the sum of squared residuals from an artificial regression of the restricted model's residuals on y_t. See Problem 10.6.

For purposes of illustration we chose a functional form for utility that led to a linear model. However, one of the nice aspects of GMM is that we could have chosen a functional form for utility that did not result in a linear model. GMM applies *mutatis mutandis* to nonlinear models as well.

10.7 READINGS

Hansen's original paper is a nice starting point for learning about GMM estimation, although it is technically difficult.[1] In addition, it considers some of the time series issues we have ignored here. Davidson and MacKinnon have very nice discussions of Hausman tests and specification tests.[13,14]

[16] J. Duesenberry, "Income-Consumption Relations and Their Implications," *Essays in Honor of Alvin H. Hansen*, by Lloyd A. Metzler and others, W.W. Norton, 1948, 54–81.

	$\Delta g68$	$\Delta g69$	$\Delta g70$	$\Delta g71$	$\Delta g72$	$\Delta g73$	$\Delta h68$	$\Delta h69$	$\Delta h70$	$\Delta h71$	$\Delta h72$	$\Delta h73$
$\Delta g68$	.17											
$\Delta g69$	−.06	.17										
$\Delta g70$	.00	−.06	.17									
$\Delta g71$	.00	.00	−.06	.17								
$\Delta g72$	.00	.00	.00	−.06	.17							
$\Delta g73$	.00	.00	.00	.00	−.06	.17						
$\Delta h68$	.07	−.02	.00	.00	.00	.00	.12					
$\Delta h69$	−.02	.07	−.02	.00	.00	.00	−.03	.12				
$\Delta h70$	.00	−.02	.07	−.02	.00	.00	.00	−.03	.12			
$\Delta h71$	.00	.00	−.02	.07	−.02	.00	.00	.00	−.03	.12		
$\Delta h72$	.00	.00	.00	−.02	.07	−.02	.00	.00	.00	−.03	.12	
$\Delta h73$	.00	.00	.00	.00	−.02	.07	−.02	.00	.00	.00	−.03	.12

leisure is a normal good, the model predicts that $\eta > 0$. It is useful to note that earnings equals hours times wages so that in log terms, $g = w + h$.

(a) Use the foregoing estimates to generate an estimate of η that is positive.

(b) Use these estimates to generate an estimate of η that is negative. How could the existence of measurement error explain this result?

(c) Consider the following "pure measurement error" model for log hours in levels:

$$h_{i,t} = \bar{h}_i + \epsilon_{it}$$

where ϵ_{it} is distributed as $N(0, \sigma^2 \boldsymbol{I}_T)$. Is the evidence in the accompanying covariance matrix of earnings and hours consistent with this model? How would you test this?

CHAPTER 11

A Smorgasbord of Computationally Intensive Methods

In this chapter, we briefly survey several econometric methods that have grown in popularity as a consequence of the great advances in computing technology: Monte Carlo methods, permutation and approximate randomization tests, bootstrapping, nonparametric density estimation, and nonparametric regression.

11.1 AN INTRODUCTION TO MONTE CARLO METHODS

We first consider *Monte Carlo* methods. The object of interest is usually an estimator or a test statistic that has unknown finite sample properties (although we may understand a great deal about its asymptotic properties.) We are interested in how this estimator, for example, two-stage least squares (2SLS), will perform 'in practice." Often the issue is whether the known asymptotic properties of an estimator provide a useful guide for the estimator's (unknown) finite sample properties. Although there may be analytical ways to study the finite sample distribution of estimators, it is often easier to do Monte Carlo experiments. In fact, a Monte Carlo study is often performed when the computational resources of a researcher are more abundant than the researcher's "mental" resources.

Very generally, a Monte Carlo experiment proceeds as follows:

1. Completely specify a "true" model. As an example, if the true model is the standard linear model, this means specifying the distribution of the error term, the explanatory variables, the coefficients, and the sample size.
2. Generate a data set using this true model.
3. Calculate the test statistic or estimator that is being evaluated with this artificially generated sample and store the results.

4. Repeat Steps 2 and 3 a large number of times. Each generation of a new data set is called a *replication*. Later we will discuss how large is "large."
5. Evaluate how well the estimator performs, or how frequently the test statistic rejects or accepts the "true" model in the set of replications.

In other words, if you are uncertain about the properties of a specific estimator or test in practice, you perform an experiment. You thus subject the estimator to a variety of different conditions (different simulated samples) and evaluate the estimator's performance.

There are no strict rules that define what makes a Monte Carlo experiment useful although we will suggest some guidelines below. Sometimes a researcher wishes merely to establish that something is "possible."

Some examples of issues studied with Monte Carlo methods include the following:

1. How well does 2SLS perform when the instrumental variable satisfies the necessary orthogonality conditions but is not highly correlated with the endogenous variable?[1]
2. How well do common tests of the existence of unit roots perform in finite samples?[2]
3. Consider estimating an error components model with panel data (discussed in Chapter 12), where the panel data are unbalanced or incomplete. How much efficiency is lost by applying traditional balanced sample methods on the subpanel that is complete (that is, the panel that results from dropping individuals for whom a complete set of observations is not available) rather than using more complicated models for unbalanced data?[3]

11.1.1 Some Guidelines for Monte Carlo Experiments

Before turning to the mechanics of Monte Carlo experiments there are some guidelines that may be helpful in thinking about them. Like a laboratory-based experiment, a good Monte Carlo experiment typically has these features:

1. Is easy to understand and economical
2. Is relevant to understanding problems with real data
3. Allows one to measure the influence of all the relevant factors
4. Is sufficiently precise for the problem at hand

Guideline 1 simply states that the Monte Carlo results themselves should be easy to understand. As Hendry has put it,

[1] C. Nelson and R. Startz, "The Distribution of the Instrumental Variables Estimator and its T-Ratio," *Journal of Business*, **63**, 1990, S125-S140.

[2] G. Rudebusch, "The Uncertain Unit Root in GNP," *American Economic Review*, **83**, 1993, 264–272.

[3] L. Mátyás and L. Lovrics, "Missing Observations and Panel Data—a Monte-Carlo Analysis," *Economic Letters*, **37**, 1991, 39-44.

> To interpret empirical evidence, users of econometric theory tend to require general but simple formulae which are easily understood and remembered, not vast tabulations of imprecise results which are specific to an unknown extent.[4]

This often means thinking carefully about both the design of the experiment and the presentation of the results. One approach, called *response surfaces,* is discussed in Hendry[4] and in Davidson and MacKinnon.[5] The basic idea is that regression analysis of the Monte Carlo results can often be used to summarize these results in a simple form. One attractive feature of this approach is that the adequacy of the simple form that has been chosen can be tested by generating more results! More frequently, histograms and tables are used to summarize the results.

Guideline 2 is self-evident but often hard to achieve. If the experimental conditions with which one confronts one's estimator in the Monte Carlo study never happen in the "real world," one can never be sure whether the results are relevant either. Of course, the real world is complicated; rarely does the Monte Carlo world precisely mimic the real world. Often there is a trade-off between generality and ease of understanding. In other ways, however, when analytical results can only be established for special cases, Monte Carlo results can in fact be more general. It is also sometimes possible to use actual data in the design of the experiment.

Guideline 3 is closely related to guideline 2. The performance of an estimator often depends on several factors. In the following example, the bias of the 2SLS estimator will depend on the quality of the instrument (the extent to which the instrumental variable is correlated with the endogenous right-hand-side regressor) and on the sample size, among other things. In this case, it would be helpful if our Monte Carlo experiment lets us judge both the effect of sample size and the correlation of the instrument with the endogenous regressor on the bias in 2SLS.

Guideline 4 recognizes that the precision of the results depends on how many replications or "experiments" are performed. Clearly, one replication is not enough; and more replications are preferred to fewer. It is also possible to use *variance reduction techniques* to minimize the number of replications necessary to achieve a certain level of precision. An accessible discussion can be found in Hendry.[4] We illustrate a simple approach to determining the number of replications necessary shortly.

These four guidelines are meant only as suggestions, and the design of the experiment will often depend on the nature of the question being asked.

11.1.2 An Example

Hall pointed out that when future income is uncertain and the representative agent has time-separable preferences, maximization of expected utility implies the following for the time series path of aggregate consumption, C_t:

[4] D. Hendry, "Monte Carlo Experimentation in Econometrics," *Handbook of Econometrics,* Vol. 2, Z. Griliches and M. D. Intriligator, Elsevier, 1984, 944.

[5] R. Davidson and J. MacKinnon, "Regression-based Methods for Using Control Variates in Monte-Carlo Experiments," *Journal of Econometrics,* **54**, 1992, 203–222.

$$U'(C_{t+1}) = \frac{(1+\delta)}{(1+r)}U'(C_t) + \epsilon_{t+1} \tag{11.1}$$

where δ = rate of time preference
r = (constant) real interest rate
ϵ_{t+1} = an error term uncorrelated with C_t.[6]

Nelson and Startz use Monte Carlo methods to investigate problems in estimating such a model.[1] They consider the case where the marginal utility can be expressed well by a quadratic. By assuming that $\delta = r$, the model becomes

$$C_{t+1} + \beta C_{t+1}^2 = C_t + \beta C_t^2 + \epsilon_{t+1} \tag{11.2}$$

$$C_{t+1} - C_t = -\beta(C_{t+1}^2 - C_t^2) + \epsilon_{t+1} \tag{11.3}$$

where β is a parameter reflecting preferences. Nelson and Startz suggest that OLS estimation of this model may be a problem because the independent variable $(C_{t+1}^2 - C_t^2)$ and the error term share a common component. They try an instrumental variable procedure, with $C_t^2 - C_{t-1}^2$ for the instrumental variable.

Now suppose that the true model is given by $\beta = 0$. That is, consumption follows a random walk. When $\beta = 0$ the following equation describes the time series pattern of consumption:

$$C_{t+1} = C_t + \epsilon_{t+1} \tag{11.4}$$

One test of whether consumption follows a random walk or some other process is a t test of the 2SLS estimate of β using the estimating Eq. (11.3). Normally, the 2SLS estimate of β is consistent. Two questions arise, however: (*i*) If the standard asymptotic theory is applied to this case, how well does the asymptotic distribution approximate the finite sample distribution of $\hat{\beta}_{2SLS}$? (*ii*) With the presence of a unit root in the data under the null, are the conventional arguments even appropriate? That is, do the asymptotic results we would normally consider for 2SLS provide a good guide for inference in the finite samples we actually encounter? Even when the standard asymptotic results would be a useful guide, how will they do in a case like this where the problem of unit roots makes deriving the asymptotic results difficult?

Although there are analytical approaches to characterizing the actual finite-sample distribution of the 2SLS estimator of β in the model described by Eq. (11.2) and (11.3), they are quite difficult and the problem is a good candidate for a Monte Carlo experiment. How might one do that?

1. Generate, say, 120 ϵ's from a normal distribution with unit variance. This number is something you might want to vary later. (The number *120* represents the length of a typical quarterly time series.)
2. Generate a C_t series using Eq. (11.4).
3. Run 2SLS on Eq. (11.3) using $(C_t^2 - C_{t-1}^2)$ as an instrumental variable for $C_{t+1}^2 - C_t^2$.

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

```
program define hall
version 3.1
set more 1
quietly macro define _ii 1
set seed 1001                   /*Choose a number to begin*/
                                /*the random number stream*/
while %_ii<= 10000              /*For i less than or equal to 10,000*/
   {quietly{                    /*Do the following loop*/
   set obs 123                  /*Sample size = 123*/
   gen e=invnorm(uniform())     /*Generate error from standard normal*/
   replace e=0 if _n==1         /*Let first error = 0*/
   gen c=100+sum(e)             /*Generate c(t)*/
```

Note: $c_t = 100 + \sum_{i=0}^{i=t} \epsilon_i$

```
       gen c1=c[_n-1]              /*Generate lagged c(t)*/
       gen y=c-c1                  /*Generate dependent variable*/
       gen x=(c^2)-(c1^2)          /*Generate independent variable*/
       gen z=x[_n-1]               /*Generate instrumental variable*/
       reg y x (z)                 /*Run 2SLS*/
       }
       display _b[x] " " _b[x]/_se[x]        /*Display beta and t-ratio*/
       clear                            /*Clear data set*/
       macro define _ii=%_ii+1          /*Increment i and start loop again*/
    }
    end
```

FIGURE 11.1
A sample program.

4. Display and store the estimate of β and its t ratio.
5. Repeat steps 1–4 many times, say 10,000.
6. Analyze the output.

Figure 11.1 shows a sample program from STATA[7] that performs the Monte Carlo experiment described above. Most of the program is straightforward, and we will use it as a basis for discussing various practical aspects of Monte Carlo simulation.[8]

11.1.3 Generating Pseudorandom Numbers

A key aspect of any Monte Carlo program is the generation of the random numbers. In the sample program (Fig. 11.1) the line is

```
gen e=invnorm(uniform())        /*Generate error from standard normal*/
```

[7] In this and the remaining chapters of the book, the computer code, output, and numerical results come from STATA, a software program available for many computers and operating systems from STATA Corporation, College Station, Texas.

[8] Note that Fig. 11.1 is not intended to be an example of a well-written program. It is not. In EViews, TSP. Rats, SAS, STATA, etc., it is easy to write a program more compactly. This program is meant solely as a pedagogic device.

There are actually two steps to the process. First, a "random" number is generated from the uniform distribution on (0,1). Next, this uniform variate is transformed into a standard normal variable by use of the inverse cumulative normal distribution function. This process of converting a uniform variate into a normal variate is often called the *transformation* method. The idea is that *any* cumulative distribution function returns a number on the (0,1) interval. Given a way to generate a uniform variable therefore, a variable from density f can be obtained by using the inverse cumulative distribution function associated with that density. STATA, like many other packages, has a function for the standard cumulative normal (mean 0, variance 1). To create a normal variable y with mean μ and variance σ^2 from a standard normal variable x, the equation is

$$y = \mu + x \cdot \sigma \tag{11.5}$$

where x is the variable generated from the standard uniform density. In STATA, if we had wanted to create a variable with mean 3 and variance 4, we would have issued the command:

```
gen e_alt=3 + (sqrt(4)*invnorm(uniform()))
```

A similar sort of transformation is necessary if one wants to take a uniform (0,1) variable and create a uniform (α, β) variable. This approach can be easily extended to the case where the researcher wants to create a set of normal variables that are correlated with each other in some prespecified way.

How is the uniform variable created in the first place? There are several ways to generate a U(0,1) variable. The first important thing to recognize is that one can only generate pseudorandom numbers. That is, we can generate series of numbers that behave as random numbers but are completely deterministic. This feature is actually useful. One would hope, for instance, to be able to replicate one's work. In that case, it would be helpful to know exactly what stream of pseudorandom numbers was used.

One popular method for generating pseudorandom numbers is called the *congruential method.* To illustrate, we consider one example of a congruential generator. At the heart of this generator is the following equation:

$$R_{n+1} = 69069 R_n \qquad (\text{mod } 2^{32}) \tag{11.6}$$

where 2^{32} is the *modulus,* and the notation signifies that we take the quantity that precedes it, divide it by the modulus, and take the remainder. Each member of this series will be a number between 0 and 2^{32}. The number 69069 is called the *multiplier.* As it turns out, the choice of multiplier is very important. A poor choice can lead to sequences with undesirable properties. To start the process, the user specifies a *seed* value for the initial value R_0. In our example program the following line,

```
set seed 1001                      /*Choose a number to begin*/
                                   /*the random number stream*/
```

specified the value 1001 as the seed. One could choose a random seed, say, by using some function of the clock time. However, if one chooses the seed personally, it is possible to replicate the Monte Carlo study exactly by choosing the same seed.

For this method, the initial seed is typically a large, positive, odd number. An element of the series R is turned into a number between 0 and 1 by division, for example,

$$x_{n+1} = \frac{R_{n+1}}{2^{32}} \tag{11.7}$$

where x is the desired random number. One of the most important aspects of a random number generator is the periodicity. Eventually, any random number generator repeats itself; a good random number generator should have a period as large as possible. An upper bound to the periodicity is given by the modulus; hence this is usually chosen to be as large as possible. Often the periodicity can be increased by combining different streams of random numbers in clever ways. Fortunately, the researcher will not have to write his or her own random number generator, but the researcher should make sure that the one used is of acceptable quality. If the researcher suspects there is a problem, the generator should be tested with an appropriate statistical test.

The rest of the program is straightforward. Once one has created a set of epsilons ϵ, one creates a series for C_t and its lag. The dependent variable and independent variable are merely functions of these two variables, and one creates the instrument by taking the lag of the independent variable.

11.1.4 Presenting the Results

One of the most difficult aspects of any Monte Carlo study is presentation of the results. The problem is not much different, however, from that which arises in any empirical study, Monte Carlo or otherwise. Some methods that are used are these:

1. Tables with summary statistics about the quantities of interest.
2. Histogram or kernel density estimates (discussed later in this chapter).
3. Response surfaces. Essentially this method uses regression techniques to summarize the sensitivity of the results to the parameters of interest. Unlike the conventional non-Monte Carlo case, if one discovers that one does not have enough variation in the parameters of interest, one can simply perform more experiments. We will not have a lot more to say about response surfaces here, but good discussions can be found in Hendry[4] and Davidson and MacKinnon.[5]

Recall that we *know* that β in our example is 0. It is interesting to ask: given that $\beta = 0$, how often will one reject that null hypothesis given conventional levels of significance? One simple way to shed some light on this and related questions is to tabulate statistics for the Monte Carlo estimates. Table 11.1 presents some summary statistics from our Monte Carlo simulation. Since the objects of our analysis do not have a known distribution, it is useful to display various percentiles.

According to conventional distribution theory, at the 95 percent confidence level we would reject the null hypothesis if the t ratio was greater than 1.98. The news is quite bad: 75 percent of the time, the t ratio is larger than 3.53! The median t ratio is also quite high at 11.7, which implies a decisive rejection of the null, despite the fact that we *know* that the null model is correct!

TABLE 11.1
Summary statistics for $\hat{\beta}_{2SLS}$ and t ratio from 10,000 replications

Percentiles	$\hat{\beta}_{2SLS}$	t statistic
1	−.0014910	−0.00327
5	.0037121	0.14706
10	.0042638	0.61606
25	.0046738	3.53549
50	.0049745	11.73876
75	.0052915	25.72853
90	.0057522	43.42638
95	.0063265	57.37334
99	.0113449	95.86761

TABLE 11.2
Number of Monte Carlo replications required

	Replications required so that a 95% CI for p has length	
p	.01	.02
.01	1,521	380
.05	7,299	1,825
.1	13,830	3,457
.15	19,592	4,898
.2	24,586	6,147
.25	28,812	7,203

The table also reveals more bad news regarding $\hat{\beta}_{2SLS}$. The median value of $\hat{\beta}_{2SLS}$ is 0.005, and most of the distribution appears to be to the right of the true value of β.

When have we done "enough" replications? The answer depends on the precise question being asked, but consider the percentage of times we (falsely) reject the null hypothesis or the size of the test.

The simplest method is to use the normal approximation to the binomial. Recall that the variance of a binomial variable is

$$\sigma_p^2 = \frac{p(1-p)}{N}$$

If we are interested in 95 percent confidence intervals of the size of the t test, p might be the percentage of times we incorrectly reject the null that $\beta = 0$. Suppose we would like a 95 percent confidence interval around the nominal level of the test to be .01. By using the normal approximation to the binomial, a 95 percent confidence interval is

$$\text{prob}\{\hat{p} - \sigma_p z_{2.5\%} < p < \hat{p} + \sigma_p z_{97.5\%}\} = .95$$

In this example we therefore require that

$$2\sigma_p \cdot 1.96 = .01$$

Table 11.2 shows the required number of replications to produce a 95 percent confidence interval of length .01 or length .02 for various levels of p. Note that the number of necessary replications depends on the true value of p, which is unknown. One approach is to treat the suggested number of Monte Carlo trials from this procedure as a lower bound to the number of required observations.

Another way to present the results of this Monte Carlo experiment is to plot the empirical density of $\hat{\beta}_{2SLS}$. This can be done using either a histogram or, in this case, a kernel density estimator (described later in this chapter).

Figure 11.2 displays an estimate of the density of the 2SLS estimates. (We calculated the density by dropping observations above the ninety-fifth percentile or below the fifth percentile since in this simulation the 2SLS estimates ranged from −258 to 659!) One interesting thing to note is that the distribution looks nothing like a bell curve centered at zero.

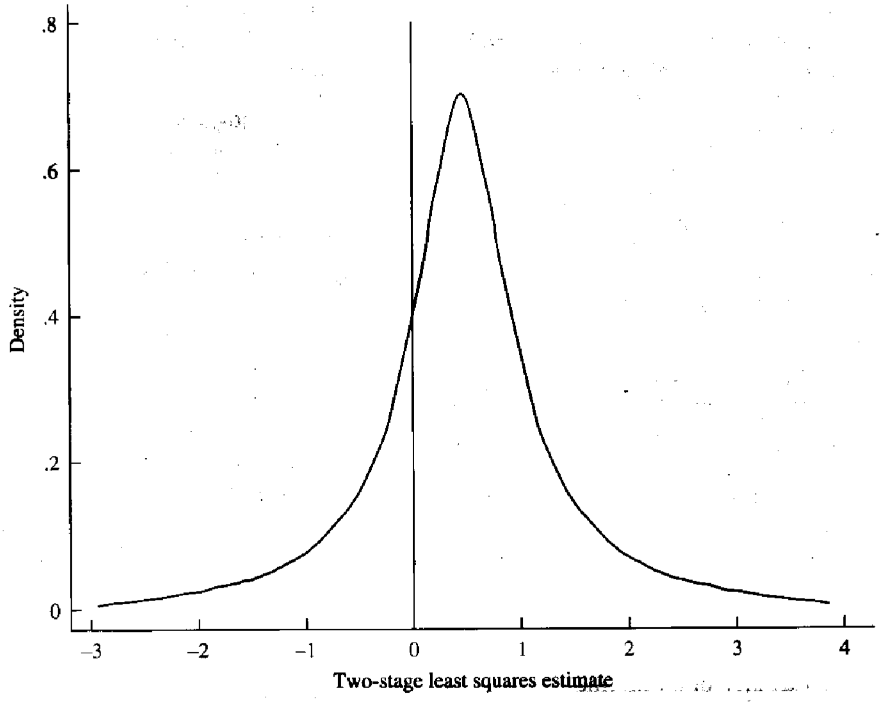


FIGURE 11.2
Distribution of $\hat{\beta}_{2SLS}$.

What makes this Monte Carlo experiment interesting is that it is a vivid demonstration of the proposition that the asymptotic distribution one might blithely assume is appropriate can be a very poor approximation to the actual finite-sample distribution. However, it is not immediately clear how the results of this Monte Carlo generalize to other instrumental variable problems. Is this true for all 2SLS estimators? Is the problem that the sample size is too small? Perhaps the problem is the low correlation of the instrumental variable with the endogenous regressor. One clear weakness of this Monte Carlo calculation is its specificity. That is, although we understand this one special case well, it is uncertain how well these results about 2SLS extend to other settings.

Nelson and Startz[1] extend this demonstration by considering the simple model

$$y = \beta x + u \tag{11.8}$$

where u is $N(0, \sigma_u^2)$. Without loss of generality, they consider the case when $\beta = 0$. The rest of the data generation process is described by

$$\begin{aligned} x &= \gamma u + \epsilon \\ z &= \rho\epsilon + \nu \end{aligned}$$

where ν and ϵ are both standard normal variates uncorrelated with each other and u. In this setup, the two parameters γ and ρ (besides N, the number of observations) allow for a wide variety of circumstances. The parameter γ calibrates the extent to which OLS is biased. When $\gamma = 0$, OLS is BLUE. The parameter ρ calibrates the quality of the instrument. Although z is a proper instrument in the sense of being

TABLE 11.3
Summary statistics for $\hat{\beta}_{2SLS}$ from 500 replications ($\gamma = 1$)

	$\gamma = 1$:	$\rho = 1$			$\rho = .05$			$\rho = .01$			$\rho = .001$		
Fractile	OLS	2SLS	ASY	p(H)	2SLS	ASY	p(H)	2SLS	ASY	p(H)	2SLS	ASY	p(H)
.01	.39	−.55	−.33	0	−32.4	−4.66	.01	−26.22	−23.26	.01	−14.09	−232.63	.01
.1	.44	−.25	−.18	0	−1.72	−2.57	.09	−1.4	−12.82	.1	−1.48	−128.16	.09
.5	.5	0	0	0	.44	0	.49	.48	0	.52	.49	0	.49
.9	.57	.15	.18	0	1.6	2.57	.92	1.8	12.82	.89	1.77	128.16	.89
.99	.62	.25	.33	0	15.86	4.66	1	22.43	23.26	.99	22.91	232.63	.99

Note: ASY column gives fractile implied by asymptotic distribution. The p(H) column gives fractile of the probability value of the Hausman test comparing OLS and 2SLS.

correlated with x and uncorrelated with u, when ρ is small, z is a poor instrument in the sense of not being highly correlated with x.

Nelson and Startz chose to examine several parameter spaces of interest. For illustration purposes, we look at the case when ρ is small relative to γ. Holding γ fixed at 1, and the number of observations fixed at 100, they experimented with varying ρ. They chose to present the results in tabular form. Table 11.3 is a modification of Table 1 in Nelson and Startz.[1]

Clearly, use of a poor instrument is to be avoided. By taking the worst case depicted in Table 11.3, when $\rho = .001$ the actual distribution of $\hat{\beta}_{2SLS}$ is quite poor. In fact, by most criteria the cure (2SLS) is worse than the disease (OLS), where the distribution is fairly tightly concentrated around .5. Note also the difference between the actual distribution of $\hat{\beta}_{2SLS}$ and the distribution predicted by asymptotic theory (the columns labeled ASY). Except for the case when $\rho = 1$ (the good instrument case) the actual dispersion tends to be much less than predicted by asymptotic theory. As we learned from Table 11.1, a high t ratio is no insurance against the possibility that the finite sample bias of $\hat{\beta}_{2SLS}$ is quite bad.

As the authors note, one possible protection against erroneous inference is to look directly at the correlation between the instrument and the explanatory variable. However, this task is not straightforward because estimates of the correlation will also be biased. The pattern of the probability value of the Hausman test [p(H)] in Table 11.3 is quite interesting. When the instrument is poor ($\rho = .05, .01, .001$), the Hausman test rarely rejects OLS in favor of 2SLS. In the case of a good instrument ($\rho = 1$), the Hausman test rejects quite frequently. Note also that when $\rho = 1$ the performance of $\hat{\beta}_{2SLS}$ is quite good relative to OLS.

Many researchers feel that the consequence of a poor instrument (an instrument that is exogenous but poorly correlated with the endogenous variable) is an imprecise 2SLS estimate. From this intuition, many have inferred that the problem of poor instruments is easily detectable by investigating the t statistic: if the standard error is not large for the 2SLS estimate, then one may infer that the instruments are adequate. Apparently, this intuition is wrong.

The Monte Carlo approach suggests instead that the consequences of poor instruments are estimates of $\hat{\beta}_{OLS}$ and $\hat{\beta}_{2SLS}$ that are not significantly different from each other even if the latter has a high t ratio. Why does 2SLS perform poorly in this case? After all, the problem satisfies the standard conditions for appropriate use of 2SLS: the instrumental variable is in fact uncorrelated with the error term (by construction!) and is correlated with the variable being instrumented. As we have learned, the problem is that when the correlation of the instrumental variable and the variable being instrumented is low, the asymptotic distribution of 2SLS is a very poor approximation to the actual finite sample distribution. Nelson and Startz discuss this case in detail and provide some analytical answers.[9]

[9] C. Nelson and R. Startz, "Some Further Results on the Exact Small Sample Properties of the Instrumental Variables Estimator,"*Econometrica*, **58**, 1990, 967–976. Note that Nelson and Startz mistakenly conclude that the distribution of 2SLS is bimodal when the instrumental variable is weakly correlated with the variable being instrumented. G. S. Maddala and J. Jeong, "The Exact Small Sample Distribution of the Instrumental Variable Estimator," *Econometrica*, **60**, 1992, 181–183, show that this is not correct, although the other results of the paper are correct.

11.2 MONTE CARLO METHODS AND PERMUTATION TESTS

A discussion of the permutation test due to Fisher[10] is a nice segue to our discussion of the bootstrap, which is closely related to the permutation test. The permutation test is commonly used to test whether two samples come from the same distribution. To illustrate, consider the following data from a well-known study on the minimum wage by David Card and Alan Krueger.[11]

The design of the study was straightforward. The simplest versions of the supply and demand model make a clear prediction about the impact of an increase in the minimum wage: If the minimum wage binds (that is, it is set high enough to affect some workers), employment should fall in response to the higher price of labor. Card and Krueger took advantage of a fortuitous natural experiment. With a Democrat-controlled New Jersey legislature in place, and on the heels of federal legislation that had raised the federal minimum to $4.25 an hour in April 1991, New Jersey voted to raise its state minimum wage above the federal minimum to $5.05, effective April 1992. In the two years between the passage of the higher minimum and its effective date, Democratic majorities in both houses of the legislature were replaced in a Republican landslide. In early 1992, before it was clear that the new law would actually go into effect, Card and Krueger surveyed fast-food restaurants in New Jersey and Pennsylvania. After a dramatic sequence of events, the minimum wage increase in New Jersey became official. Six months later, Card and Krueger resurveyed the same restaurants. First, they compared the distribution of wages across the two states before and after the minimum wage rise in New Jersey. Their wage data clearly showed that the minimum wage had an impact on the wages of many workers. Second, after having shown that the minimum wage did in fact bind for many workers, they compared employment changes in the two states. If the minimum wage change had the effect predicted by the simplest economic model, employment should fall in New Jersey relative to Pennsylvania.

For illustrative purposes a randomly chosen subset of the employment changes in their data are presented in Fig. 11.3. A standard approach to test whether the two samples of employment changes come from identical distributions is to compute a simple t test: If we assume that both sets of data come from the normal distribution with identical variances, but perhaps different means, the appropriate statistic is

$$\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{(1/n) + (1/m)}}$$

The difference in the means of the two samples is 2.14. (Employment increased in New Jersey—the state where the minimum wage was raised—relative to Pennsyl-

[10] R. A. Fisher, *The Design of Experiments*, 1935. Oliver and Boyd.

[11] For an excellent, though controversial, study on the economic effects of the minimum wage see their book *Myth and Measurement: The New Economics of The Minimum Wage*, Princeton University Press, 1995. The book is also a fine and accessible example of testing economic theories.

33 employment changes in New Jersey

```
-20, -17.5, -13, -12.5, -4.5, -4, -3.5, -2, -1.5, -1, -.5, -.5,
0, 0, .5, .5, 1.5, 2, 2, 2.25, 3, 4.5, 4.5, 5.5, 6, 6.25, 8.25,
9, 10, 10.5, 12, 14.75, 34
```

7 employment changes in Pennsylvania

```
-7, -6, -2.5, -.5, 4, 4.5, 4.5
```

FIGURE 11.3
A sample from the Card-Krueger data.

vania.) Is the difference significant? Using the standard t test already described, the value of the test statistic is 0.56. The probability that a value greater than this would be observed, under the null hypothesis that the samples come from the same distribution, is .5775: no evidence of a significant difference.

Another approach, which does not rely on assuming that the data come from a normal distribution, is called a permutation test. Consider two iid samples: X of size n, and Y of size m. If they are both drawn from the same distribution, any permutation of the elements in X and Y is equally probable. Given this fact, we could then proceed to enumerate every possible permutation of the data, calculate the difference in the means between these "X" samples and "Y" samples, and see how likely an outcome as high as 2.14 is from the set of possible permutations. We do not have to restrict ourselves to a comparison of means. We could consider other tests of whether the two sets of numbers come from the same distribution. One attractive feature of tests constructed in this way is that no assumptions about the distribution of the data have to be made.

In principle, a permutation test could be done using only a pen and paper, although this would involve enumerating an extremely large set of possibilities. An alternative is to use a Monte Carlo approach. Instead of systematically listing every permutation, we could simulate all the possibilities. Provided we choose a large enough number of replications, we can generate a very accurate distribution of the possible permutations. (When the possibilities are systematically enumerated the test is exact. When Monte Carlo methods are used these types of tests are often called *approximate randomization tests.*)

In our example, we would proceed as follows:

1. Draw two samples of size n and m, respectively, from all $n + m$ observations (pooling the two samples) *without replacement.* (For the data in Fig. 11.3, n and m would be 33 and 7, respectively.)
2. Compute the absolute difference (or the actual difference if one is interested in a two-sided test) between the means of the two samples.
3. Repeat the process a large number of times.
4. Calculate the percentage of times the value of the difference in means exceeds the value we computed from the original samples. This number is then the significance level of the difference.

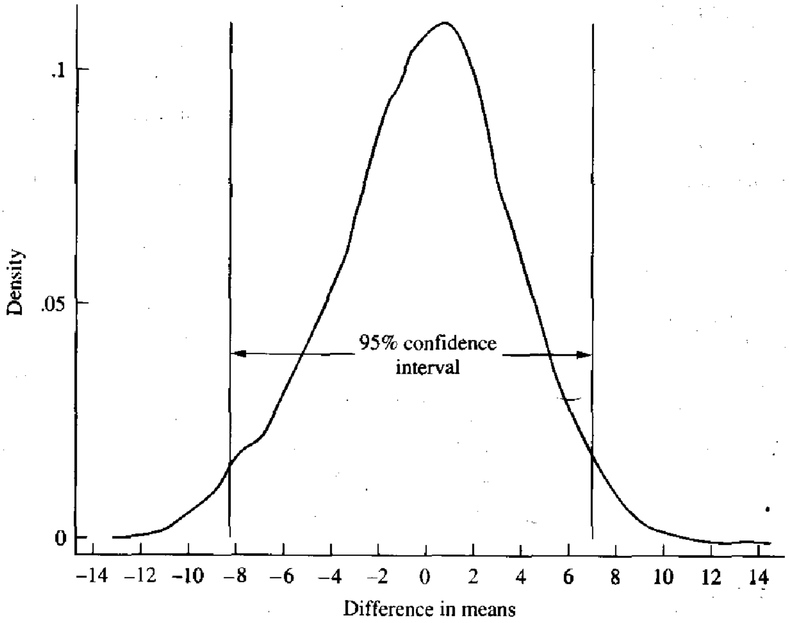


FIGURE 11.4
Estimates of density from 10,000 permutations using data from Fig. 11.3.

We could implement Step 1 in a computer program as follows:

1. Generate a random number for each of the $n + m$ observations.
2. Sort the data by this random number.
3. Label the first 33 observations of this sorted data *New Jersey* and the last 7 *Pennsylvania.*

Figure 11.4 shows a resulting estimate of the density from 10,000 Monte Carlo replications using the data from Fig. 11.3. At conventional levels of significance we would judge that the two samples are the same. In fact, the probability of observing a value of the difference between the sample means being greater than 2.14 in absolute value is .556—not very different from the value we obtained with the conventional t test. In fact, Fisher developed the permutation test to justify the use of the standard t test.

Fisher's exact test is closely related to the permutation test. In the former, all the possible permutations of a 2 × 2 table that have the same marginal values as the table being tested are calculated. The probability of the particular 2 × 2 table being observed is essentially calculated by enumerating all possible 2 × 2 tables and by finding out what fraction are the 2 × 2s of interest.

There are other applications of this general idea. Schmoyer shows how a version of the Durbin-Watson test that does not depend on normality of the error terms nor have an *inconclusive zone* can be constructed using a permutation-type test.[12]

[12] R. Schmoyer, "Permutation Tests for Correlation in Regression Errors," *Journal of the American Statistical Association*, **89**, 1994, 1507–1516.

An example of this test is provided in the exercises to this chapter. The approach also can be used to provide a spatial analog to the Durbin-Watson test. Another interesting application is from Kling and Anderson.[13] They have data on cases brought before several judges for the same jurisdiction. The cases are assigned almost at random to each of the judges. For each judge information on what charge was made in the cases that were adjudicated and what sentence was meted out is available. Presumably, if the court is "fair," the judge one faces should be irrelevant to the length of sentence. One interesting issue that can be evaluated with a permutation test is the hypothesis that the judge one faces matters in sentencing.

Although these tests can be quite valuable in some contexts, they suffer from the drawback that rejections of the null hypothesis are often not informative about the reason for the rejection. Furthermore, the permutation test just described cannot be used in many real-life situations. Taking our example, suppose we want to test whether the sample of New Jersey changes had a mean of zero. In this case, there is nothing to permute and a simple permutation test cannot be used.

11.3 THE BOOTSTRAP

The *bootstrap* due to Efron is more versatile than the permutation test and its use in applied econometric work is growing.[14] The bootstrap is often used in these circumstances:

1. An analytic estimate of the standard error of an estimator is too difficult or impossible to calculate.
2. The researcher has reason to believe that asymptotic theory provides a poor guide to the precision of a particular estimator and desires an alternative that may provide a better finite sample approximation.

Although not a panacea, the bootstrap holds great promise in many applications and is finding its way into more and more applied econometric research. The advantage of the bootstrap is that one does not have to know the underlying data generation process, unlike the Monte Carlo method.

11.3.1 The Standard Error of the Median

The classic illustration of the power of the bootstrap is the computation of the standard error of the sample median.

Consider a random variable x with distribution $f(x)$ and one sample of size n from this distribution. The standard approach to calculating the standard error of the median would be to develop an estimator analytically and then compute an estimate from the sample.

[13] J. Kling and J. Anderson, "Did Sentencing Guidelines Reduce Disparity between Judges in Prison Sentences?" mimeo, March 1996, Massachusetts Institute of Technology, Department of Economics.

[14] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, **7**, 1979, 1–26.

The formula for the standard error of the median is

$$s_{\text{median}} = \sqrt{4\widehat{f^2(0)}}$$

where $f(0)$ is a consistent estimate of the value of the probability density function at 0. One would have to use the data to generate an estimate of the distribution and then calculate $f^2(0)$. A second approach to calculating the precision of the median would be (*i*) to draw a large number of samples of size n from the distribution $f(x)$, (*ii*) to calculate the median in each of these samples, and (*iii*) to calculate the square root of the variance of these estimated medians across a large number of replications.

In both approaches we could calculate a consistent estimate of the standard error if we had precise knowledge of the distribution generating the samples in the first place. Typically, we do not. Efron's suggestion was to use the sample data to generate an estimate of the distribution. That is, use the empirical distribution to learn about the actual distribution. For a sample X and for $i = 1, \ldots, B$ the procedure amounts to the following:

1. Generate a random sample X^i *with replacement* from the original sample X.
2. Compute the median $\widehat{M}^i$ for this new sample.
3. Store the value of $\widehat{M}^i$.

The number of bootstrap replications, B, should be set as high as is practical, as in a Monte Carlo experiment. The bootstrap standard error of the median is

$$\sigma_{\text{boot}} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}[\widehat{M}^i - \widehat{M}^{(\cdot)}]^2}$$

where

$$\widehat{M}^{(\cdot)} = \frac{1}{B}\sum_{i=1}^{B}\widehat{M}^i$$

In Step 1, we are drawing a sample of size n from the original sample, thus putting a probability $1/n$ on each observation in the sample.

11.3.2 An Example

The bootstrap technique has many possible applications, for example, as an alternative to the permutation test described earlier. The example we pursue below is not a typical case in which one would use the bootstrap, but we discuss it to show the relationship between the permutation test described and the bootstrap.

By using the Card and Krueger data (Fig. 11.3), the algorithm for a bootstrapping version of the test that the employment changes come from the same distribution is as follows:

1. Draw a sample of size $n + m$ from the set of New Jersey and Pennsylvania observations *with replacement.*
2. Compute the absolute difference (or the actual difference if one is interested in a one-sided test) between the means of the New Jersey observations and the Pennsylvania observations. *Note:* This could be accomplished by running a regression of the change in employment on a constant and a dummy variable that indicates

whether the observation is from New Jersey and reading off the coefficient on the dummy variable.

3. Repeat Steps 1 and 2 a large number of times.
4. Calculate the percentage of times the value of the difference in means exceeds the value we computed from the original samples (in our example this value was 2.14). This number is then the significance level of the difference.

In this case, the bootstrap provides very similar answers to the approximately exact test, which is comforting. A symmetric 95 percent confidence interval using the bootstrap yields $(-2.6, 7.2)$ and our inference about the nature of employment changes in the two states is the same. It is interesting to compare the results from the three tests (Table 11.4).

All three lead to the same inference about the effect of the change in the minimum wage, which is that it had no impact on employment. Notice the similarity of the bootstrap approach to the permutation test. The single difference is that in bootstrapping the sampling is done *with replacement.* In the permutation test, the sampling is done *without replacement.*

An alternative bootstrap test for the difference in the means uses pivotal statistics. As we will briefly discuss, such tests have some advantages. In the present instance, the exercise amounts to calculating the t statistic,

$$\frac{\bar{X}^b - \bar{Y}^b}{\sqrt{(s_x^2/n_x) + (s_y^2/n_y)}}$$

for each of the B bootstrap samples ($b = 1, \ldots, B$), where

$$s_x^2 = \frac{\sum_i^{n_x}(x_i - \bar{x})^2}{n_x - 1}$$

and n_x is the number of observations of the x variable. The terms are defined similarly for the y data.

The result is a distribution of "t statistics." The idea is to see how the t statistic computed for the original sample compares with the set of "possible" t statistics that could have arisen. This particular pivotal method is called the *bootstrap t* or *percentile t,* for reasons that will become clear shortly. See Jeong and Maddala for a further discussion.[15] An example is provided as an exercise.

TABLE 11.4
The two-sample problem

Test	95% confidence interval
Permutation	−6.98 to 7.90
Bootstrap	−2.65 to 7.24
Asymptotic	−5.56 to 9.85

[15] J. Jeong and G. S. Maddala, "A Perspective on Application of Bootstrap Methods in Econometrics," Chapter 11, *Handbook of Statistics,* eds. G. S. Maddala, C. R. Rao, and H. D. Vinod, Elsevier, 1993, 573-610.

The potential for applying the bootstrap is much greater than the permutation test. For instance, suppose we had only the New Jersey sample of employment changes and we wished to test the hypothesis that the mean was equal to zero. In this case, there is nothing to permute so the permutation test is not applicable, but the bootstrap can still be used.

11.3.3 The Parametric Bootstrap

The greater applicability of the bootstrap has made it increasingly popular in econometric applications. There are several different types of bootstrap procedures. A common one is the *parametric bootstrap.* An example will illustrate.

Consider the ratio (or any nonlinear function) of two parameters:

$$\text{ratio} = \frac{\alpha_1}{\alpha_2} \tag{11.9}$$

Further suppose that we have no estimate of the ratio, but that we do have estimates of α_1 from one sample (or study) and α_2 from another sample (or study.) Clearly one estimate of the ratio is merely

$$\widehat{\text{ratio}} = \frac{\widehat{\alpha_1}}{\widehat{\alpha_2}} \tag{11.10}$$

where $\widehat{\alpha_i}$ are the two estimates. If we want an estimate of a confidence interval for $\widehat{\text{ratio}}$ we can use what is called the parametric bootstrap:

1. Generate 10,000 draws of $\widehat{\alpha_1}$ from the distribution $N(\widehat{\alpha_1}, \widehat{\sigma^2}_{\alpha_1})$ where $\widehat{\sigma^2}_{\widehat{\alpha_1}}$ is the estimated variance of $\widehat{\alpha_1}$. Note that we assume that the asymptotic properties hold so we can use the normal distribution.
2. Generate 10,000 draws of $\widehat{\alpha_2}$ from the distribution $N(\widehat{\alpha_2}, \widehat{\sigma^2}_{\alpha_2})$ where $\widehat{\sigma^2}_{\widehat{\alpha_2}}$ is the estimated variance of $\widehat{\alpha_2}$.
3. Compute 10,000 pairs of $\widehat{\alpha_1}$ and $\widehat{\alpha_2}$ and their ratio using the generated data.
4. Present a 95 percent confidence interval.

In a parametric bootstrap, we essentially perform a Monte Carlo, where the parameters of our simulation are calculated from the data and use a specific distribution such as the normal.

Let us consider this example in more detail: The appropriate lines of computer code resemble Fig. 11.5. In the the code, **alpha_1** = $\widehat{\alpha_1}$, **alpha_2** = $\widehat{\alpha_2}$, and **se_a1** and **se_a2** are the standard errors of $\widehat{\alpha_1}$ and $\widehat{\alpha_2}$, respectively. An example of the use of the parametric bootstrap can be found in the paper by Valletta.[16]

Calculating a standard error for the estimate is probably not a good idea. You may recall that the ratio of two standard independent normals has a *Cauchy* distribution, which has no mean or higher moments. In this example, it is perhaps simplest to note

[16] R.G. Valletta, "Union Effects on Municipal Employment and Wages—A Longitudinal Approach," *Journal of Labor Economics*, **11,** 1993, 545–574.

```
set obs 10000                                     /*10,000 Observations*/
generate epsilon1= invnorm(uniform())             /*Generate std. normals*/
generate epsilon2= invnorm(uniform())             /*Generate std. normals*/
generate top = alpha_1 +  epsilon1 * se_a1        /*Generate numerator of ratio*/
generate bottom = alpha_2 +  epsilon * se_a2      /*Do same for denominator*/
generate ratio= top/bottom                        /*Compute ratio*/
centile ratio, centile(2.5,97.5)                  /*Display 95% Confidence Interval*/
```

FIGURE 11.5
Sample parametric bootstrap.

that the denominator spends some time close to zero, which would make the ratio infinite. The appropriate method in this case is the *percentile method.* The method is quite straightforward if the distribution of the ratio is approximately symmetric. Order the estimated ratios in ascending order by size: $\{r_1, r_2, \ldots, r_{10,000}\}$. Then 95 percent confidence interval is

$$r_{251} \leq r \leq r_{9,750}$$

When the distribution is not symmetric, the appropriate procedure is the *modified percentile method.* If one is interested in a 95 percent confidence level, the procedure is to find the *smallest* interval that includes 95 percent of the observations.

One final note should be made about the use of the parametric bootstrap. We assumed that the two parameters ($\widehat{\alpha_1}$ and $\widehat{\alpha_2}$) were independent. If they are not, then the foregoing procedure should be modified to incorporate this fact.

11.3.4 Residual Resampling: Time Series and Forecasting

In time series applications the most common form of bootstrapping is based on resampling the ***residuals.*** For the model with t observations

$$y_t = X_t\beta + \epsilon_t \tag{11.11}$$

the *residual resampling **bootstrap procedure*** to compute standard errors for $\hat{\beta}$ is

1. Estimate $\hat{\beta}$.
2. Calculate $\hat{y}$ and the residual $\hat{\epsilon}$. Store the $\hat{y}$ and $\hat{\epsilon}$ in separate places.
3. Rescale or standardize the residuals as described below.
4. For the B bootstrap samples, do the following:
   a. Draw a sample of size T with replacement from the set of rescaled residuals $\hat{\epsilon}^b$.
   b. Construct new dependent variables y_t^b with the formula
   $$y_t^b = \hat{y}_t + \hat{\epsilon}^b$$
   That is, for each element in $\hat{y}$, draw an adjusted residual randomly with replacement and add it to generate a new y variable.
   c. Regress y_t^b on X_t and save the estimated coefficients.
5. Compute a 95% confidence interval using the percentile method or compute the standard error of $\hat{\beta}$ from the sample of bootstrapped $\widehat{\beta}$'s with the standard formula

$$\hat{\sigma}_{\hat{\beta}} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\beta}^i - \hat{\beta}^{(\cdot)})^2}$$

where $$\hat{\beta}^{(\cdot)} = \frac{1}{B}\sum_{i=1}^{B}\hat{\beta}^i$$

As already noted, it is not appropriate merely to use the residuals. Why are the unadjusted residuals too small? It is left as an exercise for the reader to show that when ϵ_t is iid, the OLS residual of Eq. (11.11) has variance

$$E[\hat{\epsilon}_t^2] = (1 - h_t)\sigma^2 \tag{11.12}$$

where h_t is defined as $$h_t = X_t(X'X)^{-1}X_t'.$$

Instead, one can rescale the residuals in this manner:

$$\tilde{\epsilon}_t = \frac{\hat{\epsilon}_t}{\sqrt{(1-h_t)}} - \frac{1}{N}\sum_{s=1}^{N}\frac{\hat{\epsilon}_s}{\sqrt{(1-h_s)}} \tag{11.13}$$

The rescaled residuals are sometimes called the standardized residuals, and h_t is the *t*th diagonal element of the *hat matrix.* The second term in our rescaled residual is there to ensure that the mean of the resulting residual remains zero.

Residual resampling bootstrapping is most frequently implemented in time series applications. Often the context is a situation where the error term has a (known) time series correlation. One example that exploits the usefulness of the bootstrap for time series applications is the paper by Bernard and Veall.[17] Among other things, they are interested in a confidence interval for a forecast of electrical demand y at some future point in time.

Bernard and Veall settled on the following two-equation system:

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t \tag{11.14}$$

$$x_t = Z_t\gamma + \mu_t \tag{11.15}$$

where ϵ_t is iid and μ_t follows a first-order autoregressive process

$$\mu_t = \rho\mu_{t-1} + \eta_t \tag{11.16}$$

One object of interest is the confidence interval for y_{T^*}—a forecast of future electricity demand. For purposes of illustration we will assume that future values of Z are known.

Performing the bootstrap can be thought of as involving two conceptual steps. In the first, an explicit process for calculating a forecast is written down. In the second, this process is bootstrapped and the relevant statistics are calculated.

We can specify a forecast of electricity demand at some time in the future T^*:

1. Estimate β_0 and β_1 by OLS and compute an OLS estimate of γ.
2. Use the residuals from OLS estimation of Eq. (11.16) to calculate $\hat{\rho}$.

[17] J. T. Bernard and M. Veall, "The Probability Distribution of Future Demand: The Case of Hydro Quebec." *Journal of Business and Economics Statistics,* **5,** 1987, 417–424.

3. Using the estimate $\hat{\rho}$ to calculate γ_{GLS}, predict x_{T^*} as,

$$\hat{x}_{T^*} = Z_{T^*}\hat{\gamma}_{\text{GLS}}$$

for a given set of values of Z_{T^*}.[18]

4. Forecast $\widehat{y_{T^*}}$ with the equation

$$y_{T^*} = \hat{\beta}_0 + \hat{\beta}_1\hat{x}_{T^*}$$

The bootstrap is implemented on this process in the following way:

1. Compute estimates of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\rho}$ and $\hat{\gamma}_{\text{GLS}}$.
2. Construct B bootstrap samples:

a. Draw a random sample $\widehat{\eta^i}$ of size T with replacement from the set of residuals $\hat{\eta}$.

b. Take the first element of $\widehat{\eta^i}$ and divide by $(\sqrt{1-\hat{\rho}^2})$ to yield $\hat{\mu}_1^i$.

c. Construct the remaining elements by:

$$\hat{\mu}_t^i = \hat{\rho}\hat{\mu}_{t-1}^i + \hat{\eta}_t^i \qquad \text{for } t = 2,\dots,T$$

d. The artificial sample is then developed as

$$x_t^i = Z_t\hat{\gamma}_{\text{GLS}} + \hat{\mu}_t^i \qquad \text{for } t = 1,\dots,T$$

where $\hat{\gamma}_{\text{GLS}}$ is the original GLS estimate.

e. The sample is then completed by using

$$y_t^i = \hat{\beta}_0 + \hat{\beta}_1 x_t^i + \hat{\epsilon}_t^i \qquad \text{for } t = 1,\dots,T$$

where $\hat{\beta}_0$, and $\hat{\beta}_1$ are the original OLS estimates and $\hat{\epsilon}_t^i$ are randomly resampled with replacement from $\hat{\epsilon}$.

With each of these bootstrapped samples in hand, one can calculate a set of forecasts for $\hat{x}_{T^*}^i$:

$$\hat{x}_{T^*}^i = Z_{T^*}\hat{\gamma}_{\text{GLS}}^i \tag{11.17}$$

where the i superscript denotes that the estimate is taken from the ith bootstrapped sample. This ith bootstrapped estimate, $\hat{x}_{T^*}^i$ along with the corresponding bootstrapped estimates for $\hat{\beta}$ can be used to construct an estimate of $\hat{y}_{T^*}^i$ Finally, confidence intervals can be constructed as described earlier using this sample of bootstrapped estimates.

One significant drawback of this method is that it is not well-suited to situations where the error terms are not (conditional on a known error process) identically and independently distributed. For instance, the procedure is not correct if the error terms are heteroscedastic. Recall that heteroscedasticity is a relationship between the variance of the error term and the independent variables. In residual resampling, different "error" terms get attached to different x's, thus scrambling the relationship. For such cases, a different form of bootstrapping is needed.

[18] Bernard and Veall make the (reasonable) assumption that T^* is sufficiently far in the future that they can ignore the influence of μ_T. When such an assumption is not plausible, the forecast should incorporate the effect of the serially correlated errors. See Section 6.8, "Forecasting with Autocorrelated Disturbances," for a discussion.

11.3.5 Data Resampling: Cross-Section Data

The reader who has been following the discussion so far may have thought of a more straightforward approach to bootstrapping in the regression context. Instead of resampling from the residuals, why not resample from the (y, X) pairs of the data? In fact, in cross-section work this method is the most common.

Starting again with the standard model,

$$y_i = X_i\boldsymbol{\beta} + \epsilon_i \tag{11.18}$$

the procedure is as follows:

1. Sample with replacement from the original (y, X) sample in "pairs."
2. Compute the statistic of interest.
3. Repeat Steps 1 and 2 a large number of times.

This procedure, in contrast to the previous, is robust to heteroscedasticity. That is, one does not have to assume that the errors are iid. It would then appear that this procedure is to be preferred in general. In cross-section or panel data (see Chapter 12) applications it has much to commend itself.

Unfortunately, this procedure is not implementable in the typical time series case when the error term is correlated across time. Resampling from the data in this way again scrambles the relationship between adjacent error terms.

For use in panel data, the bootstrap needs to be modified slightly. Specifically, one needs to resample *clusters* where each individual or cross-section unit is a cluster, instead of observations. Suppose one has panel data (y_{it}, X_{it}) for N individuals and T time periods for a total of NT observations. Let y_i be the $T \times 1$ vector of observations for individual i and X_i the corresponding $T \times k$ matrix of independent variables for individual i. Instead of sampling with replacement from the set of NT observations, one should sample with replacement from the set of N individuals. That is, one should sample with replacement from (y_i, X_i). In that way, the correlations within each individual are kept.

11.3.6 Some Remarks on Econometric Applications of the Bootstrap

In econometric applications, the bootstrap seems to be used most frequently in situations like those explored by Bernard and Veall,[17] especially confidence intervals for forecasts. Such a confidence interval is difficult to compute analytically, and the bootstrap provides a useful alternative. The parametric bootstrap is often used to compute confidence bands of impulse response functions in vector autoregressions.

When the asymptotic theory is tractable, however, it is not clear that the bootstrap is better. One exception is the computation of pivotal statistics—statistics like the simple t test whose distribution does not depend on the underlying parameters.[19]

[19] For example, if a t statistic is constructed by taking the ratio of an estimate of parameter β to an estimate of its standard error σ, it has a t distribution regardless of the true values of β and σ. Loosely, statistics whose distribution does not depend on the specific values of the underlying parameters are said to be pivotal.

Horowitz points out that for such statistics the bootstrap *may* provide a way to improve on the approximations of asymptotic theory.[20] Even this case is tempered by the fact that, because the output of a bootstrap is a random number, achieving such an improvement may not be practical. A good rule is that the bootstrap should be evaluated by Monte Carlo techniques (although these are rather computationally intensive as well!). Furthermore, the bootstrap may sometimes fail. Brown and Newey demonstrate, for example, that the bootstrapped Generalized Method of Moments statistic (GMM) has no power.[21]

Finally, several authors question the emphasis of much applied econometric work in calculating standard errors. Jeong and Maddala argue that "standard errors are of interest only if the distribution is normal.... If one wants to make confidence interval statements and to test hypotheses, one should use the bootstrap method directly and skip the standard errors, which are useless."[14]

11.4 NONPARAMETRIC DENSITY ESTIMATION

Nonparametric density estimation is a topic very different from the ones we have covered so far. We discuss it here only because the technique has grown in popularity with advances in computing. Nonparametric density estimation is most frequently used for exploratory data analysis although, as we will illustrate shortly, it can also be useful for more sophisticated data analyses.

The object of interest is a density. Frequently, we adopt parametric methods to describe the density of a variable. For example, it is often alleged that the distribution of male log wages is approximately normal. Consider our sample of 1000 men from the 1988 Current Population Survey (see example 6.1 "Tests for Heteroscedasticity" in Chapter 6). The parametric approach to describing the distribution would begin by specifying a distribution (in this case normal), calculating a set of sufficient statistics for the distribution from the sample (in this case the mean and the variance), and then using the equation for the normal density as follows:

$$f(x) = \frac{1}{\sqrt{2\pi\widehat{\sigma^2}}}\exp-\frac{(x-\hat{\mu})^2}{2\widehat{\sigma^2}}$$

where $\hat{\mu}$ and $\widehat{\sigma^2}$ are estimated in the usual way, that is,

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad \widehat{\sigma^2} = \sum_{i=1}^{N}\frac{(X_i-\hat{\mu})^2}{N-1}$$

We now have an equation that yields an estimate of the density for any value of x.

[20] J. Horowitz, "Bootstrap Methods in Econometrics: Theory and Numerical Performance," Working Paper Series #95-10, July 1995, University of Iowa, Department of Economics.

[21] B. Brown and W. Newey, "Bootstrapping for GMM," July 1992, Massachusetts Institute of Technology, Department of Economics seminar notes.

If the data are known to be distributed normally, the parametric procedure has much to commend itself. In particular, we have taken a complex problem of describing a sample and reduced it to the estimation of two parameters.

On the other hand, suppose we are not certain that the data are in fact distributed normally. The parametric approach would involve taking various density functions (normal, beta, gamma, log normal, etc.), estimating the sufficient statistics of those distributions, performing appropriate hypothesis tests, and finally settling on a distribution that gives us a good characterization of the data.

The nonparametric approach is quite different. The idea is to rid oneself of the need to specify in advance a particular functional form. The simplest form of nonparametric density estimation is the histogram. In Fig. 11.6, the parametric approach (assuming a normal distribution) is compared with the simple histogram.

Although our understanding of the histogram is quite intuitive, it will be helpful to be a bit more formal. Each of the rectangles is called a *bin*. In Fig. 11.6 there are five bins of equal width h. The procedure requires choosing a starting point x_0 and drawing the first rectangle with width $x_0 + h$. The second rectangle begins where the first one left off covering the interval $(x_0 + h, x_0 + 2h)$, and so on. Given the width of the bin, there is a simple algorithm for determining the height of the bins. Because the height of the bin is an estimate of the density at a particular value of x, we can write the algorithm as follows:

$$f(x) = \frac{1}{Nh}[\text{number of observations in the interval } (x, x + h)]$$

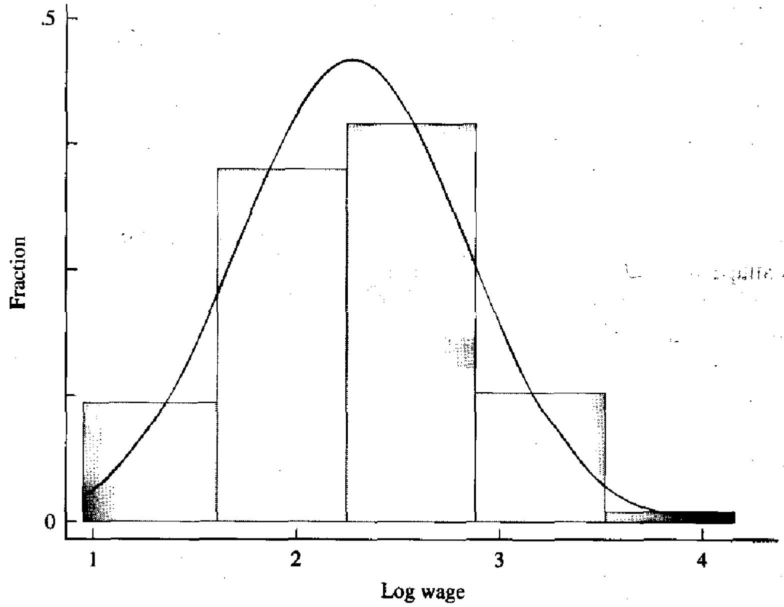


FIGURE 11.6
Histogram with five bins.

Two features of the histogram deserve mention. First, the density estimate depends on the choice of starting location x_0. Second, and more important for the discussion that follows, the shape of the histogram depends on the width of the bins. Narrower bins correspond to histograms that are less "smooth"; wider bins are smoother.

For example, consider the same data as in Fig. 11.6 but with narrower bins and with the parametric estimate again superimposed on the histogram (Fig. 11.7). Although we have used the same data, the picture looks more jagged and there appear to be spikes in the data at various locations. As previously noted, one undesirable feature of the histogram is that it depends on the choice of origin. Figure 11.8 is a histogram with 50 bins but a slightly different origin. A close inspection of the left-hand tail seems to reveal a different shape in the density than in the previous histogram. Like the previous histogram, it is not very smooth.

A simple solution that avoids this problem is the *naive* estimator.[22] Recall that the definition of the probability density function for a random variable is

$$f(x) = \lim_{h\to 0} \frac{1}{2h} P(x - h < X < x + h) \tag{11.19}$$

We can approximate this for a given value of h as

$$\widehat{f(x)} = \frac{1}{N}\frac{1}{2h}[\text{number of } X_1, \ldots, X_N \text{ falling in the interval } (x - h, x + h)] \tag{11.20}$$

As will be apparent in a moment, it will be useful to describe this estimator as

$$\widehat{f(x)} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{h} w\left(\frac{x - X_i}{h}\right) \tag{11.21}$$

where w is a ***weighting*** function defined as

$$w(z) = \begin{cases} \frac{1}{2} & \text{if } |z| < 1 \\ 0 & \text{otherwise} \end{cases}$$

This estimator is closely related to the histograms we have just described. The substantive difference is that the naive estimator does not have the undesirable property that its shape depends on the choice of origin. As Silverman describes it, the naive estimator is "an attempt to construct a histogram where every point is the center of a sampling interval."[22]

An example of this naive estimator using the same data is given in Fig. 11.9. Such a histogram is satisfactory for many purposes, but it is quite simple to do a little bit better. In particular, such an estimator is not smooth—it has jumps at the edges of bins and zero derivatives everywhere else.

Define a kernel function as a function such that

$$\int_{-\infty}^{\infty} K(x)dx = 1$$

[22] B. Silverman, *Density Estimation for Statistics and Data Analysis*, 1986, Chapman and Hall.

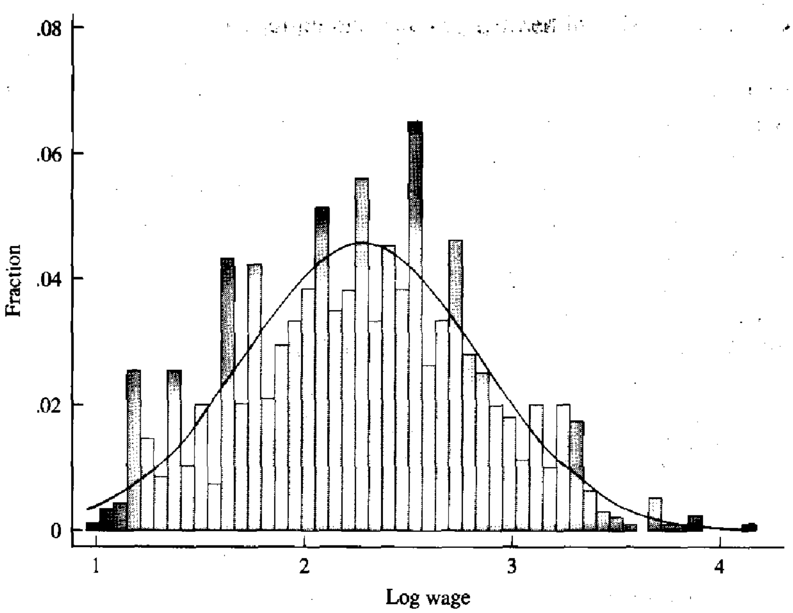


FIGURE 11.7
Histogram with 50 bins.

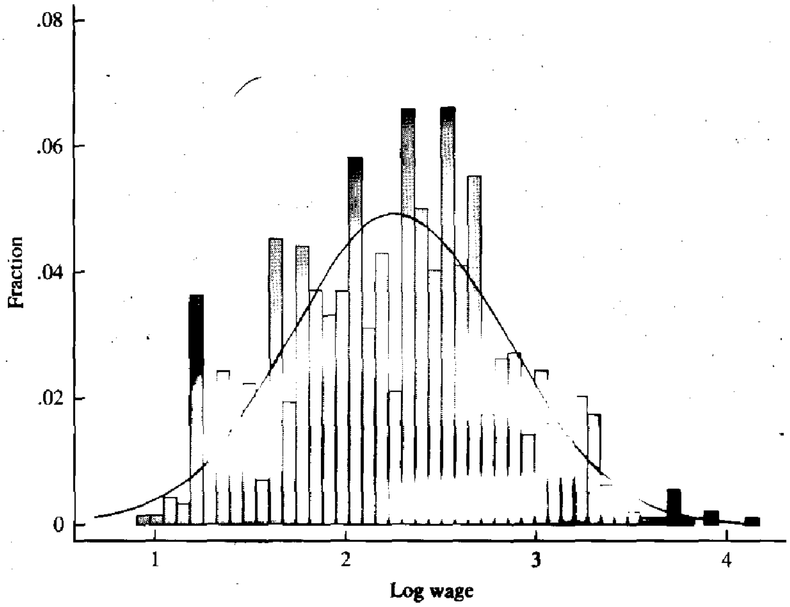


FIGURE 11.8
Histogram with 50 bins but different starting point.

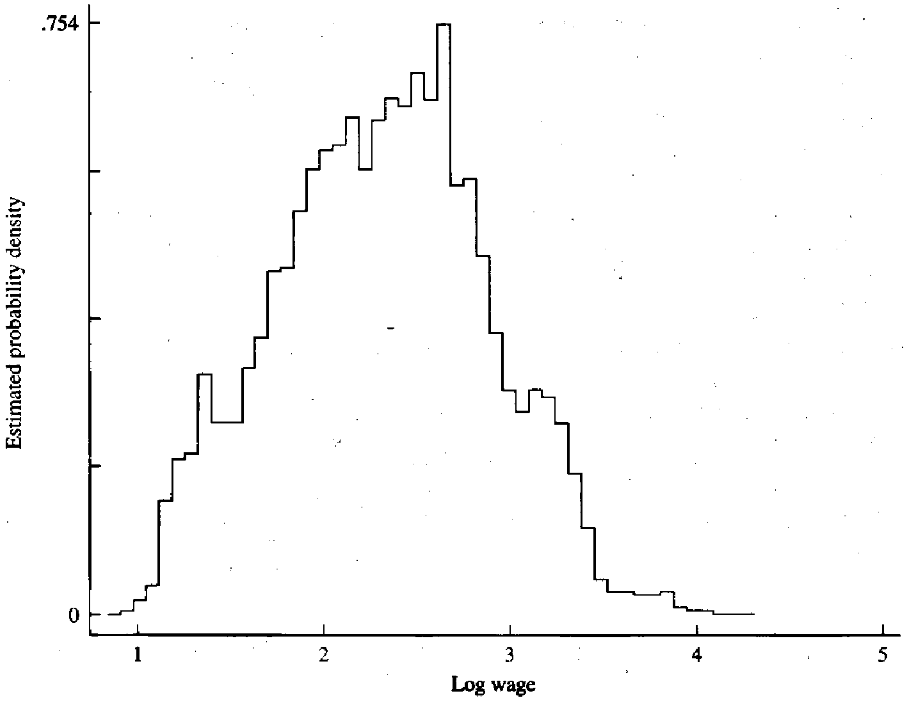


FIGURE 11.9
Naive estimator with 50 bins.

Instead of using the previous weight function (known as the *rectangular kernel*) we can replace it with a kernel function that is smooth and has derivatives. That is, in contrast to the naive estimator, the basic idea of the kernel density estimator is to put a "bump" on every point of the estimate. The most common kernels are presented in Table 11.5.

We can now define a broad class of density estimators, the *Rosenblatt-Parzen* kernel density estimator, as

$$\widehat{f(x)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{x - X_i}{h}\right) \tag{11.22}$$

TABLE 11.5
Some common kernels

Name	$K(z)$	
Biweight	$\frac{15}{16}(1 - z^2)^2$	for $\lvert z\rvert < 1$
	0	otherwise
Epanechnikov	$\frac{3}{4}\frac{1-(z^2/5)}{\sqrt{5}}$	for $\lvert z\rvert < \sqrt{5}$
	0	otherwise
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-(z^2/2)}$	
Rectangular	$\frac{1}{2}$	for $\lvert z\rvert < 1$
	0	otherwise
Triangular	$1 - \lvert z\rvert$	for $\lvert z\rvert < 1$
	0	otherwise

where K refers to the kernel (often one of those defined in Table 11.5) and h is the bandwidth.

Again in contrast to the naive estimator, kernel density estimators are locally smooth. The extent to which this smoothness is necessary will depend on the problem at hand. As a practical matter, in most exercises the choice of kernel turns out to be relatively unimportant (although it is useful to verify this in practice). This result is not entirely surprising since we noted that the *bandwidth* for the histogram was quite important to determining the amount of smoothing.

11.4.1 Some General Remarks on Nonparametric Density Estimation

Some general comments about kernel density estimates will be useful:

- Kernel density estimation methods are easily applied to data when sampling probabilities are provided. In several data sets (the Current Population Survey, the Panel Study of Income Dynamics, the Survey of Income Program Participation, for example) a weight reflecting the inverse probability of being selected for the sample is often included. If we denote this weight by θ, the kernel density estimate is modified in a straightforward way:

$$\widehat{f(x)} = \frac{1}{N}\sum_{i=1}^{N}\frac{\theta}{h}K\left(\frac{x - X_i}{h}\right)$$

- What is the right choice of bandwidth? In general the rule is that there is a tradeoff between variance and bias. The larger the bandwidth the smaller the variance but the greater the bias, and vice versa. As it turns out a number of methods are available for automatically choosing the bandwidth, ranging from simple cross-validation to various "plug-in" methods.[23] In many instances, it will be sufficient to judge the right bandwidth by the "eyeball" method, that is, whatever looks appropriate to the eye. As a general rule, it is easier to smooth with the eye than to "unsmooth" with the eye, so "oversmoothing" is to be avoided.
- Caution is warranted when applying kernel methods to data with "long tails." In these cases, there is a tendency for spurious noise to appear in the tails of the estimates. Smoothing sufficiently to deal with this problem, however, sometimes results in oversmoothing in other parts of the density. *Adaptive smoothing* methods exist: where the density is sparse, more smoothing occurs, where the density is less sparse, there is less smoothing.[21]
- If the data lie on a bounded interval, the kernel destiny estimates may have the property of estimating positive density at points that lie outside the domain of the variable. One approach is to transform the variable suitably. For many purposes we can simply ignore this problem. For other methods, again see Silverman.[22]

[23] S. Sheather and M. Jones, "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society, B*, **53**, 1991, 683–690.

11.4.2 An Application: The Wage Effects of Unions

Nonparametric density estimation can be a useful tool for exploratory data analysis. For instance, the presence of a large and moving spike (at the value of the minimum wage) can easily be seen in simple nonparametric estimates of the density of log wages for women over the period 1979–1989 (a period over which the minimum wage, adjusted for inflation, was falling), which suggests that the minimum wage plays an important role in explaining changes in wage inequality.[24] Nonparametric density estimation can also be used for more standard testing and inference problems.

Not surprisingly, nonparametric density estimation is very useful when the object of interest is a density. Consider the following problem: What effect do unions have on the distribution of wages? Although we shall focus on unions in particular, the analysis can be easily modified to consider other factors.

One way to begin would be to postulate two different wage equations, one for the union sector and another for the nonunion sector:

$$y^u = X^u\beta^u + \epsilon^u \tag{11.23}$$

$$y^n = X^n\beta^n + \epsilon^n \tag{11.24}$$

If selection issues are not a major problem (and some evidence suggests that they are not), the simplest way to proceed is to estimate separate regressions for the union sample and the nonunion sample to get estimates of β^u and β^n, respectively. The well-known Oaxaca decomposition proceeds exactly in this fashion and then computes the following:

$$\widehat{\bar{Y}_u^n} = \bar{X}_n\widehat{\beta^u}$$

$$\widehat{\bar{Y}_n^u} = \bar{X}_u\widehat{\beta^n}$$

where $\widehat{\beta^n}$ and $\widehat{\beta^u}$ are the OLS estimates from Eqs. (11.24) and (11.23), respectively, and $\bar{X}_n$ and $\bar{X}_u$ are the means of X variables in the nonunion and union sector.[25] $\widehat{\bar{Y}_n^u}$ is the mean salary of union workers had they been paid with the wage function in the nonunion sector. $\widehat{\bar{Y}_u^n}$ is the mean salary of nonunion workers if they had been paid according to the wage function in the union sector.

The effect of unions on the mean wages of union workers is then computed as the following difference:

$$\text{Union effect} = \bar{Y}_u - \widehat{\bar{Y}_n^u} \tag{11.25}$$

where $\bar{Y}_u$ is the actual mean wage in the union sector.

Although this approach is helpful if the object of interest is the mean, it is not very helpful for understanding the distributional consequences of unionism. An al-

[24] J. DiNardo, N. Fortin, and T. Lemieux, "Labor Market Institutions and the Distribution of Wages: 1973–1993. A Semi-Parametric Approach," *Econometrica*, in press.

[25] R. Oaxaca, "Male-Female Differentials in Urban Labor Markets," *International Economic Review*, **14**, 1973, 693–709.

ternative approach is to use nonparametric density estimation. Note that the definition of conditional probability yields the following representation of the overall distribution of wages:

$$g(w) = \int f(w \mid x)h(x)\,dx$$

where $f(w \mid x)$ is the conditional density of wages. It will also be useful to define two other densities. First, the observed density of wages in the nonunion sector is given by

$$g(w \mid u = 0) = \int f^n(w \mid x)h(x \mid u = 0)\,dx \tag{11.26}$$

where $f^n(w \mid x) \equiv f(w \mid x, u = 0)$. As before, $f^n(w \mid x)$ represents the structure of wages in the nonunion sector. Likewise, the observed density of wages in the union sector is given by

$$g(w \mid u = 1) = \int f^u(w \mid x)h(x \mid u = 1)\,dx$$

where $f^u(w \mid x) \equiv f(w \mid x, u = 1)$.

By analogy to the Oaxaca decomposition, we are interested in what distribution would prevail if all workers (not just nonunion workers) were paid under the wage structure in the nonunion sector or, more formally,

$$g^n(w) = \int f^n(w \mid x)h(x)\,dx \tag{11.27}$$

Estimation of the foregoing density can be made simple by noting that by Bayes' Law

$$h(x) = \frac{h(x \mid u = 0)\text{prob}(u = 0)}{\text{prob}(u = 0 \mid x)} \tag{11.28}$$

By substituting Eq. (11.28) into Eq. (11.27) we get the following:

$$g^n(w) = \int f^n(w \mid x)\frac{h(x \mid u = 0)\text{prob}(u = 0)}{\text{prob}(u = 0 \mid x)}\,dx \tag{11.29}$$

$$= \int \theta f^n(w \mid x)h(x \mid u = 0)\,dx \tag{11.30}$$

where $\theta = [\text{prob}(u = 0)]/[\text{prob}(u = 0 \mid x)]$. But notice that Eq. (11.30) is identical to Eq. (11.26) except for the *weight* θ.

We would like to know what the distribution of wages would be if everyone were paid nonunion wages. Our first choice might be to use the nonunion sample as an estimate. Unfortunately, the distribution of x characteristics of this sample does not reflect the sample of x characteristics for the population at large. For instance, the nonunion sample has too many Ph.D.'s and not enough blue-collar workers relative to a sample of the population at large. The solution: give more weight to members of the nonunion sample who are likely to be underrepresented.

The two probabilities are easy to estimate: prob(u = 0) is the proportion of nonunion members in the sample, and prob($u = 0 \mid x$) can be estimated by a discrete choice model like the Probit (discussed in Chapter 13) or nonparametrically by dividing up or grouping the sample by the characteristics x and calculating the proportion of individuals in each cell.

We illustrate the procedure using our sample of 1000 observations from the 1988 CPS. First, using the entire sample (that is, both union and nonunion workers) we estimate a simple probit of the form

$$\text{Union} = \Phi(\text{experience, experience squared, married, part-time, years of education}) \tag{11.31}$$

Next, we employ this equation to generate a predicted probability of being a nonunion worker (note that, because the prediction equation generates the probability of being a union member, we had to calculate 1 less the estimated probability of being in a union). This gives us the term prob($u = 0 \mid x$). Next we calculate prob(u = 0) as the sample proportion of nonunion members and use this, along with the previous estimate, to generate θ.

We then apply kernel density estimates (using the Epanechnikov kernel) to the sample of *nonunion workers only* using our estimates of θ. This step yields the distribution of wages in the economy if everyone were paid according to the nonunion wage structure.

Figure 11.10 displays two distributions. The first, labeled "Density without unions," is the distribution that would have prevailed in the entire population if

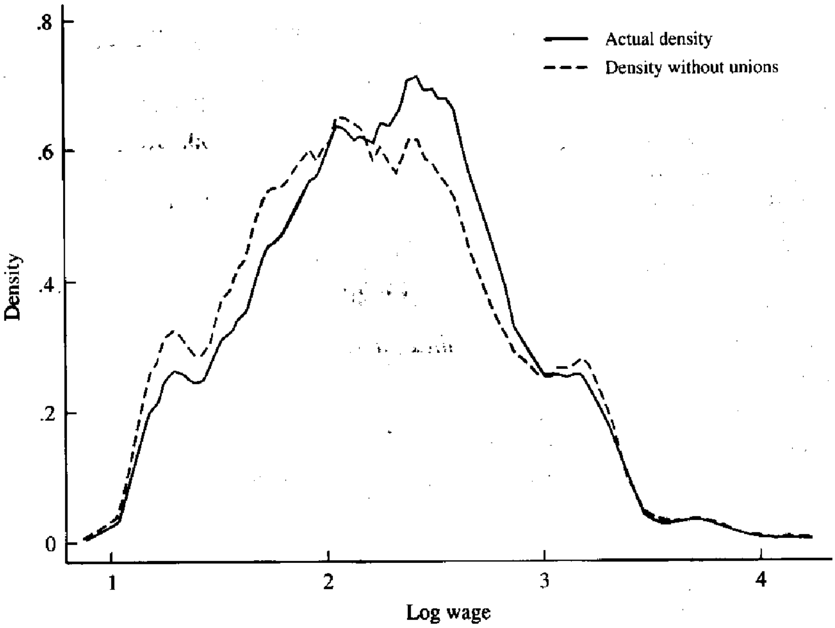


FIGURE 11.10
The effect of unions.

everyone had received the nonunion wage (the estimate described previously). The other distribution is a similar kernel density estimate using the *entire* sample *without* our θ weights. The difference between the two densities can be viewed as the *effect of unions.*

The estimates suggest that for men in the United States unions tend to equalize the distribution of wages. Note that the counterfactual density ("Density without unions") has less weight at the center of the distribution and more weight in the lower half of the distribution. A simple explanation for this result is that (*i*) the higher the wage workers would receive in the nonunion sector, the less likely they are to be unionized (the lowest-wage workers tend to be unionized) and (*ii*) unions raise the wages of low-wage workers the most relative to their nonunion counterparts. Note also that unions have little effect in the tails. This observation results from the fact that few workers at either tail are likely to belong to a union.

The hump in the lower tail of the distribution is interesting, as it happens to occur at the value of the minimum wage. Further exposition of the general approach and an analysis of the effect of the minimum wage can be found in DiNardo, Fortin, and Lemieux[24] and an application to the United States and Canada can be found in DiNardo and Lemieux.[26] The "hump" in the right tail of the distribution is interesting as well. There are several explanations, including the possibility that some persons "round" their wage to whole dollar amounts. For a fascinating discussion, see the paper by Card and Hyslop.[27]

11.5 NONPARAMETRIC REGRESSION

The heading of nonparametric regression covers a great deal of ground. We will deal here with only a small class of "smoothers" for the univariate model.

A nonparametric regression for the single regressor case can be described as follows:

$$y_i = m(x_i) + \epsilon_i \tag{11.32}$$

A nonparametric regression attempts to recover the function m which might be highly nonlinear. Nonparametric regression is quite simple when we have repeated observations on y for various levels of x. The easiest case is when x takes on a small number of discrete values. For example, let x take on three values (1, 2, 3). In that case a simple nonparametric regression would be

$$y_i = \beta_1 D_{i1} + \beta_2 D_{i2} + \beta_3 D_{i3} + \epsilon_i \tag{11.33}$$

where $D_{i,j} = 1(x_i = j)$ and where $1(\cdot)$ is the indicator function that takes the value

[26] J. DiNardo and T. Lemieux, "Diverging Wage Inequality in the U.S. and Canada: 1981–1989. Do Unions Explain the Difference?" mimeo, University of California-Irvine, Department of Economics, 1994.

[27] D. Card and D. Hyslop, "Does Inflation 'Grease the Wheels of the Labor Market,'" Industrial Relation Section Working Paper number 356, Princeton University, Dec. 1995.

of 1 if the statement inside the parentheses is true. In other words, we create a set of dummy variables for each level of the x variable. In this simple case, the coefficients β_1, β_2, and β_3 are just the means of y at each level of x. When x does not take on too many different values, this approach is often best. When x takes on a large number of values, however, such a method breaks down because typically we will not get repeated observations on the same x. In that case, other techniques may be required.

There are several different ways to recover the function $m(x_i)$. Consider a large class of "smoothers" of the form

$$\widehat{m(x)} = \frac{1}{n}\sum_{i}^{n} w_{ni}(x)y_i \tag{11.34}$$

The simplest nonparametric estimator described in Eq. (11.33) is a special case of this smoother where

$$w_{ni} = \frac{1(x_i = x)}{(1/n)\sum_i 1(x_i = x)}$$

Such a smoother also has an interpretation as a weighted regression. Note that $\widehat{m(x)}$ is the solution to the following problem:

$$\min_{\beta_0,\beta_1}\left(\frac{1}{n}\sum_{i=1}^{n} w_{ni}(x)(y_i - \beta_0 - \beta_1 x)^2\right) = \min\left(\frac{1}{n}\sum_{i=1}^{n} w_{ni}(x)[y_i - \widehat{m(x)}]^2\right) \tag{11.35}$$

That is, smoothing can be viewed as a series of weighted least-squares regressions, and $\widehat{m(x)}$ can be viewed as the *local least-squares estimate.*[28] Nonparametric regression of y on x involves running a regression at each x where the relationship $y = m(x_i)$ is to be estimated. Typically the weighting procedure gives the most weight to points close to x and less weight (or no weight) to points far from x.

Different types of smoothers can be differentiated by the choice of $w(\cdot)$. One popular method is called *loess.*[29] Loess is a *nearest neighbor method.* Instead of averaging the dependent variable across all the x's, loess averages only those y's in a "neighborhood" around x. To be more precise, loess can be summarized by two parameters, α and λ. The latter term takes on two values—1 or 2—and refers to whether the local regression function is linear or quadratic. When $\lambda = 1$ we have the conventional case, and loess can be described as follows:

$$\min_{\beta_0,\beta_1}\left(\frac{1}{n}\sum_{i=1}^{n} w_{ni}(x)(y_i - \beta_0 - \beta_1 x)^2\right)$$

[28] W. Härdle, *Applied Nonparametric Regression,* 1990, Cambridge University Press.

[29] According to W. Cleveland, *Visualizing Data,* 1993, Hobart Press, the term comes from the German *loess* which is short for *local regression.* A bit of wordplay is involved because loess is a term used by geologists to describe a deposit of silt or fine clay typically found in valleys—hence, a surface of sorts. This subtlety seems to have been lost, and *loess* is often rendered as *lowess.* Härdle,[28] for example, renders the term as the acronym LOWESS—LOcally WEighted Scatter plot Smoothing—although the procedure he describes is a slight variant of the one described here and by Cleveland.

TABLE 11.6
Data generated by $y = \sin(x) + \epsilon$

Observation no.	y	x	Observation no.	y	x
1	0.82	0.79	11	1.23	8.64
2	2.52	1.57	12	2.30	9.42
3	1.66	2.36	13	0.19	10.21
4	1.50	3.15	14	1.88	11.00
5	1.47	3.93	15	−1.71	11.78
6	1.60	4.71	16	−0.70	12.57
7	0.49	5.50	17	1.42	13.35
8	0.41	6.28	18	1.31	14.14
9	3.09	7.07	19	0.79	14.92
10	1.32	7.85	20	0.22	15.71

When $\lambda = 2$ the local linear regressions that loess calculates are quadratic in x instead of linear. In this case loess is the solution to

$$\min_{\beta_0,\beta_1,\beta_2}\left(\frac{1}{n}\sum_{i=1}^{n} w_{ni}(x)(y_i - \beta_0 - \beta_1 x - \beta_2 x^2)^2\right)$$

The most important term, however, is α, which calibrates the degree of smoothing. For a sample size n, let $q = \text{int}(\alpha n)$ where the function $\text{int}(\cdot)$ truncates its argument to the nearest integer and α is a number between 0 and 1. A nearest neighborhood of x_i is then the $2q + 1$ values of x that are nearest to (and include) x_i, that is, the set $(x_{i-q}, x_{i-(q-1)}, \ldots, x_i, \ldots, x_{i+(q-1)}, x_{i+q})$ and the set $(y_{i-q}, y_{i-(q-1)}, \ldots, y_i; \ldots, y_{i+(q-1)}, y_{i+q})$.

To illustrate, suppose we had the data in Table 11.6 sorted by x_i for convenience. If $\alpha = .1$ then loess will consider neighborhoods of size $q = 20 \times .1 = 2$ around x_i. Consider computing a value for the regression function at $x = 7.07$, or observation number 9. The neighborhood defined at this point when $q = 2$ are observations 7 through 11. The loess estimate is then the predicted value from a (weighted) linear regression of y on x for these observations.

More formally, if the data are sorted from smallest to largest, define the set

$$J_x = x_i : x_i \text{ is in a } q\text{-neighborhood}$$

and define a weight for observation in the set J_x,

$$w_j = \left[1 - \left(\frac{|x_j - x_i|}{\Delta}\right)^3\right]^3$$

where $\Delta = \max_j |x_j - x_i|$ and the weighting function is known as the *tricube*. The procedure is then quite straightforward:

1. For each x_i find its q-length neighborhood.
2. For each x in each neighborhood, calculate the foregoing weight.
3. Run a weighted least-squares regression of y on x in each neighborhood and predict $\hat{y}_i$—the smoothed value for y_i.

To give a reader a sense of the importance of choosing the smoothing parameter α, we perform a loess regression on the data in Table 11.6, which are generated by

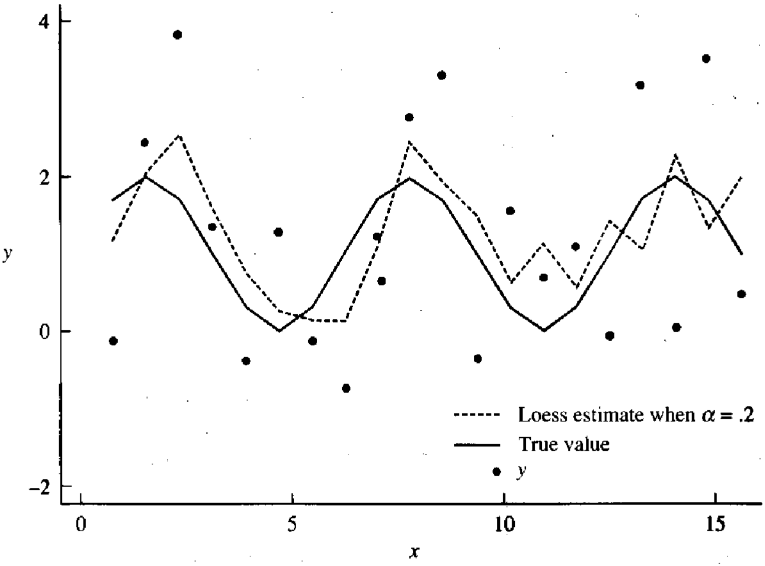


FIGURE 11.11
A simple loess estimate.

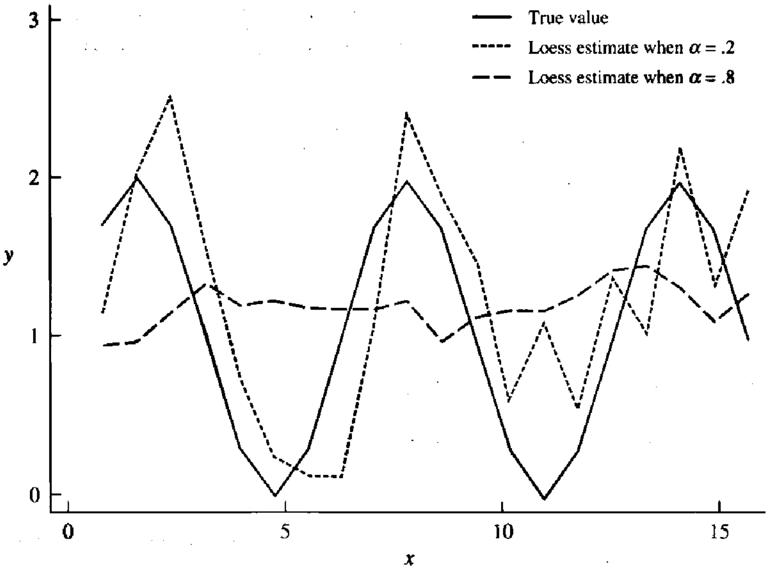


FIGURE 11.12
The importance of neighborhood size.

the equation

$$y_i = \sin(x_i) + \epsilon \tag{11.36}$$

where ϵ is a standard normal variate. Figure 11.11 displays the original data, the "true" regression function, and the loess estimate with a value of α of .2. At this level, the estimate seems to do a good job of reproducing the underlying variation.

Oversmoothing, however, can cause significant bias. Figure 11.12 compares two loess estimates ($\alpha = .2$ and $\alpha = .8$) to the truth. Apparently .8 is too large. The curve is too "smooth" and it does not do a very good job at reproducing the underlying variation. Again as we saw for nonparametric density estimation, since it is easier to smooth with the eye than to unsmooth with the eye, undersmoothing is to be preferred to oversmoothing.

Loess is not the only type of smoother, although it has the desirable property of closely *following the line.* An alternative is to use all the data and a conventional kernel estimator:

$$w_{ni}(x) = w_{h_i}(x) = \frac{(1/h)K[(x - x_i)/h]}{\widehat{f_h(x)}} \tag{11.37}$$

where

$$\widehat{f_h(x)} = \sum_i^n \frac{1}{nh} K\left(\frac{x - x_i}{h}\right)$$

This is clearly recognizable as the Rosenblatt-Parzen density we encountered in Section 11.4. These weights were originally proposed by Nadaraya and by Watson, so this estimator is often referred to as the *Nadaraya-Watson* estimator.[30] As we saw with density estimation, the choice of kernel is less important than the choice of bandwidth.

11.5.1 Extension: The Partially Linear Regression Model

One serious limitation of the nonparametric regression methods that we have considered so far is that it does not extend well to more than two variables—this is the well-known curse of dimensionality. As we extend beyond two dimensions, estimation, interpretation, and display of the results become more difficult.

One compromise is the partially linear regression model,

$$y_i = X_i\boldsymbol{\beta} + f(z_i) + \epsilon_i \tag{11.38}$$

where one part of the model is linear—the Xs—and a single variable has a potentially nonlinear relationship with y. The simplest method is to divide z into categories in the manner of Eq. (11.33) if the variable is categorical. If the variable is not categorical, the variable can still be grouped or divided into categories; otherwise another approach is necessary.

Suppose the variable z represents years of work experience. One approach would be to include z, z^2, and z^3 as regressors and perform OLS. In many applications this

[30] E. A. Nadaraya, "On Estimating Regression," *Journal of Probability Applications,* **10,** 1964, 186-190; and G. S. Watson, "Smooth Regression Analysis," *Sankhya A,* **26,** 1964, 101–116.

may be adequate. One drawback to such an approach may be that it is particularly sensitive to, say, a few unusually large values of z. It is also uncertain a priori how many polynomial terms need to be included to approximate $f(z)$ sufficiently.

Yet another approach is to approximate $f(z)$ by a piecewise linear function. For instance, if z ranged from 0 to 100, we might replace $f(z)$ by

$$\gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \gamma_4 z_4$$

where
$$\begin{aligned} z_1 &= z \text{ if } z \leq 20, \text{ and 0 otherwise} \\ z_2 &= z \text{ if } 20 < z \leq 40, \text{ and 0 otherwise} \\ z_3 &= z \text{ if } 40 < z \leq 60, \text{ and 0 otherwise} \\ z_4 &= z \text{ if } z > 60, \text{ and 0 otherwise} \end{aligned}$$

Other divisions of the data, of course, are possible. One advantage of all these approaches is that they fit comfortably into the traditional linear framework.

A recent paper by Estes and Honoré suggests that an alternative may be possible.[31] They observe that in the partially linear model it is possible to treat the nonlinear portion of the model as a *fixed effect*. The estimator proceeds as follows:

1. Sort the data by ascending values of z.
2. "First difference" adjacent values of y and x in the sorted data, constructing
$$\Delta y = y_i - y_{i-1} \qquad \Delta X = X_i - X_{i-1}$$
3. Run the OLS regression,
$$\Delta y = \Delta X\beta + \text{error}$$

Estes and Honoré show that under some conditions (in particular, when z is bounded) the $\hat{\beta}$ estimated this way is consistent. The consistency of β in this step is a consequence of the fact that as $n \to \infty$ adjacent z's are closer and closer to each other. As a consequence, if f is continuous, the difference $\Delta f(z_i)$ approaches zero at a fast rate and hence can be ignored provided n is large enough.

This result suggests the following two-step approach to estimating the shape of $f(z)$:

1. Calculate consistent estimates of β as just described.
2. Compute the "residuals,"
$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$$
3. Run a nonparametric regression of the form
$$\hat{u}_i = m(z_i) + \text{error}$$

As empirical applications of this approach currently appear to be scarce, this technique should be applied with caution. On the other hand, such a technique would seem useful for exploratory data analysis. It is especially attractive because it is simple and is computable with standard statistical software.

[31] E. Estes and B. Honoré, "Partially Linear Regression Using One Nearest Neighbor," Princeton University Department of Economics, March 1995, mimeo.

11.6 REFERENCES

For a useful discussion of Monte Carlo methods see the chapter by Hendry.[4] Use of permutation tests in econometrics is rare, but the article by Schmoyer is a good starting point.[12] For a basic discussion of the bootstrap see Efron and Tibshirani.[32] For applications of the bootstrap to econometrics (paying attention to time series applications we ignore here) see J. Jeong and G. S. Maddala.[15] For a discussion of nonparametric density estimation an easy-to-read starting point is the book by Silverman.[22] For details on the particular application of nonparametric density estimation see DiNardo and Lemieux.[26] Härdle provides a nice although technical introduction to nonparametric regression.[28] In addition, he covers many topics omitted here. Some interesting new developments in nonparametric regression have been omitted owing to considerations of space and their (current) unavailability in most current software. For example, J. Fan describes a method for "design-adaptive nonparametric regression" that, although similar to loess, dominates it asymptotically and appears to perform very well in finite samples. For those not deterred by a small amount of programming, this procedure is a worthy alternative to consider.[33] Finally, the text by Cleveland is a wonderful easy-to-read introduction to nonparametric regression and other ways to visualize data.[29]

PROBLEMS

1. It has been argued that simulation (in particular, hands-on simulation, where the simulations are computed manually) can be a helpful teaching device. In particular, consider the famous "Monty Hall" problem.[34] There are three doors, behind one of which is a desirable prize. The remaining two doors have no prize. The contestant chooses a door but does not open it. Next, the host exposes one of the remaining doors, behind which there is no prize. (That is, although the host always reveals what is behind one of the other doors, she does not reveal where the prize is.) The contestant can either remain with her first choice or choose the remaining door.
   a. Should the contestant remain with her first choice or switch, or is the choice of strategy immaterial?
   b. After the host has revealed one of the doors, if the contestant switches from her first choice, what is the probability that she wins?
   c. Create a Monte Carlo to simulate this game and compute the probability that the contestant wins if she chooses one of the following:
      i. Always switching
      ii. Never switching (i.e., remaining with her first choice)

[32] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, 1993, Chapman & Hall.

[33] J. Fan, "Design-adaptive Nonparametric Regression," *Journal of the American Statistical Association*, **87**, 1992, 998–1004.

[34] Monty Hall was the moderator of a very popular U.S. TV game show called "Let's Make a Deal," where contestants would dress in strange clothing and compete to win a large variety of consumer goods. The show had a contest similar to the one described here.

d. How many Monte Carlo simulations are necessary to construct a 95 percent confidence interval such that the extremes of the confidence interval are no greater than .01 in absolute value from the correct answer?

2. A researcher has a convenient way to generate K independent standard normal random variables $c_1, c_2, \ldots, c_k$, where

$$C = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix}$$

and he wants to generate K correlated normal variables such that

$$\boldsymbol{P} \overset{d}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} \quad$ and $\quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2K} \\ \vdots & \cdots & \ddots & \vdots \\ \sigma_{K1} & \cdots & \cdots & \sigma_K^2 \end{bmatrix}$

The matrix $\boldsymbol{P}$ can be generated as follows:

$$P = \mu + AC$$

where A is the Choleski decomposition of Σ. That is:

$$AA' = \Sigma$$

where A is a lower triangular matrix. Specialize this to the case of $K = 2$ and show that

$$\boldsymbol{P} = \begin{bmatrix} \mu_1 + \sigma_1 c_1 \\ \mu_2 + c_1 \rho c_2 \sqrt{1 + \rho^2}\sigma_2 \end{bmatrix}$$

where $\rho = \sigma_{12}/\sigma_1\sigma_1$.

3. Prove Eq. (11.12).

4. A double-blind experiment was performed to test the effect of caffeine on the capacity to tap one's fingers. Several male college students were trained in finger tapping. One group received no caffeine; the other group received 200 milligrams of caffeine. Use the Monte Carlo version of the permutation test to test whether the distribution of finger taps is the same for the two groups. Compare your results to a conventional t test.[35]

```
No caffeine:
242 245 244 248 247 248 242 244 246 242
200 milligrams of caffeine:
246 248 250 252 248 250 246 248 245 250
```

5. Perform the following Monte Carlo. Let the true model be given by

$$y = \beta_0 + \beta_1 x_t + \epsilon_t$$

where $$\epsilon_t = \rho\epsilon_{t-1} + \nu_t$$

[35] N. R. Draper and H. Smith, *Applied Regression Analysis*, 2nd edition, 1981, John Wiley & Sons, 425.

where ν_t and x_t are standard independent normal variables, β_0 and β_1 are both zero, and $\rho = .1$ for a sample size of $N = 100$. Do the following:

a. Compute the Durbin-Watson test statistic for 1000 Monte Carlo samples.

b. Use the following permutation test to compute an alternative for each Monte Carlo sample:

i. Compute

$$d = \frac{\sum_{i=1}^{i=100} \hat{e}_t \hat{e}_{t-1}}{\sum_{i=1}^{i=100} \hat{e}_t^2}$$

for the Monte Carlo sample.

ii. For each Monte Carlo sample construct L samples of size 100 of ϵ (i.e., randomly reorder the data).

iii. Compute

$$\hat{d}_l = \frac{\sum_{i=1}^{i=100} r_t r_{t-1}}{\sum_{i=1}^{i=100} r_t^2}$$

for each of these L samples, where r is the reordered sample of residuals.

iv. Compute the percentile of d in the L sample of $\hat{d}_l$'s.

c. Compute the proportion of times that you reject the hypothesis that the errors are serially uncorrelated and evaluate the power of the two tests. You may vary L and ρ to learn more.

6. Using the data in Fig. 11.3 use Monte Carlo methods to compute the distribution of the t statistic. Construct a nonparametric kernel density estimate of this distribution and compare it to the parametric t distribution you would estimate for these data under the assumption that the data are normal.

CHAPTER 12

Panel Data

In this chapter we discuss techniques for panel data. These are **repeated observations on the same set of cross-section units.** First, let us establish some notation:

y_{it} = the value of the dependent variable for cross-section unit i at time t where $i = 1, \ldots, n$ and $t = 1, \ldots, T$

X_{it}^{j} = the value of the jth explanatory variable for unit i at time t. There are K explanatory variables indexed by $j = 1, \ldots, K$.

We will restrict our discussion to estimation with *balanced* panels. That is, we have the same number of observations on each cross-section unit, so that the total number of observations is $n \cdot T$. When $n = 1$ and T is large, we have the familiar time-series data case. Likewise, when $T = 1$ and n is large, we have cross-section data. Panel data estimation methods refer to cases when $n > 1$ and $T > 1$. In this chapter, we will deal only with cases where n is large relative to T. The asymptotic theory we will employ assumes that n goes to infinity and T is fixed.

The most common way of organizing the data is by decision units. Thus, let

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix} \qquad \boldsymbol{X}_i = \begin{bmatrix} X_{i1}^1 & X_{i1}^2 & \cdots & X_{i1}^K \\ X_{i2}^1 & X_{i2}^2 & \cdots & X_{i2}^K \\ \vdots & \vdots & \ddots & \vdots \\ X_{iT}^1 & X_{iT}^2 & \cdots & X_{iT}^K \end{bmatrix} \qquad \boldsymbol{\epsilon}_i = \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{iT} \end{bmatrix} \tag{12.1}$$

where ϵ_{it} refers to the disturbance term for the ith unit at time t. Often the data are stacked to form

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_n \end{bmatrix} \tag{12.2}$$

where $\boldsymbol{y}$ is $nT \times 1$, $\boldsymbol{X}$ is $nT \times k$, and $\boldsymbol{\epsilon}$ is $nT \times 1$. The standard linear model can be expressed as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{12.3}$$

where
$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

The models we will discuss in this chapter are all variants of the standard linear model given by Eq. (12.3). The models will differ in their assumptions about the nature of the disturbance term ϵ. Models in which the coefficients vary across time or individuals and models that include lagged dependent variables are beyond the scope of the current discussion.

12.1 SOURCES AND TYPES OF PANEL DATA

Before turning to estimation issues, let us discuss some common sources and types of panel data.

One of the most frequently used panel data sets is the Panel Study of Income Dynamics (PSID), collected by the Institute of Social Research at the University of Michigan. Since 1968, researchers have collected information on more than 5000 families. Once a year family members are reinterviewed about their economic status; and information is collected about job changes, income changes, changes in marital status, and many other socioeconomic and demographic characteristics.

The Survey of Income and Program Participation (SIPP, U.S. Department of Commerce, Bureau of the Census) is similar to the PSID, although it covers a shorter time period and respondents are interviewed about their economic condition four times a year. Panel data on the economic conditions of Canadians have become available with the Canadian Labor Market Activity Survey (LMAS). Similar data sets are now available for an increasing number of countries.

Another type of panel data set consists of repeated observations on larger entities, such as individual states of the United States. One example comes from a study by David Card on the effect of minimum wage laws on employment.[1] For this study, Card collected information by state on youth employment and unemployment rates, school enrollment rates, average wages, and other factors for the period 1976–1990. The decision unit in this case is a particular state.

Yet another type of panel data involves grouping cross-sectional data into relatively homogeneous classes first. One common approach is to group individuals by age, sex, and educational status. If the process is repeated for cross-sectional data

[1] D. Card, "Using Regional Variation in Wages to Estimate the Employment Impacts of the Minimum Wage," *Industrial and Labor Relations Review*, **46** (1), 1992, 22–37.

from other time periods, these groups can be treated as a continuous albeit "synthetic" cohort. For an example of this approach see Deaton.[2]

12.2 THE SIMPLEST CASE—THE POOLED ESTIMATOR

We begin by considering the simplest estimation method, which proceeds by essentially ignoring the panel structure of the data. Stack the data as described in Eq. (12.1) and let the model be given by

$$y = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{12.4}$$

where now we assume that $\epsilon_{it} \sim \text{iid}(0, \sigma^2)$ for all i and t. That is, for a given individual, observations are serially uncorrelated; and across individuals and time, the errors are homoscedastic.

Estimation of this model is straightforward. The assumptions we have made correspond to the classic linear model. Efficient estimation proceeds by stacking the data as already shown and using OLS. By assuming each observation is iid, however, we have essentially ignored the panel structure of the data. Although this estimation method is the easiest, it is often not appropriate for reasons that we now pursue.

12.3 TWO EXTENSIONS TO THE SIMPLE MODEL

Our starting point is the following model:

$$y_{it} = \boldsymbol{X}_{it}\boldsymbol{\beta} + \epsilon_{it} \tag{12.5}$$

where for the typical case the number of individuals is large, and the number of time periods is small. We go one step further and specify the following error structure for the disturbance term:

$$\epsilon_{it} = \alpha_i + \eta_{it} \tag{12.6}$$

where we *assume* that η_{it} is uncorrelated with X_{it}. The first term of the decomposition, α_i, is called an individual effect. In this formulation, our ignorance has two parts—the first part varies across individuals or the cross-section unit but is constant across time; this part may or may not be correlated with the explanatory variables. The second part varies unsystematically (i.e., independently) across time and individuals. This formulation is the simplest way of capturing the notion that two observations from the same individual will be more "like" each other than observations from two different individuals.

A large proportion of empirical applications involve one of the following assumptions about the individual effect:

[2] A. Deaton, "Panel Data from a Series of Repeated Cross-Sections," *Journal of Econometrics*, **30**, 1985, 109–126.

1. *Random effects model:* α_i is uncorrelated with X_{it}.
2. *Fixed effects model:* α_i is correlated with X_{it}.

The nomenclature is unfortunate and not used uniformly in different literatures. We have adopted the terminology that has filtered down to most applied researchers in economics (and statistical packages). To avoid confusion it might be better if the models were given different names: the relevant distinction between the two models is not whether the effect is fixed or not. The distinction is whether the effect is correlated with the explanatory variables. The nomenclature is well-established, however, and we will adopt it here.[3]

12.4 THE RANDOM EFFECTS MODEL

The random effects model has the following structure:

$$y_{it} = X_{it}\boldsymbol{\beta} + \epsilon_{it} \tag{12.7}$$

where

$$\epsilon_{it} = \alpha_i + \eta_{it} \tag{12.8}$$

It is important to stress that the substantive assumption that distinguishes this model from the fixed effects model is that the time-invariant person-specific effect α_i is uncorrelated with X_{it}. Recall that this orthogonality condition, along with our assumption about η_{it}, is sufficient for OLS to be asymptotically unbiased (see Chapter 10). Why not then merely do OLS?

The problem is twofold. When the true model is the random effects model,

1. OLS will produce consistent estimates of β but the standard errors will be understated.
2. OLS is not efficient compared to a feasible generalized least-squares (GLS) procedure.

In essence, the random effects model is one way to deal with the fact that T observations on n individuals are not the same as observations on nT different individuals. The solution is straightforward. First, we derive an estimator of the covariance matrix of the error term. Second, we use this covariance structure in our estimator of $\boldsymbol{\beta}$.

It will be helpful to be a bit more explicit about the precise nature of the error:

$$\begin{aligned} E[\boldsymbol{\eta}] &= 0 & E[\boldsymbol{\eta}\boldsymbol{\eta}'] &= \sigma_\eta^2 \boldsymbol{I}_{nT} \\ E[\alpha_i\alpha_j] &= 0, \text{for } i \neq j & E[\alpha_i\alpha_i] &= \sigma_\alpha^2 \\ E[\alpha_i\eta_{jt}] &= 0 & E[\alpha_i] &= 0 \end{aligned} \tag{12.9}$$

where all expectations are conditional on X. Given these assumptions, we can write the error covariance of the disturbance term of each individual cross-section unit:

[3] A good discussion of the difference between fixed and random effects appears in S. Searle, G. Cassella, and C. McCullouch, *Variance Components*, John Wiley & Sons, 1992, 3.

$$E[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i'] = \sigma_\eta^2 \boldsymbol{I}_T + \sigma_\alpha^2 \boldsymbol{ii}' = \begin{bmatrix} \sigma_\eta^2 + \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\eta^2 + \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\eta^2 + \sigma_\alpha^2 \end{bmatrix} \tag{12.10}$$

where $\boldsymbol{i}$ is a $T \times 1$ vector of ones. When the data are organized as in Eq. (12.2) the covariance of the error term for all the observations in the stacked model (12.3) can be written as

$$\boldsymbol{\Omega} = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma} = E[\boldsymbol{\epsilon\epsilon}'] = \begin{bmatrix} \boldsymbol{\Sigma} & 0 & \dots & 0 \\ 0 & \boldsymbol{\Sigma} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \boldsymbol{\Sigma} \end{bmatrix} \tag{12.11}$$

where $\boldsymbol{\Sigma} = E[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i']$ is the $T \times T$ matrix given in Eq. (12.10). The block diagonality of $\boldsymbol{\Omega}$ makes finding an inverse simpler, and we can focus on finding the inverse of $\boldsymbol{\Sigma}$. It is straightforward but tedious to show that

$$\boldsymbol{\Sigma}^{-1/2} = \frac{1}{\sigma_\eta}\left[\boldsymbol{I}_T - \left(\frac{1-\theta}{T}\boldsymbol{ii}'\right)\right]$$

where

$$\theta = \sqrt{\frac{\sigma_\eta^2}{T\sigma_\alpha^2 + \sigma_\eta^2}} \tag{12.12}$$

is an unknown quantity that must be estimated.

Feasible GLS requires that we get estimates of the unknown quantities in Eq. (12.12). In particular, we need estimates of the variances σ_η^2 and σ_α^2 in θ. Simple analysis of variance arguments are enough to derive consistent estimators. However, they can be equivalently derived as the appropriately modified sum of squared errors from two different estimators. It will be useful therefore to discuss these two estimators first. In the process we will develop a simple way to compute the random effects estimator.

12.5 RANDOM EFFECTS AS A COMBINATION OF WITHIN AND BETWEEN ESTIMATORS

We consider two estimators that are consistent but not efficient relative to GLS. The first one is quite intuitive: convert all the data into individual specific averages and perform OLS on this "collapsed" data set. Specifically, perform OLS on the following equation:

$$\overline{y_{i\cdot}} = \overline{\boldsymbol{X}_{i\cdot}}\boldsymbol{\beta} + \text{error} \tag{12.13}$$

where the ith term $\overline{y_{i\cdot}}$ is

$$\overline{y_{i\cdot}} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$$

and $\overline{X_{i\cdot}}$ is defined similarly. To put this expression in matrix terms, stack the data as before and define a new $nT \times n$ matrix D, which is merely the matrix of n dummy variables corresponding to each cross-section unit. Now define $\boldsymbol{P_D} = \boldsymbol{D(D'D)^{-1}D'}$, a symmetric and idempotent matrix. Premultiplying by this matrix transforms the data into the means described in Eq. (12.13): the predicted value of $\boldsymbol{y}_i$ from a regression on nothing but the individual dummies is merely $\overline{y_{i\cdot}}$.

The $\hat{\boldsymbol{\beta}}$ estimated this way is called the *between estimator* and is given by

$$\widehat{\boldsymbol{\beta}_B} = (\boldsymbol{X'P_DX})^{-1}\boldsymbol{X'P_Dy} \tag{12.14}$$

The between estimator is consistent (though not efficient) when OLS on the pooled sample is consistent. In other contexts, this estimator is sometimes called a *Wald estimator* because, if T is long enough, such an estimator is robust to classical measurement error in the X variables (provided that the orthogonality condition is satisfied with the correctly measured data). This interpretation is easiest to understand if one notices that the estimator corresponds to 2SLS (two-stage least squares), using the person dummies as instruments.[4]

We can also use the information "thrown away" by the between estimator. Define $\boldsymbol{M_D} = \boldsymbol{I}_{nT} - \boldsymbol{D(D'D)^{-1}D'}$, which is also a symmetric idempotent matrix. If we premultiply the data by M_D and compute OLS on the transformed data we can derive the following *within estimator:*

$$\begin{aligned}\widehat{\boldsymbol{\beta}_W} &= [(\boldsymbol{M_DX})'(\boldsymbol{M_DX})]^{-1}(\boldsymbol{M_DX})'(\boldsymbol{M_Dy}) \\ &= (\boldsymbol{X'M_DX})^{-1}\boldsymbol{X'M_Dy}\end{aligned} \tag{12.15}$$

which is merely the estimator that would result from running OLS on the data including a full set of dummy variables. The matrix $\boldsymbol{M_D}$ can be interpreted as a residual-maker matrix: this interpretation represents an application of the Frisch–Waugh–Lovell theorem discussed in Appendix 3.2. Premultiplying by this matrix transforms the data into residuals from auxiliary regressions of all the variables on a complete set of individual specific constants. Since the *predicted value* from such a regression is merely the individual specific mean, the *residuals* are merely deviations from person-specific means. Specifically, Eq. (12.15) is equivalent to performing OLS on the following equation:

$$y_{it} - \overline{y_{i\cdot}} = (\boldsymbol{X}_{it} - \overline{\boldsymbol{X}_{i\cdot}})\boldsymbol{\beta} + \text{error} \tag{12.16}$$

As we discuss shortly, if the assumptions underlying the random effects model are correct, the within estimator is also a consistent estimator, but it is not efficient.

[4] The term *Wald estimator* should not be confused with the Wald test. It is so named since it was proposed in a paper by A. Wald, "The Fitting of Straight Lines if Both Variables Are Subject to Error," *Annals of Mathematical Statistics,* **II**, 1940, 284–300. A simple application of this approach can be found in O. Ashenfelter, "Macroeconomic and Microeconomic Analyses of Labor Supply," *Carnegie-Rochester Conference Series on Public Policy,* **21**, 1984, 117–156. J. Angrist, "Grouped-Data Estimation and Testing in Simple Labor-Supply Models," *Journal of Econometrics.* **47** (2/3), 1991, 243–266, elaborates further on Ashenfelter's approach and provides additional details. A similar approach is discussed in A. Deaton, "Panel Data from a Time-Series of Cross-Sections," *Journal of Econometrics,* **30**, 1985, 109–126, which also considers the bias introduced by the use of imprecise or "error-ridden" sample group means instead of population group means.

This defect is clear since we have included n unnecessary extra variables. It is called the within estimator because it uses only the variation *within* each cross-section unit.

Notice that the pooled OLS estimate is just a weighted sum of the between and within estimators:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}(X'M_Dy + X'P_Dy) \\ &= (X'X)^{-1}X'M_DX\hat{\boldsymbol{\beta}}_W + (X'X)^{-1}X'P_DX\hat{\boldsymbol{\beta}}_B\end{aligned}$$

Recall our discussion in Section 6.7 about GLS estimation in the presence of auto-correlated disturbances. Although OLS was generally consistent, it was inefficient since it did not incorporate our a priori knowledge about the form of the serial correlation. In the random effects case, OLS on the pooled data fails to use information about the heteroscedasticity that results from using repeated observations of the same cross-section units. The problem with the pooled OLS estimator is that it weights all observations equally. This treatment is not generally optimal because an additional observation on a person already in the data set is unlikely to add as much information as an additional observation from a new (independent) individual.

We are now in a position to compute the necessary quantities for a feasible GLS. Standard ANOVA suggests the following estimators:

$$\begin{aligned}\hat{\sigma}_\eta^2 &= \frac{1}{nT - nk - n}\hat{\boldsymbol{u}}_W'\hat{\boldsymbol{u}}_W \\ \hat{\sigma}_B^2 &= \frac{\hat{\boldsymbol{u}}_B'\hat{\boldsymbol{u}}_B}{n - k} \\ \hat{\sigma}_\alpha^2 &= \hat{\sigma}_B^2 - \frac{\hat{\sigma}_\eta^2}{T}\end{aligned}$$

where $\hat{\boldsymbol{u}}_W$ are the residuals from the within regression and $\hat{\boldsymbol{u}}_B$ are the residuals from the between regression. These can then be used to construct $\hat{\theta}$.

The student should be able to verify that these estimators are asymptotically unbiased estimates of the relevant variances. The formula for $\hat{\sigma}_\eta^2$ is derived, for example, by noting that the deviations-from-means transformation leaves only η in the error term. The purpose of introducing the estimators should now be clear: There is a simple way to compute the random effects estimator if one does not have access to a program that computes it automatically. A simple procedure to do so is as follows:

1. Compute the between and within estimators.
2. Use the residuals to calculate the appropriate variance terms.
3. Calculate $\hat{\theta}$.
4. Run OLS on the following transformed variables $\tilde{y}$ and $\tilde{X}$ where

$$\tilde{y}_{it} = y_{it} - \overline{y_{i\cdot}} + \hat{\theta}\overline{y_{i\cdot}} \tag{12.17}$$

$$\tilde{X}_{it} = X_{it} - \overline{X_{i\cdot}} + \hat{\theta}\overline{X_{i\cdot}} \tag{12.18}$$

The transformation is intuitively appealing. When there is no uncorrelated person-specific component of variance [$\sigma_\alpha^2 = 0$ in Eq. (12.12)], $\theta = 1$, and the random effects estimator reduces to the pooled OLS estimator.

12.6 THE FIXED EFFECTS MODEL IN THE TWO-PERIOD CASE

To the attentive reader, panel data might appear to have nothing particular to offer compared to simple cross-section data. Indeed, up to this point panel data have been presented as a more complex version of cross-section data where we have to deal with the unfortunate fact that there is not quite so much information in n individuals observed T times as there is with nT individuals. As discussed before, this limitation can be noted explicitly by observing that the random effects estimator reduces to the pooled estimator with a single cross section when the variance of the individual component is zero.

In fact, some have wryly noted that one advantage of panel data is that "it has created more work for econometricians." If this alone were true, there would certainly be little advantage to panel data! Instead, panel data estimation has grown in popularity because it has held out the promise of reducing a grave problem faced by most researchers: the lack of an adequate list of independent variables to explain the dependent variable.

To see this, let us start with an intuitive discussion of one fixed effect estimator (several are in common use).

Consider a simple two-period model ($t = 1, 2$) of the form

$$y_{it} = X_{it}\boldsymbol{\beta} + Z_i\boldsymbol{\delta} + \epsilon_{it} \tag{12.19}$$

where $\boldsymbol{X}$ = a matrix of explanatory variables that varies across time and individuals

$\boldsymbol{Z}$ = a matrix of variables observed by the econometrician that vary across individuals but for each individual are constant across the two periods

Similar to our previous development in the random effects model we define

$$\epsilon_{it} = \alpha_i + \eta_{it} \tag{12.20}$$

As before, we make the following assumptions (repeating Eq. 12.9):

$$\begin{aligned} E[\boldsymbol{\eta}] &= 0 & E[\eta\eta'] &= \sigma_\eta^2 I_{nT} \\ E[\alpha_i\alpha_j] &= 0, \text{for } i \neq j & E[\alpha_i\alpha_i] &= \sigma_\alpha^2 \\ E[\alpha_i\eta_{jt}] &= 0 & E[\alpha_i] &= 0 \end{aligned} \tag{12.21}$$

where all expectations are conditional on $\boldsymbol{X}$ and $\boldsymbol{Z}$. The substantive difference between the present case and the random effects model involves one further assumption regarding the individual specific effect. Letting $\boldsymbol{W}_{it} = [\boldsymbol{X}_{it}\ \boldsymbol{Z}_i]$ we now assume that

$$E[\boldsymbol{W}_{it}'\epsilon_{it}] \neq 0 \tag{12.22}$$

In particular we are concerned that our independent variables are correlated with α. The failure of this orthogonality assumption has important consequences. Consider OLS estimation on only the first-period data:

$$y_{i1} = \boldsymbol{X}_{i1}\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\delta} + \epsilon_{i1} \tag{12.23}$$

Unlike the random effects case, a consequence of Eq. (12.22) is that *OLS will be biased.* The extent and direction of the bias will depend on the precise nature of the relationship between the individual specific effect and the other explanatory variables. The bias can be analyzed in a fashion similar to our discussion of omitted variables in Chapter 8. A useful expedient is to imagine that we could run the following regression in the population:

$$\alpha_i = \boldsymbol{W}_{it}\boldsymbol{\pi} + \text{error} \tag{12.24}$$

The population coefficients $\boldsymbol{\pi}$ of this linear projection represent the bias. For example, if $\hat{\beta}_2$ is the OLS coefficient on the second explanatory variable from Eq. (12.23) and π_2 the population parameter from the same explanatory variable in the linear projection described in Eq. (12.24) then we can write

$$\text{plim}\,\hat{\beta}_2 = \beta_2 + \pi_2$$

where β_2 is the *true* population value of the coefficient on the second explanatory variable.

Using similar logic, we face the same problem with the OLS regression using only the second-period data:

$$y_{i2} = \boldsymbol{X}_{i2}\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\delta} + \epsilon_{i2} \tag{12.25}$$

The "magic" of panel data comes from noting that if Eqs. (12.23) and (12.25) are valid representations of the world, then any linear combination of the relationships is also true. Specifically,

$$\begin{aligned}
y_{i1} &= \boldsymbol{X}_{i1}\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\delta} + \epsilon_{i1}\\
y_{i2} &= \boldsymbol{X}_{i2}\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\delta} + \epsilon_{i2}\\
y_{i2} - y_{i1} &= (\boldsymbol{X}_{i2} - \boldsymbol{X}_{i1})\boldsymbol{\beta} + (\boldsymbol{Z}_i - \boldsymbol{Z}_i)\boldsymbol{\delta} + (\epsilon_{i2} - \epsilon_{i1})\\
\Delta y &= \Delta \boldsymbol{X}\boldsymbol{\beta} + \Delta \boldsymbol{Z}\boldsymbol{\delta} + \Delta\epsilon
\end{aligned} \tag{12.26}$$

where Δ is just the difference operator. For example, $\Delta X = Xi2 - X_{i1}$. Equation (12.26) is equivalent to

$$\Delta y = \Delta X\boldsymbol{\beta} + \Delta\eta \tag{12.27}$$

where we have noted that the time-invariant terms $\boldsymbol{Z}_i$ and α_i drop out after application of the difference operator. The key difference between Eq. (12.27) and our untransformed versions, Eqs. (12.23) and (12.25), is that the necessary orthogonality condition now holds on the transformed data. Specifically,

$$E[\Delta X'\Delta\eta] = 0 \tag{12.28}$$

The fortuitous consequence of this observation is that *the OLS regression on the transformed data yields unbiased estimates of the coefficients on the* ***X*** *variables.*

This is the essence of the fixed effects model. With panel data it is possible to obtain consistent estimates of parameters of interest even in the face of correlated omitted effects when OLS on individuals' cross sections would fail to do so! Intuitively, we are using individuals as controls for themselves. Three other lessons from this simple example will obtain in the more general fixed effects case:

1. *With fixed effects estimators, we cannot generally recover estimates of any time-invariant explanatory variables.*[5] When we remove the *unobserved* correlated effects α_i, we also remove the effects of any *observed* variable that is time-invariant. In our simple example, the differencing transformation causes both $\boldsymbol{Z}_i$ and α_i to drop out of our final estimating equation. All time-invariant effects receive the same treatment.
2. Of course, the flip side of Lesson 1 is that *the fixed effects estimator is robust to the omission of any relevant time-invariant regressors.* This indeed is the promise of fixed effect estimation. In principle, with fixed effects estimation we have greatly minimized the informational requirement necessary to satisfy the orthogonality condition. Shortly, we pursue an example to illustrate this point.
3. *When the random effects model is valid, the fixed effects estimator will still produce consistent estimates of the identifiable parameters.* That is, the orthogonality condition in Eq. (12.28) is obviously still valid when the random effects model describes the state of the world. (Although, as we will discuss below, in this case the fixed effects estimator is not efficient relative to the random effects estimator.)

12.7 THE FIXED EFFECTS MODEL WITH MORE THAN TWO TIME PERIODS

Before we discuss empirical applications of the fixed effects model, we derive it for a case with more than two time periods.

Because the fixed effects model starts with the presumption that $\text{cov}(\boldsymbol{X}_{it}, \alpha_i) \neq \boldsymbol{0}$, we must estimate the model *conditionally* on the presence of the fixed effects. That is, if we rewrite the model as

$$y_{it} = \boldsymbol{X}_{it}\boldsymbol{\beta} + \alpha_i + \eta_{it} \tag{12.29}$$

the α_i are treated as unknown parameters to be estimated. Note, however, that we cannot obtain consistent estimates of these additional parameters in the typical panel data case. In the typical case, T is small, and n is large. Our asymptotic theory is based on the idea that n gets larger and larger. In this setup, however, the number of parameters is growing at the same rate as the sample. Although we cannot estimate α_i consistently, we can estimate the remaining parameters consistently.

[5] If we have additional a priori information about the elements of the time-varying regressors it is sometimes possible to recover the coefficients on the time-invariant regressors. See J. Hausman and W. Taylor, "Panel Data and Unobservable Individual Effects," *Econometrica,* **49,** 1981, 1377–1398. Also see C. Hsiao, *Analysis of Panel Data,* 1986, Section 3.6.1.

To do so, we need only run the regression

$$y = X\boldsymbol{\beta} + D\boldsymbol{\alpha} + \boldsymbol{\eta} \tag{12.30}$$

where $D = \boldsymbol{I}_n \otimes \boldsymbol{i}_T$, as before, is a set of n dummy variables (one for each person). From the Frisch–Waugh–Lovell Theorem, we note that Eq. (12.30) is just the same as running a regression of each of our variables y and X on this set of dummies and then running the regression of the y residuals on the X residuals. The matrix that produces such residuals is the familiar $\boldsymbol{M}_D = \boldsymbol{I} - \boldsymbol{D}(\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'$. We can run OLS on the transformed variables $M_D y$ on $M_D X$ to get

$$\hat{\boldsymbol{\beta}}_W = (X'M_D X)^{-1}X'M_D y \tag{12.31}$$

This is merely the *within estimator* we derived before. The within estimator is only one possible fixed effects estimator. Any transformation that rids us of the fixed effect will produce a fixed effects estimator.

For example, consider the $T \times (T-1)$ matrix $\boldsymbol{F}$, which transforms a $1 \times T$ vector of repeated observations on the same individual to a $1 \times T-1$ vector of first differences by postmultiplication:

$$\boldsymbol{F} = \begin{bmatrix} -1 & 0 & 0 & \cdots & 0 \\ 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

This transformation is just the first difference transformation we pursued in the previous section. The reader should verify that this too rids the equation of the fixed effect.

Returning to our deviations from means approach, this purges the data of the fixed effects by removing means of these variables across individual cross-section units. That is, the predicted value of y, which belongs to group i, is just the mean of that group (in the same way that a regression of y on just a constant would yield a predicted value equal to the mean computed over the whole sample),

$$\overline{y_{i\cdot}} = \overline{X_{i\cdot}}\boldsymbol{\beta} + \overline{\alpha_i} + \overline{\eta_{i\cdot}} \tag{12.32}$$

Because the mean of α_i for individual i is merely α_i, we can difference Eqs. (12.29) and (12.32) to yield

$$y_{it} - \overline{y_{i\cdot}} = (X_{it} - \overline{X_{i\cdot}})\boldsymbol{\beta} + (\eta_{it} - \overline{\eta_{i\cdot}}) \tag{12.33}$$

It is evident then that either first-differencing or differencing from person-specific means will do the trick.[6]

In many applications, the easiest way to implement a fixed effects estimator with conventional software is to include a different dummy variable for each indi-

[6] The two estimators will not in general be numerically identical, however. In fact, if the two estimators give very different answers, it is evidence that the assumptions of the fixed effects model do not hold.

vidual unit. This method is often called the least-squares dummy variable (LSDV) method as in Eq. (12.30). If n is very large, however, it may be computationally prohibitive to compute coefficients for each cross-section unit. In that case, another way to implement a fixed effects estimator is as follows:

- Transform all the variables by subtracting person-specific means.
- Run OLS on the transformed variables.

This approach will work perfectly, apart from the fact that the standard errors need to be corrected. The correct standard errors are

$$\sigma_\eta^2 (X'M_D X)^{-1} \tag{12.34}$$

This result is almost exactly the same output one would get from the two-step procedure defined earlier. There is a problem however, in how the computer generates its estimate of σ_η^2. The correct way is

$$\hat{\sigma}_\eta^2 = \frac{u_W' u_W}{nT - n - k}$$

where the denominator is the correct degrees of freedom—nT observations minus (n computed means and k parameters). The output from the regression package, however, does not know that we estimated n means. Instead, it computes σ_η^2 as

$$\hat{\sigma}_{\text{computer}}^2 = \frac{u_W' u_W}{nT - k}$$

It is simple, however, to correct for the degrees of freedom because

$$\hat{\sigma}_\eta^2 = \frac{nT - k}{nT - n - k} \cdot \hat{\sigma}_{\text{computer}}^2 \tag{12.35}$$

The fact that the fixed effects estimator can be interpreted as a simple OLS regression of means-differenced variables explains why this estimator is often called a **within group estimator.** That is, it uses only the variation *within* an individual's set of observations.

12.8 THE PERILS OF FIXED EFFECTS ESTIMATION

Note that the fixed effects estimation solves the omitted variables problems by "throwing away" some of the variance that contaminates either OLS or the random effects estimator. So far, this has all been for the best. We now ask: can the cure be worse than the disease?

12.8.1 Example 1: Measurement Error in X

Unfortunately the answer is yes. Consider the simple linear model when the single explanatory variable is measured with error. Let the recorded value of the explanatory variable, x, equal the truth x^* plus a white noise measurement error:

$$x = x^* + \nu \tag{12.36}$$

where $x^* \sim N(0, \sigma^2_{x^*})$ and $\nu \sim N(0, \sigma^2_\nu)$. Now consider the plim of $\hat{\beta}_{\text{CS}}$ or the OLS estimate from a single cross section (CS) in the following model, when no fixed effects are present:

$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it} \tag{12.37}$$

We assume that $\text{cov}(x^*_{it}, \epsilon_{it}) = 0$, that is, apart from the measurement error in the explanatory variable, all the classical desirable properties hold. It can be shown that (see Chapter 5)

$$\text{plim}\,\hat{\beta}_{\text{CS}} = \beta - \beta\frac{\sigma^2_\nu}{\sigma^2_{x^*} + \sigma^2_\nu}$$

that is, our estimates will be *attenuated*, or biased toward zero, depending on what proportion of the total variance in our measurement $\sigma^2_x = \sigma^2_{x^*} + \sigma^2_\nu$ represents variation due to mismeasurement.

This attenuation bias can be greatly exacerbated by standard fixed effects estimation, especially if the explanatory variables are correlated across time. This result can be easily seen by considering the first-differenced (FD) version of the preceding model. It can be shown that

$$\text{plim}\,\beta_{FD} = \beta - \frac{\beta\sigma^2_\nu}{(1-\rho_x)[\sigma^2_{x^*} + \sigma^2_\nu]} \tag{12.38}$$

where $\rho_x = \text{cov}(x^*_{it}, x^*_{i,t-1})/\text{var}(x^*)$. It is now evident where the difficulty lies. Apart from the inclusion of the $1 - \rho$ term in the denominator, the bias looks like the standard case. In fact, when the x^*'s are completely uncorrelated across time ($\rho = 0$), this expression reduces to the one we derived for the plim of the OLS cross-section estimator.

Unfortunately, in many applications ρ is unlikely to be small. Consider a simple labor supply example,

$$h_{it} = w^*_{it}\beta + \epsilon_{it} \tag{12.39}$$

where h refers to hours worked in the reference period and w^* refers to the true value of the wage. However, only a mismeasured version w of the true wage exists.

The problem is this: Although there may be considerable variation in wages across individuals—$\sigma^2_{w^*}$ is reasonably large—there is typically much less variation in *changes* across t in wages. If true wages change slowly enough, changes in measured wages may mostly represent measurement error. In terms of Eq. (12.38), to the extent that a person's wages do not change much from year to year, ρ will be large. In an often-cited example, Altonji finds that as much as 80 percent of the variation in wage *changes* in the PSID may be due to measurement error.[7] Using additional data

[7] J. Altonji, "Intertemporal Substitution in Labor Supply: Evidence from Micro Data," *Journal of Political Economy*, **94**, 1986, S176–S215.

from a validation survey, Bound and Krueger find similar results for the PSID.[8] These observations suggest that, if left uncorrected, measurement error can seriously bias fixed effect estimates.

Griliches and Hausman note that with a little bit of structure it is possible to estimate the impact of the measurement error on the first-differenced estimator.[9] They observe that the problem is greatest when ρ is large, as is often the case when the duration of time between the two cross-section units is small. One can exploit this by comparing first-differenced estimators from adjacent time periods to "longer" differences. The basic idea is that, if the standard fixed effect model is correct, a first-differenced estimator using cross-section observations one period apart, Δ_1, should be the same as a first-differenced estimator using cross-section observations that are $L > 1$ periods apart, Δ_L. In the presence of measurement error, however, the attenuation bias in the longer differences should be less. That is, the signal-to-noise ratio is greater for the estimator using the longer differences. In terms of the formula we have just derived, we expect that in much panel data

$$(1 - \rho_x^1) < (1 - \rho_x^L) \tag{12.40}$$

where ρ_x^L represents the correlation between x's observed L periods apart and where $L > 1$.

It is worth noting, however, that although the cure *can* be worse than this disease, it does not necessarily follow that we should be content with the disease. It is clearly possible that there might be "correlated fixed effects" as well as measurement error. If the model is

$$y_{it} = x_{it}\beta + \alpha_i + \eta_{it} \tag{12.41}$$

where the measurement error in x is as before, it can be shown that the cross-section OLS estimator is inconsistent:

$$\text{plim}\,\hat{\beta}_{CS} = \beta + \frac{\text{cov}(x_{it}, \alpha_i)}{(\sigma_{x^*}^2 + \sigma_\nu^2)} - \frac{\beta\sigma_\nu^2}{(\sigma_{x^*}^2 + \sigma_\nu^2)} \tag{12.42}$$

Both fixed effects and measurement error are a source of bias for cross-section or pooled estimators. Only the latter is a problem for fixed effect estimation. In the class of models we have considered thus far, which estimator is less biased will depend on the extent of fixed effects, the extent of measurement error, and the extent to which the Xs are correlated across time. Sometimes independent information can be gathered on the extent of measurement error. Sometimes instrumental variable techniques can help.

[8] J. Bound and A. Krueger, "The Extent of Measurement Error in Longitudinal Earnings Data—Do Two Wrongs Make a Right?" *Journal of Labor Economics*, **9**, 1991, 1-24.

[9] Z. Griliches and J. Hausman, "Errors in Variables in Panel Data," *Journal of Econometrics*, **31**, 1986, 93-118.

12.8.2 Example 2: Endogenous *X*

In the two-period case, simple fixed effects estimators involve performing OLS on *changes* of the variables. Another possible problem with fixed effect estimators is that the variation in changes in X over time may not be exogenous. Solon provides a nice illustration of such a problem.[10] Consider estimating the extent to which unsafe jobs pay "compensating wage differentials": do otherwise similar workers receive above average wages for more dangerous jobs? The simplest model might be

$$w_{it} = \alpha + D_{it}\beta + \epsilon_{it} \tag{12.43}$$

where D is a dummy variable that equals 1 if a job is unsafe and 0 otherwise. To put the problem in its starkest terms, assume that there is no such premium and that workers in both types of jobs draw their wages from a single distribution. In this case, $\beta = 0$—there is no premium. Suppose further that workers prefer safe jobs to unsafe ones, and that one attempts to estimate β with a simple fixed effect model:

$$\Delta w_{it} = \Delta D_{it}\beta + \Delta\epsilon_{it} \tag{12.44}$$

where $\Delta w_{it} = w_{it} - w_{i,t-1}$. It is important to consider *why* we observe variation in ΔD_{it}. For consistent estimation we require that sample variation in ΔD_{it} be exogenous and uncorrelated with $\Delta\epsilon_{it}$. Unfortunately, this is almost certainly not the case. In this simple example, a worker with a safe job will only switch to an unsafe job ($\Delta D_{it} > 0$) if it happens to offer a higher wage ($\Delta\epsilon_{it} > 0$). On the other hand, workers may switch from an unsafe job to a safe job ($\Delta D_{it} < 0$) for either no increase in pay, or maybe even a small decrease. If this is the case, wage gains to people switching to more unsafe jobs will be larger than wage gains (or losses) to workers moving to safe jobs. As a consequence, fixed effect estimation will result in a positive β although the true β is zero. Put differently, ΔD_{it} is determined by $\Delta\epsilon_{it}$. In general, the direction of the bias caused by this type of self-selection will depend on the process that determines changes in the dependent variable.

It is possible to incorporate this type of dynamic self-selection directly into the estimation of panel data models. However, such models are much more complex and typically require some information to identify the selection process. A discussion of these issues is beyond the scope of this chapter.

It is worth pointing out that this dynamic selection problem is only one type of difficulty that arises from "selection." Another example of such a problem is selective attrition. In many panel data sets, persons leave the sample. Furthermore, many researchers will force their samples to be balanced (the same number of observations per cross-section unit) since estimation with balanced panels is typically much easier. Dropping individuals to balance a sample can often be a problem if the people who are dropped from the sample are different from those who remain in the sample. When this is the case, the panel data set may cease to be representative. The less representative the sample, the more difficult it is to make statements about the relevant population of interest.

[10] G. Solon, "The Value of Panel Data in Economic Research," *Panel Surveys*, eds. Daniel Kasprzyk, Greg Duncan, Graham Kalton, and M. P. Singh, John Wiley & Sons, 1989, 486–496.

12.9 FIXED EFFECTS OR RANDOM EFFECTS?

As we first pointed out in Section 12.3 the salient distinction between the two models is whether the time-invariant effects are correlated with the regressors or not. We also observed that *when the random effects model is valid, the fixed effects estimator will still produce consistent estimates of the identifiable parameters.* It would appear therefore that, in general, the fixed effects estimator is to be preferred to the random effects estimator unless we can be certain that we can measure all of the time-invariant factors possibly correlated with the other regressors.

Many researchers apparently find a precisely estimated fixed effects estimate more persuasive than a precisely estimated random effects estimate. This preference seems to be a consequence of the reasonable belief that, apart from purely experimental or quasi-experimental situations, it is unlikely that the fixed effects are uncorrelated with the regressors of interest. In the literature on the effect of union status on wages, for example, it is widely held that random effects and cross-section estimates are upward-biased estimates of the true effect: union workers are "more productive" in ways not observed by the econometrician.

As we learned from our discussion of the perils of fixed effects estimation, however, *it is possible* that the cure is worse than the disease. Whether this situation is common in applied work is debatable. More common appears to be the situation when neither the fixed effects estimator nor the random effects estimator is perfect. Consider again the literature on union wage effects. Although the evidence suggests that the random effects estimators are upward-biased estimates of the true effect, the fixed effects estimates are generally held to be *downward*-biased estimates of the true effects. This conclusion rests, among other things, on the observation that since actual changes in union status are relatively rare, a small amount of measurement error in the level of union status can have an important effect on the signal to total variance ratio of *changes* in union status.

In short, there is no simple rule to help the researcher navigate past the Scylla of fixed effects and the Charybdis of measurement error and dynamic selection. Although they are an improvement over cross-section data, panel data do not provide a cure-all for all of an econometrician's problems.

Next we turn to some simple specification tests in the panel data context and introduce a more sophisticated approach to panel data that yields additional insights.

12.10 A WU-HAUSMAN TEST

We have developed two estimators that have different properties depending on the correlation between α_i and the regressors. Specifically,

1. If the effects are uncorrelated with the explanatory variables, the random effects (RE) estimator is consistent and efficient. The fixed effects (FE) estimator is consistent but not efficient.

2. If the effects are correlated with the explanatory variables, the fixed effects estimator is consistent and efficient but the random effects estimator is now inconsistent.

This difference sets up a clear case for a Hausman test, defined simply as

$$H = (\hat{\boldsymbol{\beta}}_{\mathrm{RE}} - \hat{\boldsymbol{\beta}}_{\mathrm{FE}})'(\Sigma_{\mathrm{FE}} - \Sigma_{\mathrm{RE}})^{-1}(\hat{\boldsymbol{\beta}}_{\mathrm{RE}} - \hat{\boldsymbol{\beta}}_{\mathrm{FE}}) \tag{12.45}$$

The Hausman test statistic (discussed in Chapter 10) will be distributed asymptotically as χ^2 with k degrees of freedom under the null hypothesis that the random effects estimator is correct.

An alternative method is to perform a simple auxiliary regression. Let $\tilde{y}$ and $\tilde{X}$ be the data transformed for the random effects model as in Eqs. (12.17) and (12.18). Define the $\tilde{\tilde{X}}$ variables transformed for a fixed-effects regression as

$$\tilde{\tilde{X}}_{it} = X_{it} - \overline{X_{i\cdot}}$$

As in Chapter 10 on GMM, the Hausman test can be computed by means of a simple F test on γ in the following auxiliary regression:

$$\tilde{y} = \tilde{X}\beta + \tilde{\tilde{X}}\gamma + \text{ error} \tag{12.46}$$

The hypothesis being tested is whether the omission of fixed effects in the random effects model has any effect on the consistency of the random effects estimates.

As in previous chapters it is important to stress that if the random effects model "passes" this test, all is not necessarily well. In fact, one unfortunate result, which is not uncommon in applied work, is that the two estimators are not *significantly* different from each other. This may indicate only that there is not enough variance in the change in X to provide results precise enough to distinguish between the two sets of estimates. Furthermore, an imprecisely estimated fixed effect estimate that is not significantly different from zero is no reason to exult that the effect of the variable is zero. Indeed, if one wishes to argue that a variable does not matter, a precisely estimated zero is more convincing than an imprecisely estimated zero.

12.11 OTHER SPECIFICATION TESTS AND AN INTRODUCTION TO CHAMBERLAIN'S APPROACH

One instructive approach to understanding estimation and testing of panel data methods goes by a variety of names: Π-matrix approach, minimum distance, minimum χ^2, or Generalized Method of Moments (see Chapter 10 for a discussion). This approach also has a lot in common with simultaneous equation models (see Chapter 9). Chamberlain has published two useful references.[11]

[11] G. Chamberlain, "Multivariate Regression Models for Panel Data," *Journal of Econometrics,* **18,** 1982, 5–46; and "Panel Data," in *Handbook of Econometrics,* Volume 2, edited by Z. Griliches and M. Intriligator, North-Holland, 1984, 1247–1318.

We do not intend to treat this subject exhaustively. It is discussed here for two reasons:

1. Reporting an omnibus specification test for the fixed effects model has grown increasingly common. It cannot be understood without some familiarity with this basic approach.
2. Because much of the literature is technically demanding, the authors hope that a brief introduction will make the more complete treatments a bit more accessible to the researcher or student.

Chamberlain's insight is that the simple fixed effect model is really a large collection of restrictions on a more general model. In fact, the simplest way to understand this approach is to view panel data estimation as the estimation of a *set* of equations, much like the simultaneous equation models we considered earlier.

To keep the exposition clear, we will focus on the simplest case: one binary independent variable and two time periods. To fix ideas, consider the case when the dependent variable is log wages, and the right-hand-side variable is union status. Jakubson presents a particularly clear application of the basic methodology of this case.[12]

It is important to recognize that the fixed effects model actually embodies many restrictions. The basic model is

$$y_{it} = x_{it}\beta + \alpha_i + \epsilon_{it} \tag{12.47}$$

In the case of union wage effects, there are reasons to suspect that the omitted effect may be correlated with union status. From cross-section work, it is clear that for most workers the effect of union status is to raise wages relative to otherwise identical nonunion workers. Capitalists might respond to the inability to pay workers as little as the market will bear by trying to skim off the best workers; i.e., they may recruit more selectively. In this case, α_i may represent ability observed by the capitalist, but unobserved by the econometrician—a classic case for a correlated fixed effect.

Recall that OLS on this equation on a single cross section yields unbiased estimates of β_t if $\text{cov}(\alpha_i, x_{it}) = 0$. The problem comes when α_i is correlated with the x variable. In this case a simple expedient is the first-difference estimator introduced earlier:

$$\Delta y = \Delta x\beta + \Delta\epsilon \tag{12.48}$$

It is useful to rewrite this in an algebraically equivalent way:

$$\Delta y = \beta x_2 - \beta x_1 + \Delta\epsilon \tag{12.49}$$

Clearly this fixed effects model imposes a restriction. In Eq. (12.49), the effect of union status, entered in levels, is constrained to be opposite and equal in sign. In other words, people who are joining a union ($x_2 = 1, x_1 = 0$) should receive a premium equal and opposite in sign to someone who leaves a union ($x_2 = 0, x_1 = 1$). In this case, a standard F test can be used to test this additional restriction.

[12] G. Jakubson, "Estimation and Testing of Fixed Effects Models: Estimation of the Union Wage Effect Using Panel Data," *Review of Economic Studies*, **58**, 1991, 971–991.

12.11.1 Formalizing the Restrictions

Let us be a bit more formal. In doing so we will extend the previous example to the case where the union effect varies over time. Specifically, we analyze the following model:

$$y_{it} = x_{it}\beta_t + \alpha_i + \eta_{it} \tag{12.50}$$

We would like to formalize the notion that the fixed effect is correlated with x. Consider then the following population regression of the fixed effect on all the lags and leads of x:

$$\alpha_i = x_{i1}\lambda_1 + \cdots + x_{iT}\lambda_T + \xi_i \tag{12.51}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_T)'$ is a vector of coefficients, and where we will assume that ξ_i is an uncorrelated effect, much like the one we encountered in the random effects model—the person-specific component of variance that is uncorrelated with x. In the two-period case, $T = 2$. Notice that we include in this "linear projection" all the leads and lags of x.[13] We do this as a consequence of the fact that if the fixed effect is correlated with x in one period, it is likely to be correlated with x in another period.

We can now think of writing

$$\begin{aligned} y_{i1} &= \beta_1 x_{i1} + \lambda_1 x_{i1} + \lambda_2 x_{i2} + \xi_i + \eta_{i1} \\ y_{i2} &= \beta_2 x_{i2} + \lambda_1 x_{i1} + \lambda_2 x_{i2} + \xi_i + \eta_{i2} \\ y_{it} &= \beta_t' x_{it} + \lambda' x_i + \xi_i + \eta_{it} \end{aligned} \tag{12.52}$$

where $\boldsymbol{\lambda}$ is a vector of length two. It is straightforward to see that we can estimate a "reduced form" that involves two equations: y_1 as a function of x_1 and x_2, and y_2 as a function of the same two variables. This reduced form of Eq. (12.52) can be written as

$$\begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix} = \boldsymbol{\Pi} \begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix} \tag{12.53}$$

where $\boldsymbol{\Pi}$ is a 2×2 matrix of reduced form coefficients:

$$\boldsymbol{\Pi} = \begin{bmatrix} \Pi_1^1 & \Pi_2^1 \\ \Pi_1^2 & \Pi_2^2 \end{bmatrix}$$

The corresponding structural model can be summarized as

$$\boldsymbol{\Gamma} = \begin{bmatrix} \beta_1 + \lambda_1 & \lambda_1 \\ \lambda_2 & \beta_2 + \lambda_2 \end{bmatrix} \tag{12.54}$$

The **minimum distance approach** produces efficient estimates minimizing the distance between the structural coefficients, $\boldsymbol{\Gamma}$, and the reduced form coefficients

[13] We use the term *linear projection* here to avoid confusing Eq. (12.51) with the actual conditional expectation of α_i. That is, it is not important that Eq. (12.51) be the "true" model for α_i or that the true relationship be linear.

given by $\mathbf{\Pi}$. Formally, we need to choose $\mathbf{\Gamma}$ to minimize M, where

$$M = \text{vec}(\hat{\mathbf{\Pi}} - \mathbf{\Gamma})'[\text{var}(\text{vec}(\hat{\mathbf{\Pi}}))]^{-1}\,\text{vec}(\hat{\mathbf{\Pi}} - \mathbf{\Gamma}) \tag{12.55}$$

where the vec operator "vectorizes" a matrix and is defined as follows: Let A be a $n \times k$ matrix and let $\boldsymbol{a}_i$ be the ith column of A. Then $\text{vec}(A) = [\boldsymbol{a}_1'\boldsymbol{a}_2' \ldots \boldsymbol{a}_k']'$, a column vector of length nk. Notice that $\hat{\mathbf{\Pi}}$ is easily computed by performing GLS on the equation system given by (12.53). In this case, because each equation has the same exogenous variables, this computation amounts to running the two-equation system as seemingly unrelated regressions (SURs) or OLS. The alert student will also recognize this minimum distance approach as GMM (see Chapter 10) applied to this panel data case.

12.11.2 Fixed Effects in the General Model

If we cannot assume that ξ is uncorrelated with leads and lags of x, we merely first-difference the equation system or perform OLS on the following single equation:

$$(y_{i1} - y_{i2}) = \beta_1^{\text{FE}} x_{i1} - \beta_2^{\text{FE}} x_{i2} + (\eta_{i1} - \eta_{i2}) \tag{12.56}$$

In the case we are considering, Eq. (12.56) is exactly equal to what we get from estimating Eq. (12.53) by OLS and first-differencing. That is,

$$\hat{\Pi}_1^1 - \hat{\Pi}_1^2 = \hat{\beta}_1^{\text{FE}}$$

$$\hat{\Pi}_2^1 - \hat{\Pi}_2^2 = \hat{\beta}_2^{\text{FE}}$$

This will not be true in general, however. Instead, it will be necessary to perform a transformation (first differences, deviations from person-specific means, deviations from the last period's values, etc.). Note that in our example one can read off the estimates of λ readily. In the example we have been considering, we suggested that capitalists respond to their inability to pay low wages by more selective recruiting: we would expect such a situation to imply that λ_1 and λ_2 are both positive. Note well that $\lambda \gg 0$ also implies that the fixed effect estimate will be smaller than the cross-section estimate—the higher wages union members receive reflect not only, say, the profits the union has been able to transfer from the capitalist and shareholders to workers but also the higher productivity of workers recruited to work in unionized environments. With some caveats this is what Jakubson found, for example.[12]

12.11.3 Testing the Restrictions

So far we have considered the case when we have allowed the effect of the independent variables to vary. In this case, there are exactly four reduced form coefficients $(\Pi_1^1, \Pi_2^1, \Pi_1^2, \Pi_2^2)$ and four structural coefficients, $(\beta_1, \beta_2, \lambda_1, \lambda_2)$ and the minimand given by Eq. (12.55),

$$M = \text{vec}(\hat{\mathbf{\Pi}} - \mathbf{\Gamma})'[\text{var}(\text{vec}(\hat{\mathbf{\Pi}}))]^{-1}\,\text{vec}(\hat{\mathbf{\Pi}} - \mathbf{\Gamma})$$

will be identical to zero. There are no overidentifying restrictions to test.

Suppose we return to the case we considered earlier in this chapter—the coefficient is the same over time. This adds an additional restriction to the coefficient, and the system is overidentified even in the simple two-period case. As we learned in Chapter 10, this minimand will be asymptotically distributed as χ^2 with degrees of freedom given by the difference in the rank of vec($\mathbf{\Pi}$) and vec($\mathbf{\Gamma}$)—in this case one, because the coefficient is constrained to be equal over the two periods.

As it turns out, however, it is possible to test the overidentifying restrictions implied by this framework without having to estimate the underlying reduced form. In the model with more than two time periods, and the coefficients restricted to be the same over time, computation of the test statistic can be done by means of an auxiliary regression. As before, we use the within estimator and calculate the corresponding residuals $\hat{u}_w$. Next we construct the $nT \times kT$ matrix of all leads and lags of X, to wit,

$$\begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_T \end{bmatrix}$$

The restrictions implicit in Chamberlain's approach can now be tested by computing nTR^2, where R^2 is merely the conventional measure from the regression of the within residuals on the leads and lags of X.[14] It is straightforward to verify that this statistic will be distributed $\chi^2((T-1)k)$. Obviously, if one's estimate does not pass this "omnibus" test of overidentifying restrictions, it must be interpreted with caution. Although beyond the scope of the present discussion, it is possible to ease some of the restrictions in the fixed effects model. See Jakubson for a clear exposition of one approach.[12]

12.12 READINGS

A more complete treatment of panel data estimation issues can be found in Hsiao and the references therein.[15]

As regards Chamberlain's approach, the aforementioned articles are the standard references, although they are somewhat challenging to read.[11] Jakubson has written a clear and interesting application to the problem of union wage effects that extends this approach to the case where the binary independent variable, union status, is measured with error. Ashenfelter and Krueger have presented a clear exposition of Chamberlain's approach in an empirical example where instead of two time periods, there are two twins—the first twin is time period 1, and the second twin is time period 2.[16]

[14] The student may get some intuition for this by comparing this to Eq. (12.51).

[15] C. Hsiao, *Analysis of Panel Data*, 1986, Cambridge University Press.

[16] O. Ashenfelter and A. B. Krueger, "Estimates of the Return to Schooling from a New Sample of Twins," *American Economic Review*, **84** (5), 1994, 1157–1173.

Another interesting extension is by Lemieux, who treats observations of the same individual on two different jobs at the same time as constituting a panel.[17]

PROBLEMS

12.1. Consider the following simplified version of the wage curve,[18] that is, the relationship between contemporaneous real wages and aggregate unemployment rates using T independent cross-sections of N individuals. The wage curve is given by

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{12.57}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{i}_N \boldsymbol{x}'_1 \\ \vdots \\ \boldsymbol{i}_N \boldsymbol{x}'_T \end{bmatrix} \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_T \end{bmatrix}$$

where y_t, $\boldsymbol{i}_N$, and $\boldsymbol{\epsilon}_T$ are column vectors with N components; $\boldsymbol{i}_N$ is an N vector of ones; and $x'_t = [1 \;\; U_t]$ where U_t is the aggregate unemployment rate. Furthermore, assume that the error term is homoscedastic, equicorrelated across individuals, and uncorrelated across time periods where

$$E[\boldsymbol{\epsilon\epsilon}'] = \sigma^2 \boldsymbol{G}$$

and $\boldsymbol{G}$ is a block-diagonal matrix with T blocks of the form

$$\boldsymbol{G}_t = (1 - \rho)\boldsymbol{I}_N + \rho \boldsymbol{i}_N \boldsymbol{i}'_N$$

a. Show that the GLS and OLS estimates of β are the same.

b. Suppose that the special structure of the error term is ignored and the researcher merely computes the standard formula for the covariance matrix,

$$\boldsymbol{V}^*_{\text{OLS}} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Show that $\boldsymbol{V}^*_{\text{OLS}} = \sigma^2(1/N)(\sum x_t x'_t)^{-1}$.

c. Compute the correct covariance matrix and show that

$$\Sigma_{\text{OLS}} = \sigma^2(\theta/N)\left(\sum x_t x'_t\right)^{-1}$$

where $\theta = 1 + \rho(N - 1)$.[19]

12.2. Given the setup in the previous problem, consider the following two-step estimator of β and its associated standard error. In Step 1, the following model is estimated:

$$y = \boldsymbol{D\delta} + \epsilon$$

[17] T. Lemieux, "Estimating the Effect of Unions on Wage Inequality in a Two Sector Model with Comparative Advantage and Nonrandom Selection," 1993, Working Paper 9303, Département de sciences économiques, Université de Montréal.

[18] D. G. Blanchflower and A. J. Oswald, *The Wage Curve*, MIT Press, 1994.

[19] The use of explanatory variables that take on identical values within some groups can often lead to problems with standard errors even in cross-section data. For discussion of this problem see T. Kloek, "OLS Estimation in a Model Where a Microvariable is Explained by Aggregates and Contemporaneous Disturbances are Equicorrelated," *Econometrica* **49** (1), 1981, 205–208, and B. Moulton, "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics*, **32**, 1986, 385–397.

where $$D = [D_1 \quad D_2 \quad \cdots \quad D_T]$$

and D_t is a dummy variable that equals 1 when the observation comes from time period t. In Step 2, the coefficients are regressed on a constant and the associated x_t variables.

a. Show that the coefficient on x_t is a consistent estimator of $\boldsymbol{\beta}$.

b. If N is reasonably large, and the model is correct, approximately what would be the value of the R^2 from the regression in Step 2?

12.3. Consider the panel data case when

$$\nu_{it} = \mu_i + \lambda_t + \epsilon_{it} \tag{12.58}$$

where $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The μ_i, λ_t, and ϵ_{it} are random variables having zero means, independent among themselves and with each other, with variances σ^2_μ, σ^2_λ, and σ^2_ϵ, respectively. Show that

$$\boldsymbol{V} = E(\boldsymbol{\nu}\boldsymbol{\nu}') = \sigma^2[\rho \boldsymbol{A} + w\boldsymbol{B} + (1 - \rho - w)\boldsymbol{I}_{NT}]$$

where $$\boldsymbol{A} = \boldsymbol{I}_N \otimes \boldsymbol{J}_T$$
$$\boldsymbol{B} = \boldsymbol{J}_N \otimes \boldsymbol{I}_T$$

and $$\sigma^2 = \sigma^2_\mu + \sigma^2_\lambda + \sigma^2_\epsilon$$

and $$\rho = \sigma^2_\mu/\sigma^2 \qquad w = \sigma^2_\lambda/\sigma^2$$

and $\boldsymbol{J}_T$ is a $T \times T$ matrix of 1s.

12.4. For the disturbance term ν_{it} defined in Problem 12.3, indicate possible estimators of σ^2_μ, σ^2_λ, and σ^2_ϵ.

12.5. Suppose the true model is

$$y_{i,t} = \alpha_i^* + \beta x_{i,t}^* + \epsilon_{i,t}$$

where (for all i, t, and s)

$$\epsilon_{i,t} \sim \text{iid}(0, \sigma^2_\epsilon)$$
$$\text{cov}(x^*_{i,t}, \epsilon_{i,s}) = \text{cov}(\alpha^*_i, \epsilon_{i,t}) = 0$$
$$\text{cov}(x^*_{i,t}, \alpha^*_i) = \sigma_{x^*\alpha} \neq 0$$

and $$x_{i,t} = x^*_{i,t} + u_{i,t}$$

where (for all i, t, and s)

$$u_{i,t} \sim \text{iid}(0, \sigma^2_u)$$
$$x^*_{i,t} \sim \text{iid}(0, \sigma^2_{x^*})$$
$$\text{cov}(x^*_{i,t}, u_{i,s}) = \text{cov}(\alpha^*_i, u_{i,t}) = \text{cov}(\epsilon_{i,t}, u_{i,s}) = 0$$

The term α_i^* is the time-invariant individual specific characteristic unobserved by the econometrician, and x^* is measured with error by x. Check the consistency of the following estimators:

a. OLS of y on x

b. The fixed effect estimator

c. The between estimator

d. The random effects estimator

12.6. Derive θ for the random effects model in Eq. (12.12).

12.7. Consider the panel data model,

$$y_{i,t} = \boldsymbol{X}_{i,t}\boldsymbol{\beta} + \epsilon_{i,t}$$

and make the following assumptions:

$$\begin{aligned} E(\epsilon_{i,t}^2) &= \sigma_{ii} \\ E(\epsilon_{i,t}\epsilon_{j,t}) &= \sigma_{ij} \\ \epsilon_{i,t} &= \rho_i\epsilon_{i,t-1} + \nu_{i,t} \\ \nu_{i,t} &= \text{iid}(0, \sigma_\nu^2) \qquad |\rho_i| < 1 \end{aligned}$$

Derive the var($\boldsymbol{\epsilon}$) and discuss how a feasible GLS estimator of the parameters might be constructed.

CHAPTER 13

Discrete and Limited Dependent Variable Models

In this chapter we develop several different statistical models to handle situations for which OLS and 2SLS are generally not appropriate. Although many of the lessons we learned from our extensive analysis of OLS models apply here as well, others do not. Furthermore, the models we deal with in this chapter are generally nonlinear models (i.e., nonlinear in the parameters); so that, unlike OLS, they frequently do not maintain their desirable asymptotic properties when the errors are heteroscedastic, or nonnormal. Thus, the models appear to be less robust to misspecification in general.

With the advent of cheap computing and large microdata sets, applied use of these models has burgeoned. We will restrict our attention to cross-section applications (their most frequent use) although several of the models discussed here have also been analyzed for the time series or panel data context. The texts by Amemiya and Maddala are useful points of departure for these and other more complicated models.[1]

13.1 TYPES OF DISCRETE CHOICE MODELS

Discrete choice models attempt to explain a discrete choice or outcome. There are at least three basic types of discrete variables, and each generally requires a different statistical model.

Dichotomous, binary, or dummy variables. These take on a value of one or zero depending on which of two possible results occur. The reader has already encoun-

[1] T. Amemiya, *Advanced Econometrics,* Harvard University Press, 1985; and G. S. Maddala, ***Limited*** *Dependent and Qualitative Variables in Econometrics,* Cambridge University Press, 1983.

tered these types of variables in previous chapters. In this chapter we will deal with the case when such a variable is on the left-hand side of the relationship, i.e., when the dummy variable is an *endogenous* or dependent variable. Unlike the case when the dummy variable is exogenous, the endogenous dummy variable poses special problems that we have not yet addressed.

To take one example from labor economics, we may be interested in a person's decision to take a paying job in some reference period, say, a week. We can then define a dummy variable y as follows:

$$y_i = \begin{cases} 1 & \text{if person } i \text{ is employed in a paying job this week} \\ 0 & \text{otherwise} \end{cases}$$

Other examples from labor economics include the decision to go to college or not, or the decision to join a union or not.

Dummy variables are among the most frequently encountered discrete variables in applied work, and we will analyze these types of models in detail. An example, taken from the biometrics literature (where models for endogenous dummy variables were pioneered), is the case of evaluating an insecticide. We can imagine that tolerance y_i^* of an insect i to the insecticide is normally distributed across insects, say, $y_i^* \sim N(\mu, \sigma_\epsilon^2)$. If an insect's tolerance is less than the dose x_i of the insecticide, the insect dies.

The problem is that we cannot *observe* the tolerance y_i^* of a particular insect; instead we only observe whether the insect lives or dies. That is, we observe y_i, such that

$$y_i = \begin{cases} 1 & \text{if the insect dies} \\ 0 & \text{otherwise} \end{cases}$$

Given this setup, we can now turn to the question of interest: what is the probability that insect i dies? It is merely the probability that the insect's tolerance is less than the dose:

$$\text{prob}(y_i = 1) = \text{prob}(y_i^* < x_i) \tag{13.1}$$

In this formulation, what we observe, y_i, is generated by the following rule:

$$y_i = \begin{cases} 1 & \text{if } y^* < x_i \\ 0 & \text{otherwise} \end{cases}$$

In this example y^* is called a *latent* or *index variable*. It is called a latent variable because it is unobserved, unlike y, which we actually observe. This latent variable formulation is often analytically convenient.

Polychotomous variables. These take on a discrete number, greater than two, of possible values. Several types of polychotomous variables are encountered in applied econometric research.

1. *Unordered variables.* These are variables for which there is no natural ranking of the alternatives. For example, in transportation economics it may be of interest

to predict the mode of travel that a person chooses to get to and from his or her place of employment. The mode may be a function of several things such as the price of a subway trip or the price of gasoline. For a sample of commuters, we might want to define a variable as follows:

$$y_i = \begin{cases} 1 & \text{if person } i \text{ drives to work alone} \\ 2 & \text{if } i \text{ carpools with at least one other individual} \\ 3 & \text{if } i \text{ takes the subway} \\ 4 & \text{if } i \text{ takes the bus} \\ 5 & \text{if } i \text{ walks or uses some other method} \end{cases}$$

2. *Ordered variables.* With these variables the outcomes have a natural ranking. For instance, suppose we have a sample of medical diagnoses about the overall health of individuals. Further suppose that each person in the sample is given a diagnosis of either poor, fair, or excellent health. We might then define a variable as follows:

$$y_i = \begin{cases} 1 & \text{if person } i \text{ is in poor health} \\ 2 & \text{if person } i \text{ is in fair health} \\ 3 & \text{if person } i \text{ is in excellent health} \end{cases}$$

In contrast to the previous example on transportation mode choice, here there is a natural ordering of the variables. Excellent health is clearly "more healthy" than fair health, which in turn is healthier than poor health.

A special case of an ordered variable is a *sequential variable.* This occurs when, say, the second event is dependent on the first event, the third event is dependent on the previous two events, and so on. An example might be highest educational degree attained, where

$$y_i = \begin{cases} 1 & \text{high school diploma} \\ 2 & \text{some college} \\ 3 & \text{college degree} \\ 4 & \text{advanced degree} \end{cases}$$

Count data models. In these models the dependent variable takes on integer values. A leading example is the number of patents in a given year. In this case we might have a variable y that takes on values $y = 1, 2, \ldots,$ etc. These types of models are less common in applied econometric research and are often handled by traditional linear methods.

In the next several sections we consider binary dependent variable models, where the two values are typically denoted by 0 and 1.

13.2 THE LINEAR PROBABILITY MODEL

Consider the following example where

$$y_i = \begin{cases} 1 & \text{if person } i \text{ is a union member} \\ 0 & \text{otherwise} \end{cases}$$

If we are interested in whether collective bargaining raises the wages of workers, we might consider running the following OLS regression:

$$\ln w_i = \alpha_0 + \boldsymbol{X}_i\boldsymbol{\gamma} + y_i\delta + \epsilon_i \tag{13.2}$$

where $\ln w$ refers to log weekly wages and X is a vector of demographic characteristics thought to affect wages. If we assume that all the right-hand side variables are exogenous and ϵ is normally distributed with mean zero, the binary character of y poses no special problem for estimation.

On the other hand, consider the following descriptive regression:

$$y_i = \boldsymbol{Z}_i\boldsymbol{\beta} + \nu_i \tag{13.3}$$

where $\boldsymbol{Z}_i$ is some set of characteristics thought to determine a person's propensity to join a union, and $\boldsymbol{\beta}$ is a set of parameters. In this context, the fact that y is binary does cause some problems, particularly in interpretation.

Consider estimating Eq. (13.3) by OLS, called a *linear probability* (LP) *model*. When the dependent variable is continuous, it is convenient to interpret the regression as specifying $E[y \mid Z]$, the expectation of y given a set of Zs. Although this approach is generally appropriate when y is continuous, it is generally not so when y is binary, as the next example will illustrate.

13.3 EXAMPLE: A SIMPLE DESCRIPTIVE MODEL OF UNION PARTICIPATION

To illustrate, we use our 1988 Current Population Survey data and estimate a simple descriptive linear probability model for the likelihood of being a union member. The model we will specify is the following:

$$\begin{aligned}\text{Union} = {} & \beta_0 + \beta_1(\text{potential experience}) + \beta_2(\text{experience})^2 \\ & + \beta_3(\text{grade}) + \beta_4(\text{married}) + \beta_5(\text{high}) + \epsilon\end{aligned}$$

where Potential experience = age − years of schooling − 5, which for men is often a reasonable approximation of the number of years they have been in the labor force

Grade = number of years of schooling completed

Married = a dummy variable that equals 1 if the worker is married and 0 otherwise

High = a dummy variable that equals 1 if the worker is in a "highly" unionized industry (Natural Resources, Manufacturing, Construction, Education, Health and Welfare, Transportation, or Public Administration) and 0 otherwise

First, it is useful to consider the differences in the means of the two samples. In Fig. 13.1, we present some simple summary statistics. It is evident that in this sample, a union worker is older on average (because this is a single cross section this difference in means may reflect either an age or a cohort effect), has fewer years of school, is more likely to be married, and (not surprisingly) is more likely to work

```
→ union = 0
Variable   Obs       Mean    Std. Dev.   Min    Max

  potexp   784   17.81122    12.93831      1     55
    exp2   784    484.426    595.7631      1   3025
   grade   784   13.13903    2.676191      0     18
 married   784   .6109694    .4878415      0      1
    high   784   .5140306    .5001222      0      1


→ union = 1
Variable   Obs       Mean    Std. Dev.   Min    Max

  potexp   216   22.77778    11.41711      1     49
    exp2   216   648.5741    563.3426      1   2401
   grade   216   12.56019    2.273423      5     18
 married   216        .75    .4340185      0      1
    high   216   .7638889    .4256778      0      1
```

FIGURE 13.1
Union vs nonunion workers in 1988 Current Population Survey.

```
Source          SS        df       MS          Number of obs  =   1000
                                               F(5,994)       =  18.17
Model      14.1787575      5   2.8357515       Prob > F       = 0.0000
Residual   155.165242    994   .156101854      R-square       = 0.0837
                                               Adj R-square   = 0.0791
Total         169.344    999   .169513514      Root MSE       =  .3951

Variable   Coefficient   Std. Error        t      Prob > |t|     Mean

union                                                            .216

potexp       .0200388     .0038969     5.142        0.000      18.884
exp2        -.0003706     .0000819    -4.526        0.000     519.882
grade       -.0124636     .0051005    -2.444        0.015      13.014
married      .0133428      .030001     0.445        0.657        .641
high         .1439396     .0256785     5.605        0.000        .568
_cons        .1021368     .0749337     1.363        0.173           1
```

FIGURE 13.2
A simple linear probability model for union status.

in a highly unionized industry. About 22 percent of the sample are union members. It also appears that as a group, union workers are more homogenous since the standard deviations of each factor are smaller.

Figure 13.2 presents (slightly modified) STATA output from the simple linear probability model. The output is rather conventional. The ANOVA decomposition of the sum of squares in the dependent variable (Total) into explained (Model) and unexplained (Residual) is given, along with the R^2, the adjusted R^2, and the Root MSE (Root Mean Squared Error). The value of the F statistic and its probability refer to the null hypothesis that all of the slope coefficients are zero. Not surprisingly, this hypothesis is thoroughly rejected.

Interpretation of the coefficients is straightforward. The coefficients are merely the derivatives of the probability of union membership with respect to the element of Z (this assertion is not quite true for the two dummy variables, however, because there is no simple derivative of discrete variables). Not too surprisingly, the signs on the various coefficients correspond to the difference in means between union and nonunion members that we saw in Fig. 13.1. For example, the results of the LP model suggest that an additional year of school lowers the probability of being a union member by slightly more than 1 percent. It is also clear that workers with more experience are generally more likely to be in unions, although the effect of age has diminishing marginal returns since the coefficient on experience squared is negative.[2]

The problem with the linear probability model is highlighted in Fig. 13.3, which displays a simple histogram of the predicted values calculated from the model in Fig. 13.2. The problem is that about 5 percent of the predicted values are less than zero! These values clearly do not make sense. Taken literally, the coefficients imply that some members have a -10 percent chance of being a union member.

A major weakness of the linear probability model is that it does not constrain the predicted value to lie between 0 and 1. Because the derivative of the probability with respect to $\boldsymbol{X}$ is merely $\boldsymbol{\beta}$, nothing constrains the predicted value. Imagine the case of one explanatory variable with a positive coefficient. In such a case, there is always some value of the explanatory variable we could choose that would push the

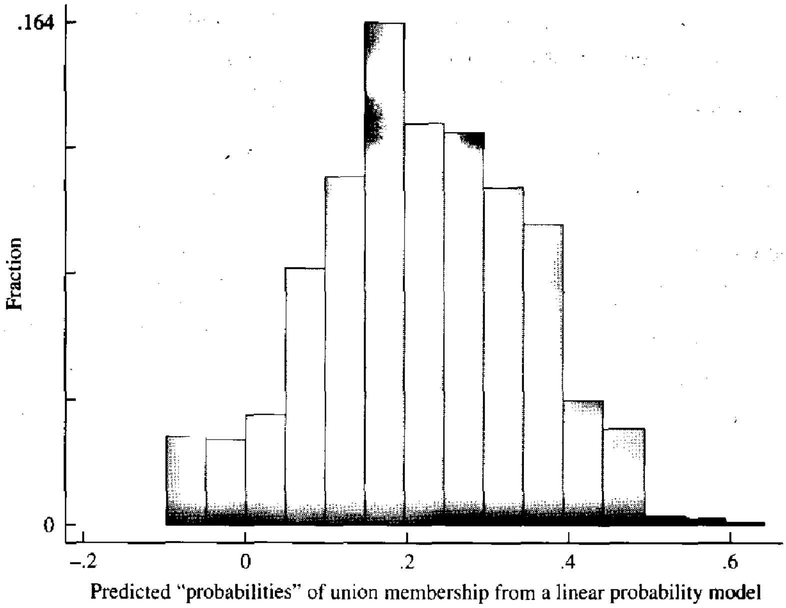


FIGURE 13.3
Histogram of predicted values for union status from a linear probability model.

[2] The value of experience that maximizes the probability of being in a union is (.02/.00037)/2 = 54.

probability above 1 (unless, of course, the range of X itself is limited, as in the case when it is a dummy variable.)

In addition, *the linear probability model is heteroscedastic.* This property of the linear probability model is easy to verify since the residual can take on only one of two values, $1 - X_i\boldsymbol{\beta}$ or $-X_i\boldsymbol{\beta}$, since y can take on one of two values, 1 or 0. The variance of ϵ for a particular observation is then

$$\operatorname{var}(\epsilon_i \mid X_i) = X_i\boldsymbol{\beta}(1 - X_i\boldsymbol{\beta}) \tag{13.4}$$

It is clear that the variance, instead of being constant, varies with the size of $\boldsymbol{X\beta}$. As a consequence, heteroscedasticity-consistent standard errors calculated by White's procedure (see Chapter 6) can be used. It is also possible to incorporate the heteroscedastic aspect of the model directly into the estimation by use of weighted least squares.

Because the linear probability model allows for predicted values outside the (0, 1) range, the model has fallen out of favor for many applications. As we will discuss briefly later, it has some advantages (primarily its simplicity) that have resulted in its continued use.

13.4 FORMULATING A PROBABILITY MODEL

Because the linear probability model has intrinsic defects for some applications, let us consider alternatives. A useful way to think about the problem is to recognize that we would like to transform $\boldsymbol{X\beta}$ into a *probability.*

That is, we need a function F such that:

$$\operatorname{prob}(y_i = 1) = F(X_i\boldsymbol{\beta}) \tag{13.5}$$

A natural choice of a function F that translates $\boldsymbol{X\beta}$ into a number between 0 and 1 in a sensible way is a distribution function, or the cumulative density. In fact, binary response models can be defined this way.

If we choose F to be the identity function, so that

$$\operatorname{prob}(y_i = 1) = X_i\boldsymbol{\beta} \tag{13.6}$$

we get the linear probability model already discussed. It is clear by inspection that such a choice for F does not yield the type of function we want, for nothing constrains $\boldsymbol{X\beta}$ to lie between 0 and 1.

Choosing F to be standard normal yields one attractive possibility, the *probit* model:

$$\operatorname{prob}(y_i = 1) = \Phi(X_i\boldsymbol{\beta}) = \int_{-\infty}^{X_i\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz \tag{13.7}$$

The standard normal transformation $\Phi(\cdot)$ constrains the probability to lie between 0 and 1, or

$$\lim_{z \to +\infty} \Phi(z) = 1 \qquad \text{and} \qquad \lim_{z \to -\infty} \Phi(z) = 0$$

Choosing F to be the logistic distribution yields another attractive possibility, the *logit* model:

$$\text{prob}(y_i = 1) = \Lambda(X_i\boldsymbol{\beta}) = \frac{\exp X_i\boldsymbol{\beta}}{1 + \exp X_i\boldsymbol{\beta}} \tag{13.8}$$

This choice of F also returns a value between 0 and 1.

We are actually not limited to these two choices. *Any* function with the right property would work, although the probit and logit are the most common models in practice. It is instructive to consider the probit and logit models in somewhat more detail.

13.5 THE PROBIT

So far we have presented the probit and logit models as convenient functional forms for models with binary endogenous variables. Both models also have a "behavioral" interpretation that is instructive and often analytically convenient. We consider the probit first. We observe some variable $\boldsymbol{y}$ that takes on one of two values, 0 and 1. Define a latent variable $\boldsymbol{y}^*$ such that

$$y_i^* = X_i\boldsymbol{\beta} + \epsilon_i \tag{13.9}$$

We do not observe $\boldsymbol{y}^*$, but rather $\boldsymbol{y}$, which takes on values of 0 or 1 according to the following rule:

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{13.10}$$

We also assume that $\epsilon_i \sim N(0, \sigma^2)$.

Remember that in contrast with the linear probability model, y_i^* (conditional on X) is distributed normally in the probit model, although its realization y_i is not. It is straightforward to show that the rule expressed in Eq. (13.10) generates a probit. First note that

$$\begin{aligned} \text{prob}(y_i = 1) &= \text{prob}(y_i^* > 0) \\ &= \text{prob}(X_i\boldsymbol{\beta} + \epsilon_i > 0) \\ &= \text{prob}(\epsilon_i > -X_i\boldsymbol{\beta}) \\ &= \text{prob}\left(\frac{\epsilon_i}{\sigma} > -X_i\frac{\boldsymbol{\beta}}{\sigma}\right) \end{aligned} \tag{13.11}$$

where σ^2 is the variance of ϵ. Dividing by σ in Eq. (13.11) is helpful because the quantity ϵ/σ is distributed as *standard* normal—mean zero and unit variance. The quantity ϵ/σ is *standard* normal because ϵ has been transformed by subtracting its mean, zero, and then dividing by its standard deviation, σ.

For the probit model (and the logit we describe shortly) the distribution is symmetric, so that Eq. (13.11) can be written as

$$\begin{aligned}
\text{prob}(y_i = 1) &= \text{prob}\left(\frac{\epsilon_i}{\sigma} > -X_i\frac{\boldsymbol{\beta}}{\sigma}\right) \\
&= \text{prob}\left(\frac{\epsilon_i}{\sigma} < X_i\frac{\boldsymbol{\beta}}{\sigma}\right) \\
&= \Phi\left(X_i\frac{\boldsymbol{\beta}}{\sigma}\right) \qquad (13.12)
\end{aligned}$$

Deriving the likelihood function is straightforward. Because

$$\text{prob}(y_i = 1) = \Phi\left(X_i\frac{\boldsymbol{\beta}}{\sigma}\right)$$

it follows that

$$\text{prob}(y_i = 0) = 1 - \text{prob}(y_i = 1) = 1 - \Phi\left(X_i\frac{\boldsymbol{\beta}}{\sigma}\right)$$

If we have iid sampling, the likelihood for the sample is the product of the probability of each observation. Denoting $1, \ldots, m$ as the m observations such that $y_i = 0$, and $m + 1, \ldots, n$ as the $n - m$ observations such that $y_i = 1$, yields

$$\begin{aligned}
L &= \text{prob}(y_1 = 0) \cdot \text{prob}(y_2 = 0) \cdots \text{prob}(y_m = 0) \\
&\quad \cdot \text{prob}(y_{m+1} = 1) \cdots \text{prob}(y_n = 1) \qquad (13.13) \\
&= \prod_{i=1}^{m}\left[1 - \Phi\left(X_i\frac{\boldsymbol{\beta}}{\sigma}\right)\right] \prod_{i=m+1}^{n} \Phi\left(X_i\frac{\boldsymbol{\beta}}{\sigma}\right) \qquad (13.14) \\
&= \prod_{i=1}^{n} \Phi\left(X_i\frac{\boldsymbol{\beta}}{\sigma}\right)^{y_i}\left[1 - \Phi\left(X_i\frac{\boldsymbol{\beta}}{\sigma}\right)\right]^{1-y_i} \qquad (13.15)
\end{aligned}$$

Typically, one works with the log-likelihood function, which is

$$\begin{aligned}
l\left(\frac{\boldsymbol{\beta}}{\sigma}\right) &= \ln(L) \qquad (13.16) \\
&= \sum_i \left\{ y_i \cdot \ln\left[\Phi\left(X_i\frac{\boldsymbol{\beta}}{\sigma}\right)\right] + (1 - y_i) \cdot \ln\left[1 - \Phi\left(X_i\frac{\boldsymbol{\beta}}{\sigma}\right)\right]\right\} \qquad (13.17)
\end{aligned}$$

Notice that the log-likelihood is bounded above by 0, because $0 \leq \Phi(\cdot) \leq 1$ implies that

$$\ln[\Phi(\cdot)] \leq 0 \qquad \text{and} \qquad \ln[1 - \Phi(\cdot)] \leq 0$$

Another important aspect of the likelihood function is that the parameters $\boldsymbol{\beta}$ and σ always appear together. Therefore, they are not separately identified: only the ratio $\boldsymbol{\beta}/\sigma$ matters. It is thus convenient to normalize σ to be one, so we can just talk about $\boldsymbol{\beta}$. (The case when σ is heteroscedastic will be discussed shortly.)

Estimating the probit is straightforward even though the model is nonlinear and no closed-form expression for $\Phi(\cdot)$ exists. $\Phi(\cdot)$, therefore, has to be evaluated numerically. One feature of the probit (and the logit) is that the likelihood functions are *globally concave.* Therefore, an optimization package does not have to worry about discriminating between local maxima and global maxima when it tries to find parameter values that maximize the log-likelihood function—they will be the same.

A standard procedure is to calculate estimates from a linear probability model and to use these as an initial "guess" with which to begin finding a solution. As each guess gets better and better, the value of the log-likelihood function rises at each step until no improvement is possible, and the solution is found. One method is the so-called method of scoring. In the method of scoring, the probit estimates are found in steps:

$$\boldsymbol{\beta}_{f+1} = \boldsymbol{\beta}_f + \hat{I}^{-1}(\boldsymbol{\beta}_f)\frac{\partial l}{\partial \boldsymbol{\beta}_f} \tag{13.18}$$

where the subscripts refer to the iteration toward finding a solution. $\hat{I}(\boldsymbol{\beta}_f)$ is an estimate of the information matrix (a square, symmetric matrix of the negative of the second-order derivatives of the log-likelihood function, or the outer product of the gradient) evaluated at the last guess. When the difference between $\boldsymbol{\beta}_{f+1}$ and $\boldsymbol{\beta}_f$ is close enough to zero, the process stops. The change in the value of the coefficients from two successive iterations will be close to zero when the *score,* $\partial l/\partial \boldsymbol{\beta}$ (the derivative of the log-likelihood function with respect to the parameters), is close to zero. (See Chapter 5 for a discussion of the score and information matrix.)[3]

```
. probit union pot exp2 grade married high

Iteration 0: Log Likelihood = -521.79487
Iteration 1: Log Likelihood = -476.40231
Iteration 2: Log Likelihood = -475.2548
Iteration 3: Log Likelihood = -475.2514

Probit Estimates                          Number of obs=  1000
                                          chi2(5)       = 93.09
Log Likelihood = -475.2514                Prob > chi2   =0.0000

Variable   Coefficient   Std. Error      t      Prob > |t|     Mean
--------------------------------------------------------------------
union                                                          .216

potexp      .0835091     .0156088     5.350      0.000       18.884
exp2       -.0015308     .0003179    -4.816      0.000      519.882
grade      -.042078      .018909     -2.225      0.026       13.014
married     .0622516     .112584      0.553      0.580         .641
high        .5612953     .0996624     5.632      0.000         .568
_cons      -1.468412     .2958126    -4.964      0.000          1
--------------------------------------------------------------------
```

FIGURE 13.4
A probit model.

[3] For an introduction to numerical optimization methods see W. Greene, *Econometric Analysis,* Chapter 12, Macmillan, 1993, or R. E. Quandt, "Computational problems and methods," in Z. Griliches and M. D. Intriligator, *Handbook of Econometrics,* Chapter 12, North-Holland, 1983.

Let us consider as an empirical illustration a model of union membership using the same data and same explanatory variables as before, except this time we estimate a probit instead of the simple linear probability model. The STATA output for the probit is in Fig. 13.4.

Unlike the linear probability model, in which no iteration is required, the output for the probit includes the value of the log-likelihood function as the program iterated to a solution. Note that the value of the likelihood increased with each guess. In this context there is no precise analog to the R^2. However, the test statistic distributed as $\chi^2(5)$ (and denoted chi2 in the output) is analogous to the usual F test. It is a test against the null hypothesis that the slope coefficients are all equal to zero.

This test is the likelihood ratio test:

$$2\,[l(\alpha, \boldsymbol{\beta}) - l(\alpha, 0)] \overset{a}{\sim} \chi^2(k-1) \tag{13.19}$$

where $L(\alpha, \boldsymbol{\beta})$ is the maximized value of the log-likelihood of the model being estimated, $L(\alpha, 0)$ is the value of the log-likelihood for a probit with only a constant term, and $k - 1$ is the number of slope coefficients. This expression is a special case of the likelihood ratio test discussed in Chapter 5. As we saw before in using the linear probability model, the null hypothesis is clearly rejected.

Observe that the sign pattern of the coefficients is the same one we observed for the linear probability model. However, calculating the change in the probability of union membership with respect to one of the right-hand-side variables is not so simple as it was in the linear probability model. In the probit, the derivative of the probability with respect to a specific X_k in the set of variables $\boldsymbol{X}$ is

$$\frac{\partial E(y)}{\partial X_k} = \phi(\boldsymbol{X\beta})\beta_k \tag{13.20}$$

where $$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

is the standard normal density. Compare this to the derivative of the probability with respect to $\boldsymbol{X}$ in the linear probability model, $\boldsymbol{\beta}$, or merely the coefficients of the model. *In the probit model, the derivative of the probability with respect to* $\boldsymbol{X}$ *varies with the level of* $\boldsymbol{X}$ *and the other variables in the model.*

A practical consequence of this difference is that it is not generally useful merely to report the coefficients from a probit (as it is for a linear probability model) unless only the sign and significance of the coefficients are of interest. To see this clearly, consider the effect of industrial affiliation as estimated from the foregoing probit. Figure 13.5 was generated by first taking the sample of workers in low-union industries, and then plotting the calculated probabilities against themselves (hence the lower line has a slope of 1). Next the probability is recalculated for these workers, now assuming that they are in a high-union industry, so that the new predicted probability becomes $\Phi(\boldsymbol{X}\hat{\boldsymbol{\beta}} + \beta_{\text{high}})$. This too is plotted against the same axis.

For all these workers, the effect of a change in industrial affiliation is positive. All else being equal, being in a high-union industry raises the chance that a person will be a union worker. However, as is clear, this "industry" effect is smaller when a person's other $\boldsymbol{X}$ characteristics suggest that the person is unlikely to join a

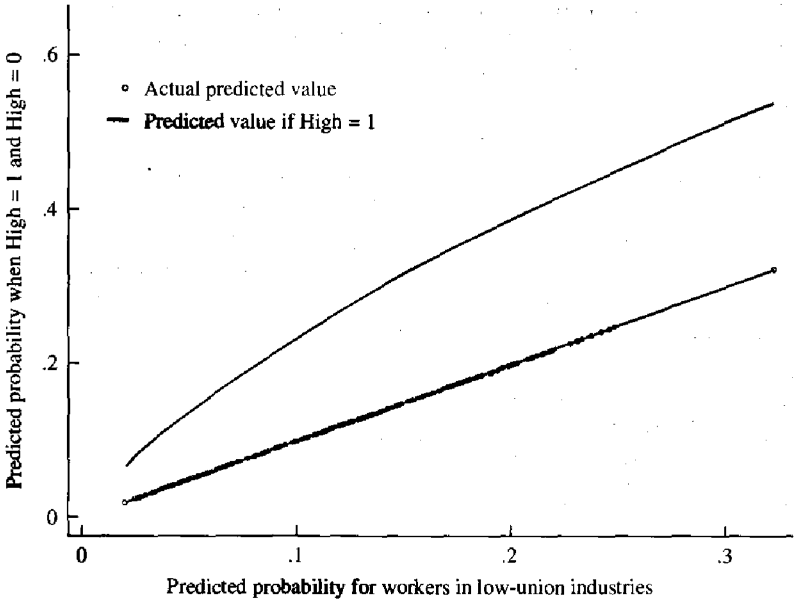


FIGURE 13.5
Effect of industrial affiliation on a sample of workers in low-union industries.

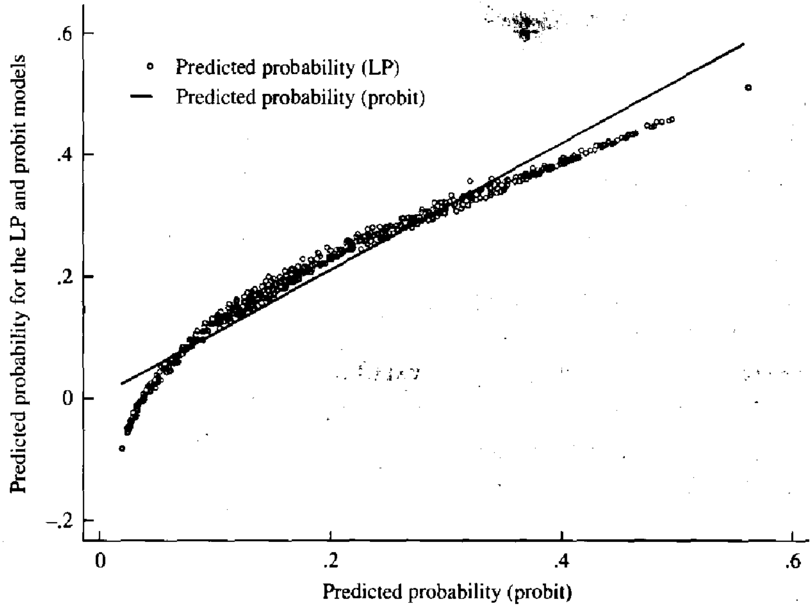


FIGURE 13.6
Probit vs the linear probability model.

union. (Over the relevant range depicted in the figure, the distance between the two lines increases as one moves away from the origin. This is one manifestation of the nonlinearity of the model.)

What this example illustrates is that reporting probit results generally requires more information than reporting coefficients from linear regressions. One useful expedient is to calculate the value of the derivatives at the mean values of all the $\boldsymbol{X}$ variables in the sample. (This is equivalent to computing the mean estimated index since the index is a linear function of the Xs). The motivation is to display the derivative for a "typical" element of the sample. We discuss the interpretation of coefficients in more detail shortly.

It is interesting to compare the probit results with those we obtained from the linear probability model. Figure 13.6 plots the predicted probabilities from the linear probability model against the predicted value for the probit. Most interestingly, the probit has taken the negative values from the LP model and moved them to values just above zero. Note also the differences between the predicted probabilities from the two models at the highest probabilities.

13.6 THE LOGIT

The development of the logit is identical to that of the probit. Recall from Eq. (13.8) that

$$\begin{aligned} \text{prob}(y_i = 1) &= \Lambda(\boldsymbol{X}_i\boldsymbol{\beta}) \\ &= \frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}_i\boldsymbol{\beta})} \end{aligned} \tag{13.21}$$

The latent variable interpretation of the logit proceeds exactly the same way as in the probit except that in Eq. (13.9) ϵ follows what is called an extreme value distribution.[4] Like the probit, and unlike the linear probability model, the formulation of the model ensures that the predicted probabilities lie between 0 and 1. The main difference between the normal distribution and the logistic distribution is that the latter has more weight in the tails.

The derivative of the probability with respect to one element of X varies with X as in the probit model:

$$\frac{\partial E(y)}{\partial X_k} = \frac{\exp(\boldsymbol{X\beta})}{(1 + \exp(\boldsymbol{X\beta}))^2}\beta_k$$

A convenient way to rewrite the derivative is

$$\frac{\partial E(y)}{\partial X_k} = p(1 - p)\beta_k \tag{13.22}$$

where

$$p = \frac{\exp(\boldsymbol{X\beta})}{1 + \exp(\boldsymbol{X\beta})}$$

[4] See D. McFadden, "Econometric Analysis of Qualitative Choice Models," Chapter 24, *Handbook of Econometrics*, eds. Z. Griliches and M. D. Intriligator, North-Holland, 1984, for a discussion.

```
. logit union pot exp2 grade married high

Iteration 0: Log Likelihood = −521.79847
Iteration 1: Log Likelihood = −478.21273
Iteration 2: Log Likelihood = −475.5873
Iteration 3: Log Likelihood = −475.55412
Iteration 4: Log Likelihood = −475.55411

Logit Estimates                          Number of obs =   1000
                                         chi2(5)       =  92.49
Log Likelihood = −475.55411              Prob > chi2   = 0.0000

Variable   Coefficient   Std. Error       t    Prob > |t|      Mean
--------------------------------------------------------------------
union                                                          .216

potexp        .1474021      .028097   5.246       0.000      18.884
exp2         −.0026869     .0005654  −4.752       0.000     519.882
grade        −.0703209      .032142  −2.188       0.029      13.014
married        .115463      .196779   0.587       0.557        .641
high          .9801411      .180049   5.444       0.000        .568
_cons        −2.581436     .5186859  −4.977       0.000           1
--------------------------------------------------------------------
```

FIGURE 13.7
A logit model of union membership.

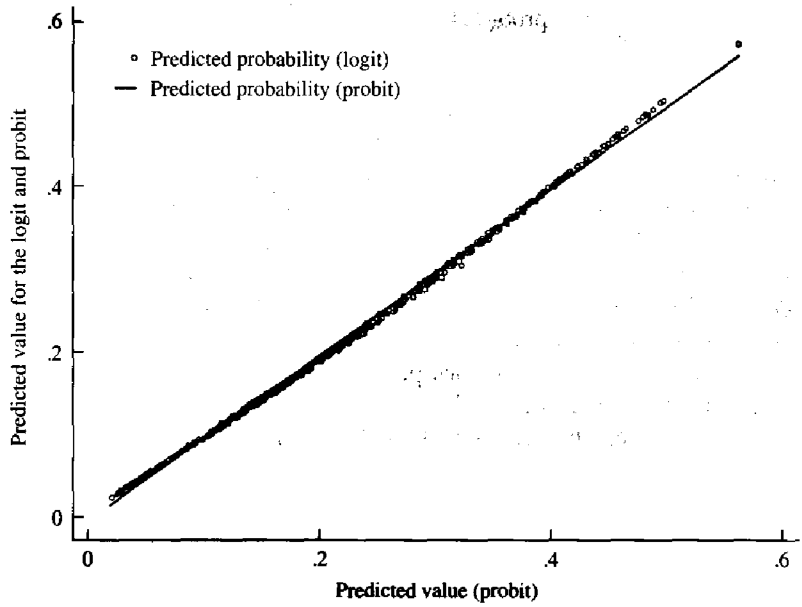


FIGURE 13.8
Logit vs probit.

Figure 13.7 presents the results from a logit specification. Again, the model is estimated by maximum likelihood methods, and the STATA output includes the values of the log-likelihood function as it iterates to its maximum. Also, the χ^2 test of this model against the null that the appropriate model contains only a constant rejects the null decisively. The sign pattern of the coefficients is the same. For example, schooling lowers the probability of union membership.

Figure 13.8 is a comparison of the predicted probabilities estimated by the probit with those of the logit. Clearly, the differences are rather minor.

13.7 MISSPECIFICATION IN BINARY DEPENDENT MODELS

13.7.1 Heteroscedasticity

As we have already discussed, the linear probability model is inherently heteroscedastic. Additional heteroscedasticity in a linear model is not a particular problem if the standard White procedure is used.

In the case of the probit and the logit, however, the problem requires some additional discussion. To make the discussion as general as possible, let us denote the regression function as $f(X)$ (we have usually assumed that this is linear, e.g., $f(X) = X\boldsymbol{\beta}$), and let $F(\cdot)$ be the appropriate cumulative distribution function (the cumulative normal and the cumulative logistic in the probit and logit, respectively):

$$\text{prob}(y = 1) = F\left[\frac{f(X)}{\sigma}\right] \tag{13.23}$$

Consider the problem of heteroscedasticity first. Recall that in discussing these two models, we *assumed* that σ was constant, so that normalizing it to 1 was harmless. If the assumption of constant variance is dropped so that $\sigma = \sigma_i$, it is easy to see why heteroscedasticity is a problem: when $f(X) = X\boldsymbol{\beta}$ and $\boldsymbol{\beta}$ is $k \times 1$, for example, we would be faced with a likelihood function of the form $L(\boldsymbol{\beta}/\sigma_i)$, which has $n + k$ parameters—$\sigma_1, \ldots, \sigma_n$, $\boldsymbol{\beta}$—and it is impossible to estimate this function without further restrictions. This outcome should be contrasted with the standard linear model in the presence of heteroscedasticity, where it is meaningful to think about the problem as estimating $\boldsymbol{\beta}$ and then "correcting" the standard errors. Of course, if the heteroscedasticity were of a *known* parametric form, we could incorporate this directly into our likelihood function.[5]

Since we are rarely blessed with knowledge of the precise form of the heteroscedasticity, this might seem to be a grave problem, for heteroscedasticity is common. The problem is more apparent than real, however.

Suppose that the heteroscedasticity is of the form $\sigma_i = \sigma g(X_i)$: then

$$\text{prob}(y_i = 1) = F\left[\frac{f(X_i)}{\sigma g(X_i)}\right] \tag{13.24}$$

[5] See Greene, op. cit., Chapter 21.4.1, especially Example 21.7.

The presence of heteroscedasticity causes inconsistency because the *assumption* of a constant σ_i clearly is what allows us to identify the regression function $f(X)$. To take a very particular but informative case, suppose that the regression function is linear $f(X) = X\boldsymbol{\beta}$, and that heteroscedasticity takes the form $g(X) = X\boldsymbol{\beta}/X\boldsymbol{\gamma}$.

In this case,

$$\text{prob}(y_i = 1) = F\left(\frac{X_i\boldsymbol{\gamma}}{\sigma}\right) \tag{13.25}$$

and it is clear that our estimates will be inconsistent for $\boldsymbol{\beta}$ (although consistent for $\boldsymbol{\gamma}$!). Because the empirical problem is to identify the effect of the covariates on the probability, it is not apparent why it should matter if the effect of the Xs is through the regression function $f(X)$ or through the "scedastic" function $g(X)$. That is, whether the Xs work through the means or the variances does not generally matter. Of course, if for some reason the function $f(\cdot)$ is the object of interest, then the problem remains.[6]

We do not want to leave the impression that misspecification is impossible because our concern is generally the *ratio* of $f(X)$ to $g(X)$. One unfortunate feature of nonlinear models in general is that the situation is never straightforward. For instance, although appending a white noise error causes no problems of bias in the linear model, it will in general cause a bias in the nonlinear models we discuss in this chapter.

13.7.2 Misspecification in the Probit and Logit

A detailed discussion of the consequences of misspecification in nonlinear models would take us too far afield, but a sketch of some of the problems involved may be useful.

The consequence of misspecification in models estimated by maximum likelihood would appear to be a straightforward matter. If we maximize the "correct" likelihood function, we get consistent estimates. If we maximize the wrong function, we get biased estimates. Although these statements are true in general, consider the familiar standard linear model:

$$y = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{13.26}$$

As discussed in Chapter 5, $\hat{\boldsymbol{\beta}}$ can be found by maximizing the likelihood function. As we also learned in Chapter 5 and in Chapter 10 (on GMM), the MLE (or OLS, which is equivalent in this case) is consistent even in the face of heteroscedas-

[6] One case where $f(\cdot)$ is the object of interest is a *random utility* model. In this class of models, our index y^* can be interpreted as the level of utility achieved by an individual. It is therefore a matter of some concern whether the effect of the independent variables is through the regression function, which therefore affects utility, or whether it works through the variance.

ticity, nonnormal errors, and serial correlation—i.e., as long as $\text{plim}(1/N)X'\epsilon = 0$. In other words, we have some latitude in choosing a likelihood function that produces consistent estimates of β. In fact, we typically compute the OLS coefficients for β that correspond to maximizing the likelihood, assuming perfectly spherical disturbances, and then later "correct" the standard errors.

An estimate obtained from maximizing an "incorrect" likelihood function is often referred to as a *quasi-maximum likelihood* (QML) estimate and is quite general. White has analyzed this case.[7] OLS in the linear model, as it turns out, is a bit of an exception because it represents a case where misspecification such as heteroscedasticity does not generate inconsistency in $\hat{\boldsymbol{\beta}}$. In some circumstances maximum likelihood for a misspecified model will produce consistent estimates of the parameters, although the standard errors will have to be modified to account for the misspecification.

This statement is not true in general, however. The difficulty with the probit or logit is that *any* misspecification of the likelihood will result in inconsistency. If the true model is a probit and we maximize the likelihood function associated with the logit, our estimates will be inconsistent. We can, however, continue to talk about the QML estimator and derive its standard errors. The salient question then becomes, is the QML estimate interesting?

Cases where the answer to this question is yes are easy to think of. Consider a case where the true model is given by

$$y = 0.01 + x + 0.00001x^2 + \epsilon \tag{13.27}$$

In this example suppose the range of x is small, say (0,1), and ϵ is standard normal. One can consider the result of maximizing the likelihood function while ignoring the quadratic term. Although it would be preferable to estimate the model with both x and x^2, it is clear that for many purposes our ML estimate of the coefficient x will not be "too bad" (note also that the errors will be heteroscedastic even though they are homoscedastic in the true model).

Returning to the probit and logit models, we see that *any* misspecification results in inconsistency in general; it appears, however, that in the vast majority of empirical cases the probit, logit, and linear probability models seem to produce similar answers. One way to reconcile the similarity of different estimates from what are almost certainly "incorrect" specifications is to consider each of them as QML estimates of some other true model. If that is the case, it seems advisable to take that into consideration.

Let the density for a particular observation t be given by $f(y_t, X_t, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ refers to the k parameters of the model. Following White, we write the log-likelihood function for a sample of size n as

$$l(y_t, X_t, \boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} \ln f(y_t, X_t, \boldsymbol{\theta}) \tag{13.28}$$

[7] H. White, "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, **50**, 1982, 1–16.

Define the following $k \times k$ matrix:

$$\boldsymbol{A}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2 \ln f(y_t, \boldsymbol{X}_t, \boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} \tag{13.29}$$

where $i = 1, \dots, k$ and $j = 1, \dots, k$. A_n is simply the average Hessian. Let

$$\boldsymbol{A}(\boldsymbol{\theta}) = E\left[\frac{\partial^2 \ln f(y_t, \boldsymbol{X}_t, \boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right]$$

Define another $(k \times k)$ matrix that is just the average of outer products of contributions to the first derivative, that is,

$$\boldsymbol{B}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n}\frac{\partial \ln f(y_t, \boldsymbol{X}_t, \boldsymbol{\theta})}{\partial\theta_i}\cdot\frac{\partial \ln f(y_t, \boldsymbol{X}_t, \boldsymbol{\theta})}{\partial\theta_j} \tag{13.30}$$

and in a similar fashion, let

$$\boldsymbol{B}(\boldsymbol{\theta}) = E\left[\frac{\partial \ln f(y_t, \boldsymbol{X}_t, \boldsymbol{\theta})}{\partial\theta_i}\cdot\frac{\partial \ln f(y_t, \boldsymbol{X}_t, \boldsymbol{\theta})}{\partial\theta_j}\right]$$

We will need one more set of definitions:

$$\boldsymbol{C}_n(\boldsymbol{\theta}) = \boldsymbol{A}_n^{-1}(\boldsymbol{\theta})\boldsymbol{B}_n(\boldsymbol{\theta})\boldsymbol{A}_n^{-1}(\boldsymbol{\theta}) \tag{13.31}$$

and in similar fashion let $\boldsymbol{C}(\boldsymbol{\theta}) = \boldsymbol{A}(\boldsymbol{\theta})^{-1}\boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{A}^{-1}(\boldsymbol{\theta})$.

White showed that, even when the model is misspecified, under suitable conditions the estimator $\hat{\boldsymbol{\theta}}$ that is the solution to the maximization of the likelihood converges to some parameter $\boldsymbol{\theta}_*$.[8] In particular, $\hat{\boldsymbol{\theta}}$ is distributed asymptotically normally, with the covariance matrix $\boldsymbol{C}(\boldsymbol{\theta}_*)$.

Therefore, the appropriate standard errors for the model can be calculated by using the empirical analog to $\boldsymbol{C}(\boldsymbol{\theta}_*)$ evaluated at the estimated parameter $\hat{\boldsymbol{\theta}}$.

How does this discussion relate to the formulae discussed in Chapter 5? Recall that we calculated the standard errors for MLE by evaluating the Hessian matrix of second derivatives at the ML parameter estimates. Now, when the model is correctly specified, a fundamental result, which we will state without proof, is the information matrix equality.[9] At the true value of the parameters $\boldsymbol{\theta}_0$,

$$\boldsymbol{B}(\boldsymbol{\theta}_0) = -\boldsymbol{A}(\boldsymbol{\theta}_0) \tag{13.32}$$

where $-\boldsymbol{A}(\boldsymbol{\theta}_0)$ is merely the information matrix, which is simply the negative of the expected value of the Hessian of the log-likelihood function. Furthermore, when the model is correctly specified, Eq. (13.32) implies

$$\boldsymbol{C}(\boldsymbol{\theta}_0) = -\boldsymbol{A}(\boldsymbol{\theta}_0)^{-1} \tag{13.33}$$

which is exactly what we derived before for ML estimation! When the model is correct, the standard errors can be calculated as we did in Chapter 5. If the model is

[8] The parameter $\boldsymbol{\theta}_*$ corresponds *loosely* to the parameter that represents the best we can do given the misspecification.

[9] See Eq. (5.3) in Chapter 5.

misspecified, however, these standard errors will not be correct and $C_n(\hat{\boldsymbol{\theta}})$ should be used.

Unfortunately, there is little empirical experience using standard errors calculated this way for the probit and logit models. However, if the standard errors calculated under the assumption of misspecification diverge greatly from the ones calculated from standard ML, this difference is the sign of a problem.[10] Perhaps one fruitful approach is to add quadratic or higher-order terms as explanatory variables to the original specification, or specify piecewise linear functions of the explanatory variables, as in Chapter 11.5.1 to make the specification more flexible.

13.7.3 Functional Form: What Is the Right Model to Use?

Again let us adopt the notation of the previous subsection and make the (heroic) assumption of homoscedasticity:

$$\text{prob}(y_i = 1) = F\left[\frac{f(X_i)}{\sigma}\right] \tag{13.34}$$

The problem is often posed as follows: given that $f(X) = X\boldsymbol{\beta}$, what is the correct form for F, cumulative normal, logistic, or something else? Clearly, it is no less heroic to assume F takes on some particular form than it is to assume that $f(X)$ is linear. Given the form for F, one solution to this problem is to be agnostic about how the covariates enter the regression function.

This approach begs the question, what is the right functional form to use? As the foregoing discussion has made clear, there is no simple answer. In some cases it is possible to reject the logit or probit by computing the difference between the maximized value of the likelihood functions, essentially by treating the two models as part of a larger model.[11] Twice the difference between the two log-likelihood functions will be distributed $\chi^2(1)$ although in practice the difference is rarely large enough to discriminate between the two models.

Fortunately, the three models seem to produce similar answers in most empirical applications. Perhaps the best approach is to stick with what is convenient in a particular application, making certain that one's inference does not depend unduly on the particular choices.

One good rule of thumb is to compare the derivatives of the probabilities with respect to X for the linear probability model versus the logit at the mean of one's sample. Recall that the derivative of the probability with respect to X_k in the linear probability model is merely the coefficient β_k. Furthermore, the derivative is constant everywhere. This derivative can be compared to the derivative from the logit

[10] A formal test, called the Information Matrix test, is available although there are some problems with its use. See R. Davidson and J. MacKinnon, *Estimation and Inference in Econometrics,* Oxford University Press, 1993, 578–581.

[11] Ibid., Section 14.3.

model at a typical value of X. Thus, one wishes to compare

$$\beta_k^{\text{LP}} \quad \text{vs} \quad \beta_k^{\text{Logit}}\overline{p}(1-\overline{p}) \tag{13.35}$$

where $\overline{p}$ is merely the proportion of 1s in the sample. Next one might compare the logit to the probit:

$$\phi(\overline{X_i}\boldsymbol{\beta})\beta_k^{\text{Probit}} \quad \text{vs} \quad \beta_k^{\text{Logit}}\overline{p}(1-\overline{p}) \tag{13.36}$$

Notice that $\phi(\overline{X_i}\boldsymbol{\beta})$ is a number that can be read off a standard z table once the mean index has been calculated. An alternative approach that does not require computing the mean index is to compare

$$\phi[\Phi^{-1}(\overline{p})]\beta_k^{\text{Probit}} \quad \text{vs} \quad \beta_k^{\text{Logit}}\overline{p}(1-\overline{p}) \tag{13.37}$$

Generally, the estimates of the derivatives should be roughly similar. For the case when $\overline{p} = .4$,

$$\beta_{\text{Logit}} \approx 1.6\beta_{\text{Probit}} \tag{13.38}$$

The linear probability model (like the logit) has another property that is sometimes important. Consider calculating $\hat{p}$ using the estimated coefficients from the logit, linear probability, and the probit. Next, sum these estimated probabilities over the sample, and consider the following two quantities:

$$\text{Sum}_1 = \sum_{i=1}^{N} \hat{p}_i \tag{13.39}$$

$$\text{Sum}_2 = \sum_{i=1}^{N} y_i \tag{13.40}$$

For the logit and linear probability models,

$$\text{Sum}_1 = \text{Sum}_2 \tag{13.41}$$

whereas in general, for the probit,

$$\text{Sum}_1 \neq \text{Sum}_2 \tag{13.42}$$

It is left as an exercise to the reader to show that Eq.(13.41) in fact holds for the logit and linear probability models (if X includes a constant). In some applications, if the equality in Eq. (13.41) is important, this may argue for the use of the logit or linear probability model instead of the probit.

In sum, the key issue seems to be convenience. All three models generally yield qualitatively similar results. The linear probability model is still frequently used in empirical applications despite the defect of predicted values not being constrained in the (0, 1) interval. When fixed effects are a concern, the linear probability model is easiest to implement with standard statistical software. Likewise, when some of the right-hand-side variables are endogenous and an instrumental variable scheme is required, the linear probability model is often convenient.

On the other hand, the linear probability model is not perfect. If its use can be avoided, one should in general use the logit or probit and evaluate the extent to which one's inference depends on the particular specification of the probability model.

13.8 EXTENSIONS TO THE BASIC MODEL: GROUPED DATA

One frequently encountered extension of the models we have reviewed thus far is their application to grouped data. That is, instead of *individual* data, the unit of observation represents the aggregation of many individuals, often at the state level. In some applications, the unit of observation may be all persons of a specific demographic class, say, women who are 35–45 years old. In other applications, the data come in the form of multiple observations on the same cross-sectional unit.

13.8.1 Maximum Likelihood Methods

The canonical grouped data problem involves J classes, where the X variables are constant within a class. Let y_i be a binary variable that equals 1 when the event occurs, and 0 otherwise. As before, we assume some probability model for the underlying individual (ungrouped data) where:

$$\text{prob}(y_i = 1) = F(X_i\boldsymbol{\beta}) \tag{13.43}$$

where the choice of function F will be described next.

Analogous to our development of the probit in the individual data model [Eqs. (13.13) to (13.17)], given a choice of F, we can write the log-likelihood as follows:

$$l = \sum_{i \in N} \{ y_i \ln[F(X_i\boldsymbol{\beta})] + (1 - y_i)\ln[1 - F(X_i\boldsymbol{\beta})]\} \tag{13.44}$$

Assuming that X variables are constant in each of the J cells allows us to rewrite this as

$$l = \sum_{j \in J} \{ p_j \ln[(F(X_j\boldsymbol{\beta})] + (1 - p_j)\ln[1 - F(X_j\boldsymbol{\beta})]\}\, n_j \tag{13.45}$$

where
$$p_j = \frac{1}{n_j}\sum_{j} y_i$$

where p_j is just the proportion of 1s in the jth class and $n_1, \ldots, n_J$ are the number of observations in each class. Note well that this likelihood function is just the sum over J terms. Given a choice for $F(\cdot)$, we proceed exactly as before. The most common choices for $F(\cdot)$ are the probit and the logit.

Because $J < N$, where N is the number of observations, in the grouped data case we can consider *a fully saturated* model with J parameters. That is, for each class of X's we assign a different parameter, δ_j for $j = 1, \ldots, J$, imposing no restriction on how the covariates might affect the probability. In this case, the log-likelihood function becomes

$$l = \sum_{j \in J} [p_j \ln(\delta_j) + (1 - p_j)\ln(1 - \delta_j)]\, n_j \tag{13.46}$$

The maximum likelihood estimator of this model is $\hat{\delta}_j = p_j$. This fully saturated model represents the best we can do in terms of maximizing the likelihood.

Denote the true probability that a class j experiences the event as π_j. We can write

$$\pi_j = F(X_j\boldsymbol{\beta}) \tag{13.47}$$

where $\dim(\boldsymbol{\beta}) = K$ and $K < J$. If the grouped data model is successful, it is because it summarizes the J cells parsimoniously as a function of a limited number of X variables.

It is now apparent that we can perform a likelihood ratio test, comparing the fully saturated model to our proposed model (the null):

$$\mathrm{LR} = -2\left(\sum_j n_j\{p_j \ln[F(X_j\hat{\boldsymbol{\beta}})] + (1 - p_j)\ln[1 - F(X_j\hat{\boldsymbol{\beta}})]\} - \sum_j n_j[p_j \ln p_j + (1 - p_j)\ln(1 - p_j)]\right)$$

which will be distributed χ^2 with $J - K$ degrees of freedom.

13.8.2 Minimum χ^2 Methods

One option that is often employed on grouped data is minimum χ^2 methods. The point of departure for this type of estimation is the fact that in grouped data we need to fit a finite number of cells. With individual data, the number of observations—the individual y_i's—grows at the same rate as the sample. In the grouped data case, the number of cells [J in Eq. (13.45)] remains fixed. The structure of the grouped data problem allows us the option of suitably transforming the dependent variable and using (weighted) OLS. Table 13.1 describes the most popular minimum χ^2 models. Each can be calculated with conventional OLS software by using the dependent variable described in the table and weighting by the inverse of the square root of the variance given in the last column of the table. The reader familiar with Chapter 10 will also recognize that minimum χ^2 methods are easily cast as GMM.

Consider the linear case first. Let π_j signify the true *population* proportion of people who have experienced the event in the jth class. If we assume the number of observations in each cell grows at a constant rate ($n_j/N \to q_j$) and we let the total

TABLE 13.1
Various minimum χ^2 models for grouped data

Model	Probability	Dependent Variable	Variance(ε)
Linear	$p_j = X\boldsymbol{\beta}$	p_j	$\dfrac{p_j(1 - p_j)}{n_j}$
Log-linear	$p_j = \exp(X\boldsymbol{\beta})$	$\log(p_j)$	$\dfrac{(1 - p_j)}{n_j p_j}$
Probit or "normit"	$p_j = \Phi(X_j\boldsymbol{\beta})$	$\Phi^{-1}(p_j)$	$\dfrac{p_j(1 - p_j)}{n\phi(p_j)^2}$
Logit	$p_j = \Lambda(X_j\boldsymbol{\beta})$	$\log\left(\dfrac{p_j}{1 - p_j}\right)$	$\dfrac{1}{n_j p_j(1 - p_j)}$

number of observations grow large ($N \to \infty$), it is clear that p_j will approach its true value π_j. It is also evident that

$$E[p_j] = \pi_j \tag{13.48}$$

and

$$\operatorname{var}(p_j) = \frac{\pi_j(1-\pi_j)}{n_j} \tag{13.49}$$

Since $E[p_j] = X\beta$ in the linear model, Eq. (13.49) makes clear that the model is heteroscedastic. In fact, given a value for n_j, Eq. (13.49) has a maximum when $\pi = \frac{1}{2}$. When the heteroscedasticity is of a known form, we need only reweight the data so that the weighted error term is homoscedastic.

Replacing the π_j with its estimated value p_j yields a consistent estimate of the variance. Weighted OLS proceeds by taking the *inverse* of the square root of this estimated variance as the appropriate weight.

Similar derivations can be given for the other models presented in Table 13.1. They can be derived straightforwardly using Taylor series expansions to get the approximate variance. These minimum χ^2 methods are consistent and have the same estimated variance as the maximum likelihood models when correctly specified. If we have specified the correct model, our estimators improve as p_j gets closer to the truth π_j, which happens as our sample grows in size.

13.9 ORDERED PROBIT

One simple extension of the framework we have developed so far is the ordered probit. To illustrate, consider an example where the outcome of interest is whether a person works full-time, part-time, or not at all. Define three dichotomous variables as follows:

$$y_i^n = \begin{cases} 1 & \text{if the person does not work} \\ 0 & \text{otherwise} \end{cases}$$

$$y_i^p = \begin{cases} 1 & \text{if the person works part-time} \\ 0 & \text{otherwise} \end{cases}$$

$$y_i^f = \begin{cases} 1 & \text{if the person works full-time} \\ 0 & \text{otherwise} \end{cases}$$

If part-time is defined as between 1 and 35 hours a week, we might consider modeling the choice of work status as arising from the value of a single indicator variable y^*. The higher the value of y^*, the more likely the person is to work. Note that in this case, the outcomes are ordered: no work is less than part-time, and part-time is less than full-time.

The model can be written as

$$\begin{aligned} y_i^n &= 1 \qquad \text{if } y^* < c_1 \\ y_i^p &= 1 \qquad \text{if } c_1 < y^* < c_2 \\ y_i^f &= 1 \qquad \text{if } y^* > c_2 \end{aligned}$$

where c_1 and c_2 are the thresholds that the latent variable must cross to change the value of y. Analogous to our previous development, we choose an appropriate function F (either the logit or probit is computationally convenient) and compute the relevant probabilities:

$$\begin{aligned}\text{prob}(y_i^n = 1) &= F(c_1 - X\boldsymbol{\beta})\\ \text{prob}(y_i^p = 1) &= F(c_2 - X\boldsymbol{\beta}) - F(c_1 - X\boldsymbol{\beta})\\ \text{prob}(y_i^f = 1) &= 1 - \text{prob}(y_i^n = 1) - \text{prob}(y_i^p = 1)\end{aligned}$$

Note that the last line implies that

$$\text{prob}(y_i^f = 1) = 1 - F(c_2 - X\boldsymbol{\beta})$$

For the remainder of our discussion let us focus on the *ordered probit* case instead of the ordered logit so that F is just the cumulative standard normal density. (In practice, there is apparently little difference between the two.)

What is identified in this model? When X includes just a constant and a single covariate, for example, we can write

$$X_i\beta = \alpha + \delta z_i$$

Our probabilities are

$$\begin{aligned}\text{prob}(y_i^n = 1) &= \Phi\left(\frac{c_1 - \alpha - \delta z_i}{\sigma}\right)\\ \text{prob}(y_i^p = 1) &= \Phi\left(\frac{c_2 - \alpha - \delta z_i}{\sigma}\right) - \Phi\left(\frac{c_1 - \alpha - \delta z_i}{\sigma}\right)\\ \text{prob}(y_i^f = 1) &= 1 - \Phi\left(\frac{c_2 - \alpha - \delta z_i}{\sigma}\right)\end{aligned}$$

When there are only three choices, we can without loss of generality set $c_1 = 0$, leaving only one threshold to estimate. If we denote the only remaining threshold c_2 as c, it is clear that we can identify

$$\frac{c}{\sigma}, \frac{\alpha}{\sigma}, \frac{\delta}{\sigma}$$

Therefore, just like the probit, we can identify the parameters up to some factor of proportionality. This property makes sense because we could obviously multiply c, α, δ, and σ by 2 and leave the decision probabilities the same. As in the standard probit, it is the *ratio* of the parameters to σ that matters. It is therefore often convenient to normalize σ to equal 1.

One complication arises in the ordered probit case that does not in the simple probit. Consider the derivatives of the probabilities with respect to z:

$$\begin{aligned}\frac{\partial}{\partial z_i}\text{prob}(y_i^n = 1) &= -\phi(-\alpha - \delta z_i)\delta\\ \frac{\partial}{\partial z_i}\text{prob}(y_i^p = 1) &= -\left[\phi(c - \alpha - \delta z_i) - \phi(-\alpha - \delta z_i)\right]\delta\\ \frac{\partial}{\partial z_i}\text{prob}(y_i^f = 1) &= \phi(c - \alpha - \delta z_i)\delta\end{aligned}$$

where we have imposed the normalization $\sigma = 1$. In the case when δ is positive, for example, an increase in z unambiguously decreases the probability of not working and increases the probability of working. The probability of part-time work, however, will depend on the size of the threshold c, among other things; and the sign of the effect is in general ambiguous.

This example raises another interesting question. Why are part-time and full-time distinct states? Another possible model might be

$$y_i^n = \begin{cases} 1 & \text{if the person does not work} \\ 0 & \text{otherwise} \end{cases}$$

$$y_i^w = \begin{cases} 1 & \text{if the person works} \\ 0 & \text{otherwise} \end{cases}$$

That is, part-time and full-time are not distinct states, and there is no second threshold (c_2 in the previous example). The question is an empirical one that would seem to set up a possible specification test. In particular, it would seem that a Hausman test (see the discussion in Chapter 10) or merely an eyeball comparison of the estimates derived for the three-state model using the ordered probit versus the estimates of the two-state model using the probit would be useful. The two sets of estimates should be roughly similar, if the two-state model is correct.

13.10 TOBIT MODELS

So far we have dealt with models that are purely categorical—the outcomes are discrete. There is also a broad class of models that have both discrete and continuous parts. One important model in this category is the Tobit.[12] The Tobit is an extension of the probit, but as we will see it is really one approach to dealing with the problem of censored data.

13.10.1 The Tobit as an Extension of the Probit

First consider a standard probit model for the decision to buy an automobile in a given week. Define a variable y^* that is a simple index of a person's desire for an automobile, and define a variable y_i that equals 1 if the person buys a car and 0 otherwise. Formally,

$$y_i^* = X_i\boldsymbol{\beta} + \epsilon_i \tag{13.50}$$

where $\quad \epsilon \sim N(0, \sigma^2), \quad$ and $\quad y_i = \begin{cases} 1 & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0 \end{cases} \quad (13.51)$

Now suppose that, instead of observing merely the decision to buy a car, we also have data on the actual amount spent on the car. One natural extension of this probit

[12] The name is a reference to Tobin, who developed the model. See J. Tobin, "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, **26**, 1958, 24–36.

is called the Tobit (Tobin's probit) and is given by the following:

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

where y^* is defined in Eq. (13.50). This model is called a *censored regression model* because it is possible to view the problem as one where observations of y^* at or below zero are *censored.* That is, we could write the model as

$$y_i = \max(0, X_i\boldsymbol{\beta} + \epsilon_i) \tag{13.52}$$

This representation should be contrasted to *truncation,* which occurs when we do not observe the X variables either. That is, in truncated data both the dependent and independent variables are missing.

The likelihood function for the Tobit is instructive. For all observations such that $y^* \leq 0$ the contribution to the likelihood will be given by prob$(y^* < 0)$, which is

$$\begin{aligned} &\text{prob}(-X_i\boldsymbol{\beta} \leq \epsilon_i) \\ &\quad = \text{prob}\left(\frac{-X_i\boldsymbol{\beta}}{\sigma} \leq \frac{\epsilon_i}{\sigma}\right) \\ &\quad = 1 - \Phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right) \end{aligned}$$

For an observation $y_i^* > 0$, the contribution to the likelihood is

$$\text{prob}(y^* > 0)\phi(y_i^* \mid y_i^* > 0) = \Phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)\frac{1}{\sigma}\frac{\phi[(y_i - X_i\boldsymbol{\beta})/\sigma]}{\Phi(X_i\boldsymbol{\beta}/\sigma)} \tag{13.53}$$

Putting both parts together, we get the likelihood function:

$$L = \prod_{y_i|y_i=0}\left[1 - \Phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)\right] \cdot \prod_{y_i|y_i>0}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\frac{(y_i - X_i\boldsymbol{\beta})^2}{\sigma^2}\right] \tag{13.54}$$

Several points are worth noting. First, the second part of the likelihood resembles the likelihood for conventional OLS on those sample points that are not censored (i.e., greater than 0). The first part resembles the probit. The log-likelihood is

$$\begin{aligned} l = &\sum_{y_i|y_i=0}\ln\left[1 - \Phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)\right] \\ &+ \sum_{y_i|y_i>0}\left[\ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2}\frac{(y_i - X_i\boldsymbol{\beta})^2}{\sigma^2}\right] \end{aligned} \tag{13.55}$$

Second, note that, unlike the probit where normalizing $\sigma = 1$ is harmless, the same is not so for the Tobit. (As we discuss shortly, this difference also causes serious problems for the Tobit in the presence of heteroscedasticity.) It is true that in the first part of the likelihood only the ratio β/σ is "identified." In the second part of the likelihood, however, both the slope coefficients and σ are separately identifiable. This observation should be no surprise, for the same is true for an OLS regression.

Third, it may not always be sensible to interpret the coefficients of a Tobit in the same way as one interprets coefficients in an uncensored linear model. Consider the following three derivatives with respect to a particular variable x_k for observation i:

$$\frac{\partial E[y_i^* \mid X_i]}{\partial x_k} = \beta_k$$

$$\frac{\partial E[y_i \mid X_i]}{\partial x_k} = \Phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)\beta_k$$

$$\frac{\partial E[y_i \mid X_i, y_i^* > 0]}{\partial x_k} = \beta_k\left[1 - \frac{X_i\boldsymbol{\beta}}{\sigma}\frac{\phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)} - \left(\frac{\phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)}\right)^2\right]$$

Any of these could be of interest! The simple coefficient, $\partial E[y_i^* \mid X_i]/\partial x_k$, is most likely to be of interest in cases like top-coding discussed shortly, where censoring is more of an annoyance than a fundamental aspect of the relationship that one is interested in.

McDonald and Moffit proposed the following decomposition that some find useful:[13]

$$\begin{aligned}\frac{\partial E[y_i \mid X_i]}{\partial x_k} &= \left\{\Phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)\beta_k\left[1 - \frac{X_i\boldsymbol{\beta}}{\sigma}\frac{\phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)} - \left(\frac{\phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)}\right)^2\right]\right\} \\ &\quad + \left\{\frac{\beta_k}{\sigma}\phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)\cdot\left(X_i\boldsymbol{\beta} + \frac{\phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{X_i\boldsymbol{\beta}}{\sigma}\right)}\right)\right\} \\ &= \text{prob}(y_i^* > 0)\cdot\frac{\partial E[y_i \mid X_i, y_i^* > 0]}{\partial x_k} + \frac{\partial\,\text{prob}(y_i^* > 0)}{\partial x_k}E[y_i \mid X_i, y_i^* > 0]\end{aligned}$$

The interpretation is that the change in the mean of y with respect to x_k has two components. One effect works by changing the conditional mean of y, the first part, and the other by changing the probability that an observation will be positive, the second part. Whether or not one finds this decomposition useful depends on the nature of the problem.

Finally, *if the true model is a Tobit* then merely ignoring the censoring problem and performing OLS is incorrect. A useful way to see this is to consider a very special case. Consider the standard latent variable formulation:

$$\boldsymbol{y}^* = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad (13.56)$$

and let x and y^* be distributed *jointly* normal. (Note that this is, generally speaking, an unrealistic assumption.) Consider ignoring the censoring problem and running OLS on all the observations. Then

$$\begin{aligned}\text{plim}\,\hat{\beta}_{\text{OLS}} &= \Sigma_{xx}^{-1}\Sigma_{xy^*}\cdot\text{prob}(y^* > 0) + 0\cdot\text{prob}(y^* \leq 0) &\qquad (13.57)\\ &= \beta\cdot\text{prob}(y^* > 0) &\qquad (13.58)\end{aligned}$$

Because prob($y^* > 0$) is less than 1, OLS will be attenuated. Given the joint normality assumption, a simple consistent method of moments estimate can be cal-

[13] J. McDonald and R. Moffit, "The Uses of Tobit Analysis," *Review of Economics and Statistics*, **62**, 1980, p. 318–321.

culated. Note that a consistent estimate of prob$(y^* > 0)$ is merely n_1/N, the ratio of uncensored observations n_1 to the total number of observations. It is straightforward to observe that a consistent estimate of $\boldsymbol{\beta}$ can be constructed by "undoing" this bias:

$$\hat{\boldsymbol{\beta}}_{\text{Consistent}} = \hat{\boldsymbol{\beta}}_{\text{OLS}} \cdot \frac{N}{n_1}$$

Although the consistency of this estimator is not guaranteed for the case where y^* and x are not jointly normally distributed (if some of the explanatory variables contain dummy variables, for example, joint normality is ruled out) it apparently performs reasonably well for non-normal distributions. We consider it here, however, not because it should be used but because it provides some intuition for the consequences of ignoring the censoring problem.[14]

13.10.2 Why Not Ignore "The Problem"?

Although we have developed the Tobit as a natural extension of linear regression, it is not really so. For example, in some instances one can blithely (and correctly) ignore the zeros problem. Consider the case of a tobacco company executive interested in the effect of workplace restrictions on the quantity of cigarettes smoked. Let T_i be a binary variable that equals 1 if restrictions on smoking are in effect at individual i's workplace, and let C_i be the number of cigarettes smoked by individual i. Let the model be

$$C_i = T_i\gamma + \epsilon_i$$

To keep the example manageable, suppose that it is appropriate to treat T_i as exogenous. It would seem that the tobacco company is interested in $E[C_i \mid T_i]$, not $E[C_i \mid T_i, C_i > 0]$ nor prob$(C_i > 0 \mid T_i)$, the usual objects of interest in the Tobit. That is, it is only interested in the total number of cigarettes smoked, not whether the number of smokers or the amount smoked conditional on smoking has changed. In that case, OLS will consistently estimate the average *treatment* effect. This particular case is easy since the simple difference in means between treated and untreated groups is the most obvious relationship of interest. When we go beyond this simple case, say, when there are explanatory variables that are not binary, the problem is more vexing since a simpler linear relationship is less likely to be adequate. Some have suggested that radical alternatives to the Tobit be considered, but further discussion would lead us too far astray.[15]

Even when we are interested in both parts of the likelihood, so to speak, it may still be the case that the Tobit imposes a structure on the data that is not always appropriate. One way to understand this is to consider a specification test based on the following fact: *If the Tobit is the correct specification, then the ratio of the Maximum*

[14] *Note:* This is not a hard-and-fast rule when the variables are not jointly normally distributed.

[15] See A. Deaton, "Econometric Issues for Survey Data," a chapter in *The Analysis of Household Surveys: A Microeconomic Approach to Development Policy,* Johns Hopkins University Press for the World Bank, forthcoming, for a lucid exposition of the issues involved.

Likelihood estimates from the Tobit, $\hat{\beta}_T/\hat{\sigma}_T$, should be the same as the probit coefficients from the same data, treating nonzero values as 1 and 0 values as 0. That is, the Tobit imposes the condition that the relationship generating the ones and zeros is the same as the process that produces the positive values.

Although it is uncommon in applied work, a Hausman test based on this fact may be useful. At a minimum, an "eyeball" test comparing the ratio $\hat{\beta}_T/\hat{\sigma}_T$ to the probit estimate of the same quantity, $\hat{\beta}_P/\hat{\sigma}_P$, is recommended. If they are very different, it suggests that the Tobit is misspecified. A test based on this idea can be found in Fin and Schmidt.[16]

The simplest case where such a specification test might be expected to reveal a problem is when the explanatory variables have different effects on the participation decision and the decision to consume conditional on participation. One example is the effect of advertising on smoking. For example, perhaps advertising has a strong influence on whether someone begins smoking, but marginal changes in advertising have little effect on the consumption of someone who already smokes. Since the Tobit constrains the participation equation and the consumption equation to have the same parameters, the Tobit is misspecified in this case, and this misspecification may have profound and undesirable consequences for your estimates. Such a case is perhaps better handled by viewing the problem as involving two equations, and we discuss such an approach in Section 13.12.

One case where there are fewer alternatives to the Tobit involves *top-coding*. A prime example is wage or earnings data. For reasons of confidentiality, data on an individual's wage is often top-coded if it exceeds some value. For example, in some data sets a person's wage is only recorded if it is less than or equal to \$999/hour. If it exceeds \$999/hour, no wage is recorded, but another variable records whether or not this top-coding occurred.

If we are interested in a standard wage regression, one approach to this problem would be to treat it as a Tobit,

$$y_i = \min(999, X_i\boldsymbol{\beta} + \epsilon_i) \tag{13.59}$$

Development of the likelihood for this case is identical to the previous case we have examined and is a feasible alternative to, say, "throwing away" the observations that are censored, an approach that almost always leads to biased estimates.

13.10.3 Heteroscedasticity and the Tobit

Suppose we have decided that the Tobit is an appropriate model. Now consider the effect of heteroscedasticity in this model. As it turns out, this problem is much more vexing than in the probit or the standard linear model.

A simple Monte Carlo illustration will help. We first need to describe the data generation process. The true model is given by the following:

$$y_i^* = x_i - 10 + \epsilon_i \tag{13.60}$$

[16] T. Fin and P. Schmidt, "A Test of the Tobit Specification against an Alternative Suggested by Cragg," *Review of Economics and Statistics*, **66**, 1984, 35–57.

TABLE 13.2
Heteroscedasticity in the Tobit (results from 500 Monte Carlo simulations)

Statistic	Mean	Standard deviation	Minimum	Maximum
Percent censored	.423	.029	.340	.495
OLS slope estimate	.975	.125	.631	1.427
Tobit slope estimate	1.678	.211	1.133	2.329

where x takes on 200 equally spaced values from 0.02 to 40, and $\epsilon_i \sim N(0, x_i^2)$. To finish the data generation process, we define

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \le 0 \end{cases}$$

That is, we have the standard Tobit model except that we have allowed the errors to be heteroscedastic.

The results from a simulation study with 500 replications are presented in Table 13.2. They are not very encouraging. Since the true slope coefficient is 1, notice that OLS outperforms the Tobit in terms of both bias and variance! Unfortunately this lesson is not a general one. It is easy to create situations where the Tobit does better than OLS under heteroscedasticity, or when both perform quite badly. Problem 8 at the end of this chapter asks you to verify this for yourself by comparing the results from estimation where none of the data are censored, to results obtained from using the Tobit, where you censor the data yourself. In any event, the key lesson to take from this illustration is that heteroscedasticity in the Tobit is likely to be a much more serious problem than in the logit or probit. In particular, the problem is that misspecification of σ has profound consequences for β, and the two are separately identifiable.

13.11
TWO POSSIBLE SOLUTIONS

The message from the foregoing Monte Carlo is rather gloomy, although it has not prevented widespread use of the Tobit. Having no wish merely to "dump the problem into the student's lap," we discuss briefly two recent developments due to Jim Powell that allow consistent estimation of the Tobit, even in the face of heteroscedasticity.[17] Interest in these models is increasing as computational burdens have become more manageable, although neither of these techniques has become commonplace among applied researchers.

[17] J. Powell, "Least Absolute Deviations Estimation for the Censored Regression Model," *Journal of Econometrics*, **25**, 1984, 303–325; and "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, **54**, 1986, 1435–1460.

13.11.1 Symmetrically Trimmed Least Squares

The idea behind symmetrically trimmed least squares is quite intuitive. As before, consider the standard index model:

$$y^* = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{13.61}$$

where we do not observe y^* but rather y, where

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

We can write this as $$y_i = \begin{cases} y_i^* & \text{if } \epsilon_i > -X_i\boldsymbol{\beta} \\ 0 & \text{if } \epsilon_i \leq -X_i\boldsymbol{\beta} \end{cases}$$

Powell first notes that if y^* were observed, and if the error term was *symmetrically* distributed around 0, then standard OLS would produce consistent estimates of the parameters. Censoring is a problem because it introduces an asymmetry into the distribution. The situation is displayed in Fig. 13.9. For a given observation X_i we do not observe only of y^*. Instead, we observe only the area to the right of 0. That is, all observations where $\epsilon_i < -X_i\boldsymbol{\beta}$ are omitted. Now imagine we *also* truncated observations such that $\epsilon_i > X_i\boldsymbol{\beta}$. That is, we would delete points to the right of $2X_i\boldsymbol{\beta}$ in Fig. 13.9. If the data were "trimmed" this way, the resulting error distribution would again be symmetric.

Powell notes that if we knew the true value of $\boldsymbol{\beta}_0$ we could merely replace y_i with the minimum of the quantities $\{y_i, 2X_i\boldsymbol{\beta}_0\}$ and generate a consistent estimate of $\boldsymbol{\beta}_0$. Equivalently, we could define

$$\epsilon_i^* = \max(\epsilon_i, -X_i\boldsymbol{\beta}_0) \tag{13.62}$$

and

- Replace ϵ_i^* with $\min\{\epsilon_i^*, X_i\boldsymbol{\beta}_0\}$ if $X_i\boldsymbol{\beta}_0 > 0$
- Delete the observation otherwise

Furthermore, the true value of the coefficient $\boldsymbol{\beta}_0$ would satisfy the following normal equation:

$$\sum_{i=1}^{n} 1(X_i\boldsymbol{\beta}_0 > 0) \cdot (\min\{y_i, 2X_i\boldsymbol{\beta}_0\} - X_i\boldsymbol{\beta}_0)X_i' = 0 \tag{13.63}$$

which can be viewed as the minimand of the following objective function:

$$\begin{aligned} M(\beta) &= \sum_{i=1}^{n}\left[y_i - \max\left(\frac{1}{2}y_i, X_i\boldsymbol{\beta}\right)\right]^2 \\ &\quad + \sum_{i=1}^{n} 1(y_i > 2X_i\boldsymbol{\beta})\left\{\left(\frac{1}{2}y_i\right)^2 - [\max(0, X_i\boldsymbol{\beta})]^2\right\} \end{aligned} \tag{13.64}$$

Of course, we do not observe $\boldsymbol{\beta}_0$. Powell makes use of a notion called self-

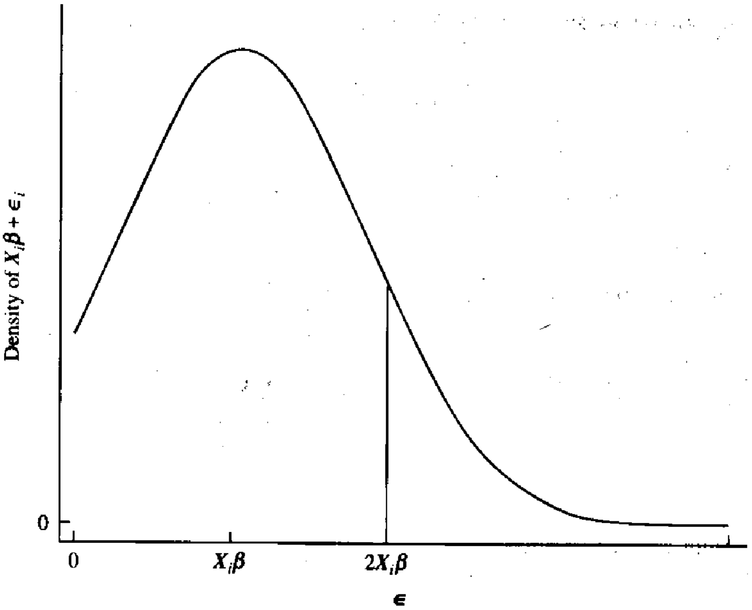


FIGURE 13.9
Density function of y and "symmetrically trimmed sample."

consistency to show that an estimate of β_0 that is consistent with being a solution to Eq. (13.63), the normal equation, will produce a consistent estimate of β_0.[18]

It is straightforward to find a consistent estimate of $\hat{\beta}$ by use of the following iterative algorithm:

1. Compute an initial estimate $\hat{\beta}$, say, OLS, on the original data.
2. Compute the predicted value:
   - If the predicted value is negative, set the observation to missing.
   - If the value of the dependent variable is greater than twice the predicted value, set the value of the dependent variable equal to $2X_i\beta$.
3. Run OLS on these altered data.
4. Use this β on the original data and repeat until β stops changing.

The covariance matrix of $\hat{\beta}$ is a bit more involved but straightforward. Define

$$C_n = \frac{1}{n}\sum_{i=1}^{n} E[1(-X_i\beta_0 < \epsilon_i < X_i\beta_0) \cdot X_i'X_i] \tag{13.65}$$

$$D_n = \frac{1}{n}\sum_{i=1}^{n} E[1(X_i\beta_0 > 0) \cdot \min\{\epsilon_i^2, (X_i\beta_0)^2\}X_i'X_i] \tag{13.66}$$

[18] Note that even when $\beta_0 \neq 0$, inconsistent solutions such as $\hat{\beta} = 0$ will satisfy the normal equations. Under certain regularity conditions, however, Powell shows that the global minimum of Eq. (13.64) is consistent.

The appropriate covariance estimator is

$$\hat{C}^{-1}\hat{D}\hat{C}^{-1}$$

where $\hat{C}^{-1}$ and $\hat{D}$ are consistent estimates of the matrices in Eqs. (13.65) and (13.66), respectively.

One attractive feature of this method is that it is robust to the presence of heteroscedasticity as long as the true distribution of the error term is symmetric and unimodal. It is most useful when the amount of censoring is not "too severe." The procedure also should not be used with limited amounts of data, as the evidence suggests there can be an important loss in efficiency.

13.11.2 Censored Least Absolute Deviations (CLAD) Estimator

Another approach also due to Powell is the censored least absolute deviations (CLAD) estimator. It requires weaker assumptions on the error term than the symmetrically trimmed estimator and is also easy to compute with standard software. Again it is not a procedure that should be implemented with small amounts of data, but Monte Carlo evidence suggests that it performs well, especially when any of the distributional assumptions of the Tobit are violated.[19] A complete discussion of the iterative solution described here would require a discussion of quantile regression, which we will not pursue, but again the basic ideas are intuitive.[20]

We are again interested in the basic censored regression model, but imagine for the moment that y_i^* in Eq. (13.61) is observed. We can write

$$E\left[y_i^* \mid X_i\right] = X_i\boldsymbol{\beta} + E[\epsilon_i \mid X_i] = X_i\boldsymbol{\beta} \tag{13.67}$$

A consistent estimate can be obtained by OLS that is the solution to

$$\min_{\hat{\boldsymbol{\beta}}}\left[\sum_{i=1}^{n}(y_i^* - X_i\hat{\boldsymbol{\beta}})^2\right] \tag{13.68}$$

That is, $\hat{\boldsymbol{\beta}}$ is the estimator that minimizes the sum of squared errors. Suppose we instead chose to minimize the sum of the *absolute value* of the errors:[21]

$$\min_{\hat{\boldsymbol{\beta}}}\left[\sum_{i=1}^{n}\left|y_i^* - X_i\hat{\boldsymbol{\beta}}\right|\right] \tag{13.69}$$

The estimator formed this way is called the **least absolute deviations** (LAD) estimator for obvious reasons. Some insight into what this estimator does can be gleaned

[19] See, for example, H. Paarsch, "A Monte Carlo Comparison of Estimators for Censored Regression Models," *Journal of Econometrics*, **24**, 1984, 197–213.

[20] The algorithm described here is from M. Buchinsky, "Changes in the U.S. Wage Structure 1963–1987: Application of Quantile Regression," *Econometrica*, **62**, 1994, 405–458, especially Section 3.3. The paper also contains an extended example of its use with the March Current Population Survey.

[21] Note that without the absolute value operator the minimand goes off to $-\infty$.

by noting that Eq. (13.69) can be rewritten as

$$\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{n} (y_i^* - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}) \cdot \operatorname{sgn}(y_i^* - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}) \tag{13.70}$$

where the sign function sgn(·) takes on values of 1, 0, or −1 as the argument is positive, zero, or negative. The corresponding normal equation is given by

$$\mathbf{0} = \sum_{i=1}^{n} \boldsymbol{X}_i' \cdot \operatorname{sgn}(y_i^* - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}) \tag{13.71}$$

In this formulation it is apparent that it is the sign of the residuals and not their magnitude that matters. As it turns out the LAD estimator corresponds to **median** regression, which is consistent for $\boldsymbol{\beta}$ because

$$q_{50}[y_i^* \mid \boldsymbol{X}_i] = \boldsymbol{X}_i \boldsymbol{\beta} + q_{50}[\epsilon_i \mid \boldsymbol{X}_i] = \boldsymbol{X}_i \boldsymbol{\beta} \tag{13.72}$$

In Eq. (13.72), q_{50} denotes the median or fiftieth quantile. For the reader who has suffered with us this far, this observation has a payoff for the censored regression model. In particular, OLS, which corresponds to mean regression, is inconsistent in the censored regression model because

$$E[\max\{0, y_i^*\} \mid \boldsymbol{X}_i] = \boldsymbol{X}_i \boldsymbol{\beta} + E[\epsilon \mid \boldsymbol{X}_i, \epsilon_i > -\boldsymbol{X}_i \boldsymbol{\beta}] \neq \boldsymbol{X}_i \boldsymbol{\beta} \tag{13.73}$$

The median, unlike the mean, is not affected by the "max" transformation. In particular,

$$q_{50}[\max\{0, y_i^*\} \mid \boldsymbol{X}_i] = \boldsymbol{X}_i \boldsymbol{\beta} + q_{50}[\epsilon \mid \boldsymbol{X}_i, \epsilon_i > -\boldsymbol{X}_i \boldsymbol{\beta}] = \boldsymbol{X}_i \boldsymbol{\beta} \tag{13.74}$$

Note that Eq. (13.74) will be true for very general forms of the error. In particular, no assumptions about homoscedasticity are necessary, and normality is also dispensed with.

This observation also suggests a simple iterative procedure for calculating a consistent estimate of $\boldsymbol{\beta}$ in the censored regression model, provided one's software can compute median regression, a feature that is becoming increasingly popular. In particular,

1. Run LAD on the entire sample to generate an initial estimate of $\boldsymbol{\beta}$.[22]
2. Use this estimate of $\boldsymbol{\beta}$ to drop observations for which the predicted value is negative.
3. Run LAD on this new sample, to calculate a new estimate of $\boldsymbol{\beta}$.
4. Repeat Steps 2 and 3 using the $\hat{\boldsymbol{\beta}}$ in Step 3 as the new initial estimate.
5. Continue until the estimate stops changing.

One difficulty with this iterative algorithm is that it is not guaranteed to find the global minimum. This problem does not appear to be significant in practice, but care should be exercised. One approach is to start the iterative routine at different starting

[22] *Note:* In some packages a general routine for calculating quantile regressions is available. In such packages, LAD corresponds to a quantile regression at the fiftieth quantile.

values to ensure a global minimum is found. Finally, standard errors are also a bit difficult. The most practical approach may be to bootstrap the entire process (see Chapter 11), although the limited empirical experience with this model suggests caution.

13.12 TREATMENT EFFECTS AND TWO-STEP METHODS

Listing the various models that combine limited and continuous variables would be an arduous task. Amemiya has a nice summary of many of these models.[23] The class of problems where continuous and discrete variables are conflated constitutes an immense (and confusing!) literature.

In this section we take a brief look at a class of problems involving a treatment (typically a dichotomous variable) and an outcome (typically a continuous variable). The terminology derives from biology and medicine, where the dichotomous variable of interest is often a new drug or a particular type of therapeutic regime and the outcome is some measure of the effect of the treatment—an increase in life span or the quantity of healthy red blood cells, for example.

Heckman has developed a general model that nests many of these models and is sometimes used in its most general form.[24] The model is given by

$$y_{1i} = X_{1i}\boldsymbol{\beta}_1 + \epsilon_{1i} \tag{13.75}$$

$$y_{2i} = X_{2i}\boldsymbol{\beta}_2 + \epsilon_{2i} \tag{13.76}$$

$$T_i = 1(Z_i\gamma + \epsilon_{0i} > 0) \tag{13.77}$$

$$y_i = T_i y_{1i} + (1 - T_i)y_{2i} \tag{13.78}$$

where T_i, the treatment, is an indicator variable that takes on values of 1 or 0 as the statement $1(\cdot)$ is true or false, respectively. The continuous measures y_1 and y_2 describe the relationship between the outcome and the covariates if the individual does or does not, respectively, get the treatment.

Two issues are at the core of this model, which we will denote (somewhat arbitrarily) as follows:

1. **Treatment effect heterogeneity.** The effect of a treatment often varies across individuals depending on their characteristics.

   Consider again the effect of unionization on wages. Whereas it is generally true that, other things being equal, union wages are higher than nonunion wages, the effect of union status is more subtle than merely changing the intercept in a wage equation. Let Eq. (13.75) be the equation that determines wages in the union sector, and Eq. (13.76) the corresponding equation for the nonunion sector. It has generally been found that the returns for further schooling tend to be lower in unionized jobs. If the kth column of X was schooling, for example, this observa-

[23] T. Amemiya, "Tobit Models: A Survey," *Journal of Econometrics*, **24**, 1984, 3–63.

[24] J. Heckman, "Varieties of Selection Bias," *American Economic Review*, **80**, 1990, 313–318.

tion suggests that $\beta_1^k < \beta_2^k$ where the superscript denotes that we are referring to the kth element of the vector. In this case, the union effect would vary with X. In particular,

$$\text{Effect on worker } i = X_i(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \tag{13.79}$$

2. **Selectivity.** "Selectivity concerns the presence of some characteristic of the treatment (or control) group that is both associated with receipt of the treatment and associated with the outcome so as to lead to a false attribution of causality regarding treatment and outcomes."[25]

   Consider the evaluation of the "quality" of a given private school. Suppose there is agreement on the outcome measure, say, postschool wages. For purposes of illustration, further suppose that the econometrician does not have information on family background. What if the school administrator does not admit students randomly? For example, suppose the administrator admits students selectively, preferring students from wealthy families to those from poor families. If students from wealthy families are more likely to be wealthy for reasons other than their "superior" schooling, then ignoring this selection may lead the econometrician to confound the effect of family background with the effect of training received at a particular private school.

Sometimes both problems can be present at once. Consider the union example again. Heterogeneity in treatment is the acknowledgment that there are different wage equations in the union and nonunion sectors. The simplest way to account for this is to run separate regressions for both sectors. If assignment into the union sector is not random, however, our estimates in each equation may be "contaminated" by selectivity bias. One of Heckman's insights was that it is sometimes possible to control for this simply. The reader is again forewarned, however, that the estimation methods described next are, like the Tobit, very sensitive to violations of underlying assumptions.

13.12.1 The Simple Heckman Correction

In a highly influential article, Heckman proposed a simple two-step method to deal with many of these models.[26] This two-step method is often used in situations where "selectivity bias" may be present.

A classic example illustrating the possible consequences of selectivity bias is due to Gronau, where the outcome is a woman's wage, and the treatment is her decision to participate in market work.[27] What are the determinants of women's wages?

[25] B. Barnow, G. Cain, and A. Goldberger, "Issues in the Analysis of Selectivity Bias," *Evaluation Studies Review Annual*, **5**, 1976, 43–59.

[26] J. Heckman, "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, **5**, 1976, 475–492.

[27] R. Gronau, "Wage Comparisons: A Selectivity Bias," *Journal of Political Economy*, **82**, 1974, 1119–1155.

Although the model fits into the foregoing, more general framework, it will simplify matters if we take a slightly different tack.

The simplest idea would be to fit the following equation on a sample of working women:

$$w_i = X_i\boldsymbol{\beta} + \epsilon_{1i} \tag{13.80}$$

where w is the log wage, and X is the vector of characteristics such as work experience, years of schooling, etc. It is argued, however, that the sample of women involved in "market work" (i.e., those who work for wages) is not a random sample of women, and that this selectivity may bias the coefficients. Formally, we can write down a participation equation:

$$T_i = 1(\boldsymbol{Z}_i\gamma + \epsilon_{0i} > 0) \tag{13.81}$$

where Z includes variables that predict whether or not a woman works. A woman works if $\boldsymbol{Z}_i\gamma > -\epsilon_{0i}$ or, equivalently, if $-\boldsymbol{Z}_i\gamma < \epsilon_{0i}$. Note that Z and X may include common variables, and in some empirical examples they are identical. In Gronau's case, Z also included the number of small children. Presumably the presence of small children might affect a woman's decision to work but should not have an effect on her wages. The selectivity problem is apparent by taking expectations of Eq. (13.80) over the sample of working women:

$$E[w_i \mid X_i, T_i = 1] = X_i\boldsymbol{\beta} + E[\epsilon_{1i} \mid \epsilon_{0i} > -\boldsymbol{Z}_i\gamma] \tag{13.82}$$

If ϵ_0 and ϵ_1 are jointly normally distributed we can write

$$\epsilon_{1i} = \frac{\sigma_{0,1}}{\sigma_0^2}\epsilon_{0i} + \nu_i \tag{13.83}$$

where ν_i is uncorrelated with ϵ_{0i}, $\sigma_{0,1}$ is the covariance between ϵ_{0i} and ϵ_{1i}, and σ_0^2 is the variance of ϵ_{0i}.[28] This last observation is useful because we can now write

$$\begin{aligned} E[\epsilon_{1i} \mid \epsilon_{0i} > -\boldsymbol{Z}_i\gamma] &= \frac{\sigma_{01}}{\sigma_0}E\left[\frac{\epsilon_{0i}}{\sigma_0} \mid \frac{\epsilon_{0i}}{\sigma_0} > \frac{-\boldsymbol{Z}_i\gamma}{\sigma_0}\right] \\ &= \frac{\sigma_{01}}{\sigma_0}\frac{\phi(\boldsymbol{Z}_i\gamma/\sigma_0)}{\Phi(\boldsymbol{Z}_i\gamma/\sigma_0)} \end{aligned} \tag{13.84}$$

where $\phi(\cdot)$ is the standard normal density and $\Phi(\cdot)$ its cumulative distribution function. It is now evident why OLS estimates of Eq. (13.80) may be biased. In particular, the last expectation in Eq. (13.82) may not be zero. Selectivity bias is said to occur whenever σ_{01} is not zero.

Heckman noted that the problem with using OLS on Eq. (13.80) is that $\hat{\boldsymbol{\beta}}$ is generally biased owing to the presence of an omitted variable, where the quantity (sometimes called the *inverse Mills ratio*)

$$\frac{\phi(\boldsymbol{Z}_i\gamma/\sigma_0)}{\Phi(\boldsymbol{Z}_i\gamma/\sigma_0)} \tag{13.85}$$

[28] For further discussion of this and related formulas see Maddala, op. cit., 365–370.

is the omitted variable. If this omitted variable were included in the OLS regression, as in

$$w_i = X_i\beta + \frac{\phi(Z_i\gamma/\sigma_0)}{\Phi(Z_i\gamma/\sigma_0)}\tilde{\sigma} \tag{13.86}$$

then consistent estimates would be straightforward. Heckman noted that such a model could be easily estimated with the following two-step estimator:

1. Run a probit of the treatment on the vector Z to obtain estimates of γ/σ_0.
2. Use these estimates to construct the inverse Mills ratio.
3. Run OLS of the outcome on X as in Eq. (13.86), using the *estimated* inverse Mills ratio as an additional regressor.

An estimate of σ_{01}/σ_0 can be read off as the coefficient $\tilde{\sigma}$ on the inverse Mills ratio. Standard errors are a bit more complicated because the resulting model is heteroscedastic and uses estimated values. Merely adjusting the standard errors for heteroscedasticity will not be adequate in general, because such a correction fails to account for the loss in precision that results from using estimates of the inverse Mills ratio instead of actual values. A discussion can be found in Amemiya.[29]

The application to the more general model, when all the coefficients are allowed to vary between the treatment and control groups, is approached the same way, except that there will be two equations to estimate, each with its own selectivity correction. Note that the selectivity regressor for the control group is of the same form as the one given before, that is, $\phi(\cdot)/\Phi(\cdot)$, except that the negative of the index is used:

$$\frac{\phi(-Z_i\gamma/\sigma_0)}{\Phi(-Z_i\gamma/\sigma_0)} = \frac{-\phi(Z_i\gamma/\sigma_0)}{1-\Phi(Z_i\gamma/\sigma_0)}$$

13.12.2 Some Cautionary Remarks about Selectivity Bias

The use of the simple Heckman correction or one of its many variants, has proliferated enormously. Many software packages now provide canned "heckman" procedures that allow the user to implement variants of Heckman's two-step method or its maximum likelihood equivalent without the need to resort to extensive computer programming.

At the same time, many econometricians and applied researchers have come to feel that indiscriminate use of these techniques should be avoided. H. Gregg Lewis, for example, in an influential survey of the effect of unions on wages, summarized his review of estimates computed using some type of selectivity bias corrections this way:

> I admire the ingenuity that has gone into development of [these techniques]. Yet in the present context the techniques are not working. I know little more about the magnitude of the selectivity bias in OLS wage gap estimates after completing the survey in this chapter than if I had ignored the... estimates reported here.[30]

[29] T. Amemiya, op. cit., Section 4.3.

[30] H. Gregg Lewis, *Union Relative Wage Effects: A Survey*, University of Chicago Press, 1986, p. 59.

Among other things, Lewis noted that estimates using these techniques seemed to exhibit much greater variability across studies than estimates produced by authors using generally simpler techniques. Heckman notes that part of the apparent variability is merely the consequence of misinterpreting the estimates generated by these models. Heckman also notes, however, that in many contexts, simpler estimation techniques (including instrumental variables) may perform as well in answering interesting economic questions as more complicated selectivity bias methods.[31]

Although a consensus on the value of selectivity bias methods and when their use is appropriate does not exist, a few remarks may be in order:

- In our examples, we have distinguished between *X*, the covariates that affect the outcome, and *Z*, the covariates (which may partially overlap with *X*) that determine whether or not the treatment is given. In principle the model is identified even when the variables in *X* and *Z* are the same. When this is the case, identification depends exclusively on the model and the normality assumption being exactly correct, assumptions which are almost certainly too thin a reed upon which to base inference.
- Though desirable, it is often difficult to find variables that affect the probability of receiving the treatment but also do not enter the wage equation. Gronau's case, where additional identification rests on the ability to exclude the presence of small children from the wage equation, probably represents the exception rather than the rule.
- Even with sufficient identifying information, the parameters of the model appear to be sensitive to the presence of heteroscedasticity, or departures from normality. In light of our discussion of heteroscedasticity in the Tobit, this may not come as a surprise. Some have suggested that it may be possible to make these two-step methods less sensitive to violations of some assumptions by combining nonparametric techniques and parametric techniques—for example, by including as additional regressors the squared value or higher powers of the inverse Mills ratio. Heckman reviews some more sophisticated semiparametric methods, but there is little empirical experience with these methods as well.[32]
- Finally, even if the model is correctly specified, the two-step approach may be very inefficient compared with the full-blown maximum likelihood counterpart. Davidson and MacKinnon, for example, recommend using the two-step procedure only to *test* for the presence of selectivity bias; if the null hypothesis of no selectivity bias is rejected they recommend using ML estimation provided it is not computationally prohibitive to do so.[33]

13.12.3 The Tobit as a Special Case

Note that the framework outlined at the beginning of this section subsumes many common estimation exercises. It may help to consider another special case. Let y

[31] See Heckman, op. cit., 1990, for a nice discussion of the issues involved.

[32] J. Heckman, ibid.

[33] R. Davidson and J. MacKinnon, op. cit., 545.

refer to consumption of cigarettes; let the vector of covariates include, say, the price of cigarettes; and let T be an indicator variable indicating whether the individual chooses to smoke. Formally,

$$y_i = X_i\boldsymbol{\beta}_1 + \epsilon_{1i} \tag{13.87}$$

$$y_{2i} = 0 \tag{13.88}$$

$$T_i = 1(Z_i\gamma + \epsilon_{0i} > 0) \tag{13.89}$$

$$y_i = T_i y_{1i} + (1 - T_i)y_{2i} \tag{13.90}$$

If we further specialize this model by assuming $X = Z, \gamma = \boldsymbol{\beta}_1$, and $\epsilon_{0i} \equiv \epsilon_{1i}$ (i.e., so that the selection equation and consumption equation are the same) we have the Tobit model! In this framework, however, it is much easier to see that the restrictions we need to impose on the system of equations in (13.87) to (13.90) to yield a conventional Tobit are not necessarily innocuous or "minor." Consider the elasticity of cigarette consumption with respect to price. Do changes in the price of cigarettes affect smokers and nonsmokers in exactly the same way? Perhaps they do, but unfortunately, the Tobit specification *assumes* they do.

Suppose the Tobit model is not correct because the covariates affect the participation decision differently from the decision of how much to consume conditional on consuming a positive quantity, as in Eqs. (13.87) to (13.90) with $\boldsymbol{\beta}_1 \neq \gamma$. It is most straightforward to estimate such a system by maximum likelihood.

Our discussion thus far suggests that one alternative is to treat the Tobit as a selectivity problem, and nonzero consumption as the "treatment." Consider the consequences of OLS regression on the nonzero values of y. We could write

$$E[y_i \mid X_i, T_i = 1] = X_i\boldsymbol{\beta} + E[\epsilon_{1i} \mid \epsilon_{0i} > -\boldsymbol{Z}_i\boldsymbol{\gamma}] \tag{13.91}$$

Again, if $\epsilon_{1i} = \epsilon_{0i}$, $\boldsymbol{\beta} = \gamma$, and $X = Z$, this equation is merely the standard Tobit. Note that we now have an expression that looks much like our example on the wages of working women. In fact, the problem has the same two-step solution. In the first step, a probit predicting whether or not the observation is "observed" is used to calculate the inverse Mills ratio. In the second step, an OLS regression of $\boldsymbol{y}$ on X and the inverse Mills ratio is run using only the nonzero observations.

Note, however, that in formulating the problem this way the interpretation has changed slightly. One can view the participation and consumption decisions as resulting from two different latent variables:

$$y_1^* = X\boldsymbol{\beta} + \boldsymbol{\epsilon}_1$$

$$y_0^* = Z\gamma + \boldsymbol{\epsilon}_0$$

$$y_i = \begin{cases} y_{1i}^* & \text{if } y_{0i}^* > 0 \\ \text{not observed} & \text{otherwise} \end{cases}$$

In this setup, we never observe y_{0i}^*—only whether it is positive or negative—and we only observe y_{1i}^* when y_{0i}^* is positive. One difference in interpretation arises from the fact that this alternative does not strictly limit the range of $\boldsymbol{y}$. For an illustration, see Hsiao.[34] In his case, y_1 is the observed amount of an insurance claim, and y_0 is whether or not a claim is filed. For more extensive discussion, again see Amemiya.[23]

The two-step method would appear to have little to commend itself compared to a full-blown ML estimation of the model apart from its simplicity. Experience suggests that the two-step method will be inefficient compared to its ML equivalent. On the other hand, when a specification test rejects the assumption that the process determining participation is the same one that determines consumption, and ML estimation is not practical, the two-step method would seem preferable to application of the simple Tobit.

13.13 READINGS

Our survey of this area has been brief and selective. Several other topics deserving mention are the multinomial logit and probit, disequilibrium models, and the truncated regression model. Other worthy topics include estimation methods based on the propensity score and hazard models. A good starting point for many of these models and those surveyed in this chapter is the book by Maddala.[1] See also the surveys by Amemiya on the Tobit models[23] and on qualitative response models.[35] A useful starting point for hazard models is the survey by Kiefer.[36] Some of the discussion in this chapter was based on work by Deaton who presents a lucid discussion of heteroscedasticity in binary choice models and related issues.[15]

PROBLEMS

1. Prove that the likelihood function of the logit is globally concave.

2. For the logit and linear probability models, show that the sum of the predicted probabilities equals the empirical sum of ones in the sample.

3. Show that the LR test that involves the comparison with the fully saturated model discussed at the end of 13.8.1 is distributed asymptotically χ^2. *Hint:* Recall that the sum of squared standard normal variates has a χ^2 distribution.

[34] C. Hsiao, "A Statistical Perspective on Insurance Rate-Making," *Journal of Econometrics*, **44**, 1990, 5–24.

[35] T. Amemiya, "Qualitative Response Models: A Survey," *Journal of Economic Literature*, **19**, 1981, 481–536.

[36] N. Kiefer, "Economic Duration Data and Hazard Functions," *Journal of Economic Literature*, **26**, 1986, 646–679.

4. Consider the following data for two dichotomous variables:

	x		
y	0	1	Total
0	40	60	100
1	60	40	100
Total	100	100	200

(*a*) Compute the regression coefficients and predicted values for the following linear probability model:

$$y = \beta_0 + \beta_1 x$$

(*b*) Consider running the same model, except this time using a probit. Using just the standard normal tables, compute the coefficients of the probit model and the predicted values for the model.

5. Consider estimating the standard probit model where $\text{prob}(y_i = 1) = \Phi[X_i(\beta/\sigma)]$ and y is a binary variable that takes on the value 0 or 1.

(*a*) How would the estimated coefficients compare if one ran a probit model on the same data, except that y has been recoded to take on a value of 0 or 10?

(*b*) Repeat the preceding question for the case when the model is a logit.

(*c*) Do the same for a linear probability model and discuss how the coefficients should be interpreted in this case.

6. The binary dependent variable models in this chapter are of the form $\text{prob}(y = 1) = F(X\boldsymbol{\beta})$. Describe the model that would result if F were the cumulative uniform distribution.

7. Consider the standard Tobit model, where y (the observed data) is censored below by 0 and

$$y^* = x\beta + \epsilon$$

Call the estimate of β, $\hat{\beta}$. Consider another estimate of β from the Tobit, replacing y with z, where

$$z_i \begin{cases} = y_i & \text{if } y_i > c \\ = c & \text{otherwise} \end{cases}$$

where $c > 0$. Informally compare the two estimators and suggest a specification test. In particular, comment on what happens to the second estimator as the value of c increases.

8. Using the data from the 1988 CPS on the data diskette, calculate the following linear regression for log wages:

$$\begin{aligned} \text{lwage} = {} & \beta_0 + \beta_1(\text{potential experience}) + \beta_2(\text{experience})^2 \\ & + \beta_3(\text{grade}) + \beta_4(\text{married}) + \beta_5(\text{high}) + \epsilon \end{aligned}$$

Next, generate a new variable, say, clwage such that

$$\text{clwage} = \begin{cases} \text{lwage} & \text{if lwage} > 1.87 \\ 0 & \text{otherwise} \end{cases}$$

Now perform a Tobit on the same model, replacing lwage with clwage. How do your estimates of the relevant coefficients compare? Try increasing the censoring point and see what happens.

APPENDIX A

Matrix Algebra

As far as possible we follow the convention of using boldface, lowercase type for vectors and boldface, uppercase type for matrices. The sequence of topics in this appendix attempts to mirror the order in which the topics appear in the main text to facilitate cross reference between them.

A.1 VECTORS

A vector is an ordered sequence of elements arranged in a row or column. In this book the elements are generally real numbers or symbols representing real numbers. As an example,

$$\boldsymbol{a} = \begin{bmatrix} 5 \\ 1 \\ 3 \end{bmatrix}$$

is a 3-element column vector, and $\boldsymbol{b} = [-2 \quad 0]$ is a 2-element row vector. **The order** of a vector is the number of elements in the vector. Changing the sequence of elements in a vector produces a different vector. Thus, permuting the elements in $\boldsymbol{a}$ would yield six different vectors. In general we will interpret the vector symbol as a column vector. Column vectors can be transformed into row vectors and vice versa by the operation of **transposition**. We will denote the operation by a prime, although some authors use a T superscript. Thus,

$$\boldsymbol{a}' = [5 \quad 1 \quad 3] \qquad \text{and} \qquad \boldsymbol{b}' = \begin{bmatrix} -2 \\ 0 \end{bmatrix}$$

Clearly, repetition of the operation will restore the original vector, so that $(\boldsymbol{a}')' = \boldsymbol{a}$.

A.1.1 Multiplication by a Scalar

Multiplication of a vector by a scalar simply means that each element in the vector is multiplied by the scalar. For example, $2b = [-4 \quad 0]$.

A.1.2 Addition and Subtraction

In this operation corresponding elements of the vectors are added or subtracted. This can only be done for vectors that (*i*) are all column vectors or all row vectors and (*ii*) are all of the same order. Clearly, one cannot add a row vector to a column vector, nor add a 3-element column vector to a 6-element column vector. To illustrate,

$$\begin{bmatrix} 6 \\ -3 \\ 4 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 7 \\ -1 \\ 7 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 6 \\ -3 \\ 4 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 5 \\ -5 \\ 1 \end{bmatrix}$$

For $\boldsymbol{a}' = [a_1\ a_2\ \cdots\ a_n]$ and $\boldsymbol{b}' = [b_1\ b_2\ \cdots\ b_n]$

$$\boldsymbol{c} = \boldsymbol{a} + \boldsymbol{b} = \begin{bmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \vdots \\ a_n + b_n \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} \tag{A.1}$$

A.1.3 Linear Combinations

Combining the two operations of scalar multiplication and vector addition expresses one vector as a linear combination of other vectors. For instance,

$$3\begin{bmatrix} 6 \\ -3 \\ 4 \end{bmatrix} + 2\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 20 \\ -5 \\ 18 \end{bmatrix}$$

In general

$$\boldsymbol{b} = \lambda_1 \boldsymbol{a}_1 + \lambda_2 \boldsymbol{a}_2 + \cdots + \lambda_k \boldsymbol{a}_k = \sum_{i=1}^{k} \lambda_i \boldsymbol{a}_i \tag{A.2}$$

defines a $\boldsymbol{b}$ vector as a linear combination of the $\boldsymbol{a}_i$ vectors with scalar weights λ_i.

A.1.4 Some Geometry

Vectors may be given a geometric as well as an algebraic interpretation. Consider a 2-element vector $\boldsymbol{a}' = [2 \quad 1]$. This may be pictured as a **directed line segment,** as shown in Fig. A.1. The arrow denoting the segment starts at the origin and ends

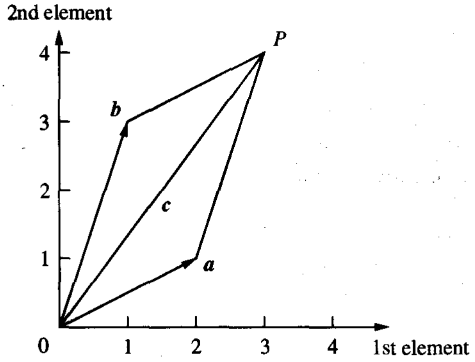


FIGURE A.1

at the point with coordinates (2, 1). The vector $\boldsymbol{a}$ may also be indicated by the point at which the arrow terminates. If we have another 2-element vector $\boldsymbol{b}' = [1 \quad 3]$, the geometry of vector addition is as follows. Start with $\boldsymbol{a}$ and then place the $\boldsymbol{b}$ vector at the terminal point of the $\boldsymbol{a}$ vector. This takes us to the point P in Fig. A.1. This point defines a vector $\boldsymbol{c}$ as the sum of the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, and it is obviously also reached by starting with the $\boldsymbol{b}$ vector and placing the $\boldsymbol{a}$ vector at its terminal point. The process is referred to as completing the parallelogram, or as the **parallelogram law** for the addition of vectors. The coordinates of P are (3, 4), and

$$\boldsymbol{c} = \boldsymbol{a} + \boldsymbol{b} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

so there is an exact correspondence between the geometric and algebraic treatments.

Now consider scalar multiplication of a vector. For example,

$$2\boldsymbol{a} = 2\begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

gives a vector in exactly the same direction as $\boldsymbol{a}$, but twice as long. Similarly,

$$-3\boldsymbol{a} = \begin{bmatrix} -6 \\ -3 \end{bmatrix}$$

gives a vector three times as long as $\boldsymbol{a}$, but going in the opposite direction. The three vectors are shown in Fig. A.2. All three terminal points lie on a single line through the origin, that line being uniquely defined by the vector $\boldsymbol{a}$.

In general

$$\lambda\boldsymbol{a}' = [\lambda a_1 \quad \lambda a_2 \quad \cdots \quad \lambda a_n] \tag{A.3}$$

It is clear from the parallelogram rule that any 2-element vector can be expressed as a *unique* linear combination of the $\boldsymbol{a}$ and $\boldsymbol{b}$ vectors in the preceding numerical example. For instance

$$\boldsymbol{d} = \begin{bmatrix} 4 \\ -3 \end{bmatrix} = 3\begin{bmatrix} 2 \\ 1 \end{bmatrix} - 2\begin{bmatrix} 1 \\ 3 \end{bmatrix} = 3\boldsymbol{a} - 2\boldsymbol{b}$$

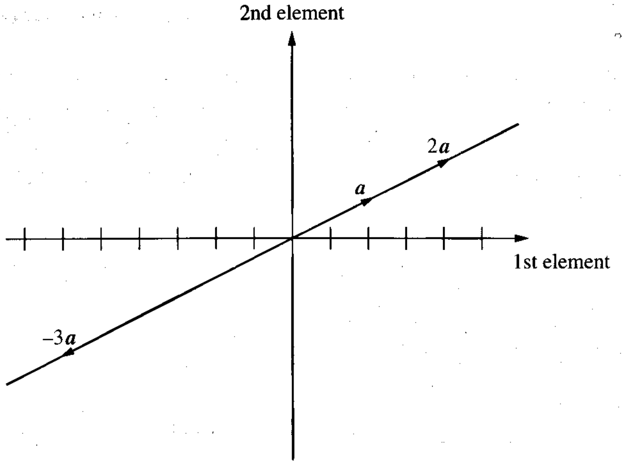


FIGURE A.2

A.1.5 Vector Multiplication

The **scalar, dot, or inner product** of two vectors is defined as

$$a'b = [a_1 \quad a_2 \quad \cdots \quad a_n]\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = a_1b_1 + a_2b_2 + \cdots + a_nb_n = \sum_{i=1}^{n} a_ib_i = b'a \tag{A.4}$$

The operation is only defined for vectors of the same order. Corresponding elements are multiplied together and summed to give the product, which is a scalar. A special case of Eq. (A.4) is the product of a vector by itself, which gives

$$a'a = \sum_{i=1}^{n} a_i^2$$

In the 2-element case this quantity is $(a_1^2 + a_2^2)$, which, by Pythagoras' Theorem, is the squared length of the vector a. The length of a vector is denoted by $\|a\|$. Extending through three and higher dimensions gives, in general, the length of a vector as

$$\|a\| = \sqrt{a'a} \tag{A.5}$$

where the positive square root is always taken.

The **outer product** of two n-element column vectors is ab', which is an $n \times n$ matrix, each element being the product of an element from a and an element from b.

A.1.6 Equality of Vectors

If two vectors of the same order are equal, they are equal element by element. The difference of the two vectors then gives the **zero vector,** in which every element is zero.

A.2 MATRICES

A matrix is a *rectangular array* of elements. The *order* of a matrix is given by the number of rows and the number of columns. In stating the order, the number of rows is always given first, and the number of columns second. Thus, a matrix A of order $m \times n$ appears as

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Clearly a column vector is a special case of a matrix, namely, a matrix of order, say, $m \times 1$, and a row vector is a matrix of order $1 \times n$. The $m \times n$ matrix may be regarded as an ordered collection of m-element column vectors or as an ordered collection of n-element row vectors. Multiplication of a matrix by a scalar means that each element in the matrix is multiplied by the scalar. The addition of two matrices of the same order, as with vectors, is achieved by adding corresponding elements.

The **transpose** of A is denoted by A'. The first row of A becomes the first column of the transpose, the second row of A becomes the second column of the transpose, and so on. The definition might equally well have been stated in terms of the first column of A becoming the first row of A', and so on. As an example

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 4 \end{bmatrix} \qquad A' = \begin{bmatrix} 1 & 2 \\ 2 & 0 \\ 3 & 4 \end{bmatrix}$$

A **symmetric matrix** satisfies

$$A' = A$$

that is, $$a_{ij} = a_{ji} \qquad \text{for} \qquad i \neq j$$

where a_{ij} is the element in A at the intersection of the ith row and jth column. This property can only hold for square matrices, $(m = n)$, since otherwise A and A' are not even of the same order. An example of a symmetric matrix is

$$A = \begin{bmatrix} 1 & -1 & 4 \\ -1 & 0 & 3 \\ 4 & 3 & 2 \end{bmatrix} = A'$$

From the definition of a transpose it follows that repetition of the operation returns the original matrix. It also follows directly that

$$(A + B)' = A' + B' \tag{A.6}$$

that is, the transpose of a sum is the sum of the transposes.

A.2.1 Matrix Multiplication

Matrix multiplication is achieved by repeated applications of vector multiplication. If A is of order $m \times n$ and B is of order $n \times p$, then a matrix $C = AB$ of order $m \times p$

can be found. The typical element c_{ij} of $\boldsymbol{C}$ is the inner product of the ith row of $\boldsymbol{A}$ and the jth column of $\boldsymbol{B}$, that is,

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj} \qquad i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, p \tag{A.7}$$

These inner products only exist if A has the same number of columns as $\boldsymbol{B}$ has rows. Thus the order in which matrices enter the product is of vital importance. When $p \neq m$ the inner products of rows of $\boldsymbol{B}$ and columns of A do not exist, and $\boldsymbol{BA}$ is not defined. The following example illustrates a case where both product matrices exist $(p = m)$:

$$\begin{aligned}
\boldsymbol{AB} &= \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 4 \end{bmatrix} \begin{bmatrix} 1 & 6 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1(1) + 2(0) + 3(1) & 1(6) + 2(1) + 3(1) \\ 2(1) + 0(0) + 4(1) & 2(6) + 0(1) + 4(1) \end{bmatrix} \\
&= \begin{bmatrix} 4 & 11 \\ 6 & 16 \end{bmatrix} \\
\boldsymbol{BA} &= \begin{bmatrix} 1 & 6 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 4 \end{bmatrix} \\
&= \begin{bmatrix} 1(1) + 6(2) & 1(2) + 6(0) & 1(3) + 6(4) \\ 0(1) + 1(2) & 0(2) + 1(0) & 0(3) + 1(4) \\ 1(1) + 1(2) & 1(2) + 1(0) & 1(3) + 1(4) \end{bmatrix} \\
&= \begin{bmatrix} 13 & 2 & 27 \\ 2 & 0 & 4 \\ 3 & 2 & 7 \end{bmatrix}
\end{aligned}$$

A.2.2 The Transpose of a Product

Let the product $\boldsymbol{AB}$ be denoted by

$$\boldsymbol{C} = \boldsymbol{AB} = \begin{bmatrix} \cdots & \boldsymbol{a}_1 & \cdots \\ & \vdots & \\ \cdots & \boldsymbol{a}_m & \cdots \end{bmatrix} \begin{bmatrix} \vdots & & \vdots \\ \boldsymbol{b}_1 & \cdots & \boldsymbol{b}_p \\ \vdots & & \vdots \end{bmatrix}$$

where $\boldsymbol{a}_j$ indicates the jth row of $\boldsymbol{A}$, and $\boldsymbol{b}_i$ the ith column of $\boldsymbol{B}$. Thus

$$c_{ji} = \boldsymbol{a}_j\boldsymbol{b}_i \qquad j = 1, \ldots, m; \quad i = 1, \ldots, p$$

denotes the jith element in $\boldsymbol{C}$. Transposition of $\boldsymbol{C}$ means that the jith element in $\boldsymbol{C}$ becomes the ijth element in $\boldsymbol{C}'$. Denoting this element by c'_{ij} gives

$$c'_{ij} = c_{ji} = \boldsymbol{a}_j\boldsymbol{b}_i$$

Referring to the definition of vector multiplication in Eq. (A.4), we see that $a_j b_i = b_i' a_j'$. Thus

$$c_{ij}' = b_i' a_j' = \text{inner product of the } i\text{th row of } B' \text{and the } j\text{th column of } A'$$

and so, from the definition of matrix multiplication,

$$C' = (AB)' = B'A' \tag{A.8}$$

The transpose of a product is the product of the transposes in reverse order. This rule extends directly to any number of conformable matrices. Thus,

$$(ABC)' = C'B'A' \tag{A.9}$$

The associative law of addition holds for matrices; that is,

$$(A + B) + C = A + (B + C) \tag{A.10}$$

This result is obvious since matrix addition merely involves adding corresponding elements, and it does not matter in what order the additions are performed.

We state, without proof, the associative law of multiplication, which is

$$(AB)C = A(BC) \tag{A.11}$$

A.2.3 Some Important Square Matrices

The **unit** or **identity matrix** of order $n \times n$ is

$$I_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

with ones down the principal diagonal and zeros everywhere else. This matrix plays a role similar to that of unity in scalar algebra. Premultiplying an n-vector, y, by I leaves the vector unchanged, that is, $Iy = y$. Transposing this last result gives $y'I = y'$, that is, postmultiplying a row vector by I leaves the row vector unchanged. For a matrix A of order $m \times n$ it follows that

$$I_m A = AI_n = A$$

Pre- or postmultiplication by I leaves the matrix unchanged. There is usually no need to indicate the order of the identity matrix explicitly as it will be obvious from the context. The identity matrix may be entered or suppressed at will in matrix multiplication. For instance,

$$y - Py = Iy - Py = My$$

where $M = I - P$.

A **diagonal matrix** is like the identity matrix in that all off-diagonal terms are zero, but now the terms on the principal diagonal are scalar elements, of which at least one is nonzero. The diagonal matrix may be written

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

or, more compactly, $\Lambda = \text{diag}\{\lambda_1 \; \lambda_2 \; \cdots \; \lambda_n\}$. Examples are

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

A special case of a diagonal matrix occurs when all the λ's are equal. This is termed a **scalar matrix** and may be written

$$\begin{bmatrix} \lambda & 0 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda \end{bmatrix} = \lambda I = I\lambda$$

A scalar multiplier may be placed before or after the matrix it multiplies.

Another important square matrix is an **idempotent matrix.** If A is idempotent, then

$$A = A^2 = A^3 = \cdots$$

that is, multiplying A by itself, however many times, simply reproduces the original matrix. An example of a symmetric idempotent matrix is

$$A = \frac{1}{6}\begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix}$$

as may be verified by multiplication.

A very useful transformation matrix is

$$A = I - \frac{1}{n}(ii')$$

where i is a column vector of n ones. The product ii' is a matrix of order $n \times n$, in which every element is one. Given a column vector of n observations on a variable Y,

$$\frac{1}{n}(ii')y = \frac{1}{n}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \bar{Y} \\ \bar{Y} \\ \vdots \\ \bar{Y} \end{bmatrix}$$

and so

$$Ay = \begin{bmatrix} Y_1 - \bar{Y} \\ Y_2 - \bar{Y} \\ \vdots \\ Y_n - \bar{Y} \end{bmatrix}$$

The matrix thus transforms raw data into deviation form. If the data series has zero mean, it is unaffected by the transformation. Finally we note that $\boldsymbol{Ai} = \mathbf{0}$.

Another important matrix, though not necessarily square, is the **null matrix 0** whose every element is zero. Obvious relations are

$$\boldsymbol{A} + \mathbf{0} = \boldsymbol{A} \qquad \text{and} \qquad \boldsymbol{A}\mathbf{0} = \mathbf{0}$$

Similarly we may encounter null row or column vectors.

An important property of a square matrix is the **trace**, which is the sum of the elements on the principal diagonal; that is,

$$\operatorname{tr}(A) = \sum_i a_{ii}$$

If the matrix A is of order $m \times n$ and $\boldsymbol{B}$ is of order $n \times m$ then $\boldsymbol{AB}$ and $\boldsymbol{BA}$ are both square matrices, and

$$\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$$

Repeated application gives

$$\operatorname{tr}(\boldsymbol{ABC}) = \operatorname{tr}(\boldsymbol{CAB}) = \operatorname{tr}(\boldsymbol{BCA}) \tag{A.12}$$

provided the products exist as square matrices.

A.2.4 Partitioned Matrices

A matrix may be partitioned into a set of submatrices by indicating subgroups of rows and/or columns. For example,

$$\boldsymbol{A} = \left[\begin{array}{ccc:c} 4 & 0 & 2 & -1 \\ 6 & 5 & 1 & 1 \\ \hdashline -3 & 2 & 0 & 5 \end{array}\right] = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \tag{A.13}$$

where

$$\boldsymbol{A}_{11} = \begin{bmatrix} 4 & 0 & 2 \\ 6 & 5 & 1 \end{bmatrix} \qquad \boldsymbol{A}_{12} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$\boldsymbol{A}_{21} = \begin{bmatrix} -3 & 2 & 0 \end{bmatrix} \qquad \boldsymbol{A}_{22} = 5 \tag{A.14}$$

The dashed lines indicate the partitioning, yielding the four submatrices defined in Eq. (A.14). The previous rules for the addition and multiplication of matrices apply directly to partitioned matrices *provided the submatrices are all of appropriate dimensions.* For instance, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are both written in partitioned form as

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix}$$

then

$$\boldsymbol{A} + \boldsymbol{B} = \begin{bmatrix} \boldsymbol{A}_{11} + \boldsymbol{B}_{11} & \boldsymbol{A}_{12} + \boldsymbol{B}_{12} \\ \boldsymbol{A}_{21} + \boldsymbol{B}_{21} & \boldsymbol{A}_{22} + \boldsymbol{B}_{22} \end{bmatrix}$$

provided $\boldsymbol{A}$ and $\boldsymbol{B}$ are of the same overall order (dimension) and each pair $\boldsymbol{A}_{ij}, \boldsymbol{B}_{ij}$ is of the same order. As an example of the multiplication of partitioned matrices,

$$AB = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

$$= \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \\ A_{31}B_{11} + A_{32}B_{21} & A_{31}B_{12} + A_{32}B_{22} \end{bmatrix}$$

For the multiplication to be possible and for these equations to hold, the number of columns in $\boldsymbol{A}$ must equal the number of rows in $\boldsymbol{B}$, and the same partitioning must be applied to the columns of $\boldsymbol{A}$ as to the rows of $\boldsymbol{B}$.

A.2.5 Matrix Differentiation

As seen in Chapter 3, OLS requires the determination of a vector $\boldsymbol{b}$ to minimize the residual sum of squares,

$$\boldsymbol{e'e} = \boldsymbol{y'y} - 2\boldsymbol{b'X'y} + \boldsymbol{b'X'Xb}$$

The first term on the right-hand side does not involve $\boldsymbol{b}$, whereas the second term is linear in $\boldsymbol{b}$ since $\boldsymbol{X'y}$ is a k-element column vector of known numbers, and the third term is a symmetric quadratic form in $\boldsymbol{b}$. To differentiate a linear function, write it as

$$f(\boldsymbol{b}) = \boldsymbol{a'b} = a_1b_1 + a_2b_2 + \cdots + a_kb_k = \boldsymbol{b'a}$$

where the a's are given constants. We may partially differentiate $f(\boldsymbol{b})$ with respect to each of the b_i. The resultant partial derivatives are arranged as a column vector,

$$\frac{\partial(\boldsymbol{a'b})}{\partial \boldsymbol{b}} = \frac{\partial(\boldsymbol{b'a})}{\partial \boldsymbol{b}} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} = \boldsymbol{a} \qquad \text{(A.15)}$$

These derivatives might equally well have been arranged as a row vector. The important requirement is consistency of treatment so that vectors and matrices of derivatives are of appropriate order for further manipulation. For the linear term in $\boldsymbol{e'e}$ it follows directly that

$$\frac{\partial(2\boldsymbol{b'X'y})}{\partial \boldsymbol{b}} = 2\boldsymbol{X'y}$$

which is a k-element vector.

The general quadratic form in $\boldsymbol{b}$ may be written $f(\boldsymbol{b}) = \boldsymbol{b'Ab}$, where the matrix $\boldsymbol{A}$ of known constants may be taken as symmetric. As a simple illustration consider

$$f(\boldsymbol{b}) = [b_1 \quad b_2 \quad b_3] \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$= a_{11}b_1^2 + a_{22}b_2^2 + a_{33}b_3^2 + 2a_{12}b_1b_2 + 2a_{13}b_1b_3 + 2a_{23}b_2b_3$$

The vector of partial derivatives is then

$$\frac{\partial f(\boldsymbol{b})}{\partial \boldsymbol{b}} = \begin{bmatrix} \dfrac{\partial f}{\partial b_1} \\ \dfrac{\partial f}{\partial b_2} \\ \dfrac{\partial f}{\partial b_3} \end{bmatrix} = \begin{bmatrix} 2(a_{11}b_1 + a_{12}b_2 + a_{13}b_3) \\ 2(a_{12}b_1 + a_{22}b_2 + a_{23}b_3) \\ 2(a_{13}b_1 + a_{23}b_2 + a_{33}b_3) \end{bmatrix} = 2\boldsymbol{Ab}$$

This result obviously holds for a symmetric quadratic form of any order; that is,

$$\frac{\partial(\boldsymbol{b}'\boldsymbol{Ab})}{\partial \boldsymbol{b}} = 2\boldsymbol{Ab} \tag{A.16}$$

for symmetric $\boldsymbol{A}$. Applying this result to the OLS case gives

$$\frac{\partial(\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Xb})}{\partial \boldsymbol{b}} = 2(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{b}$$

which is a k-element column vector.

A.2.6 Solution of Equations

The OLS coefficient vector, $\boldsymbol{b}$, is the solution of $(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{y}$. We need to establish the conditions under which a unique solution vector exists. Consider the set of equations

$$\boldsymbol{Ab} = \boldsymbol{c} \tag{A.17}$$

where $\boldsymbol{A}$ is a square, but not necessarily symmetric, matrix of order $k \times k$, and $\boldsymbol{b}$ and $\boldsymbol{c}$ are k-element column vectors. The elements of $\boldsymbol{A}$ and $\boldsymbol{c}$ are known and $\boldsymbol{b}$ is to be determined. The simplest case occurs when $k = 2$. The equations may then be written

$$b_1 \begin{bmatrix} \vdots \\ \boldsymbol{a}_1 \\ \vdots \end{bmatrix} + b_2 \begin{bmatrix} \vdots \\ \boldsymbol{a}_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \boldsymbol{c} \\ \vdots \end{bmatrix}$$

where the $\boldsymbol{a}_i$ $(i = 1, 2)$ are the 2-element column vectors in $\boldsymbol{A}$. If the type of situation illustrated in Fig. A.1 exists, it follows directly that there is a unique linear combination of the $\boldsymbol{a}_i$ that gives the $\boldsymbol{c}$ vector. However, if the situation pictured in Fig. A.2 obtains, where one column vector is simply a scalar multiple of the other, say, $\boldsymbol{a}_2 = \lambda\boldsymbol{a}_1$, then any linear combination of the two vectors can only produce another multiple of $\boldsymbol{a}_1$. Should the c vector also lie on the ray through $\boldsymbol{a}_1$, Eq. (A.17) will have an infinity of solutions, whereas if c lies elsewhere there will be no solution. The difference between (*i*) a unique solution and (*ii*) no solution or an infinity of solutions is that in the first case the column vectors are **linearly independent,** and in the second case they are **linearly dependent.** If the only solution to $\lambda_1\boldsymbol{a}_1 + \lambda_2\boldsymbol{a}_2 = \boldsymbol{0}$ is $\lambda_1 = \lambda_2 = 0$, the vectors are said to be linearly independent; otherwise they are linearly dependent.

The extension of this definition to higher dimensions is as follows. If the only solution to

$$\lambda_1 \boldsymbol{a}_1 + \lambda_2 \boldsymbol{a}_2 + \cdots + \lambda_k \boldsymbol{a}_k = \boldsymbol{0}$$

is $\lambda_1 = \lambda_2 = \cdots = \lambda_k = 0$, the k-element $\boldsymbol{a}_i$ vectors are linearly independent. Any k-element vector can then be expressed as a unique linear combination of these vectors, and so Eq. (A.17) has a unique solution vector, $\boldsymbol{b}$. This set of linearly independent vectors serves as a **basis** for the k-dimensional **vector space,** which contains all k-element vectors with real elements. The vector space is denoted by $\mathbf{E^k}$, the symbol for Euclidean space of dimension k. The sum of any two vectors in the space also lies in the space, the multiple of any vector in the space is also in the space, and distance in the space is measured as in Eq. (A.5). A basis is not unique. Any set of k linearly independent vectors will do. The basis vectors are said to **span** the space. A useful basis is the set of unit vectors. In $\mathbf{E}^3$ the unit vectors are

$$\boldsymbol{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad \boldsymbol{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \qquad \boldsymbol{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

so that any vector $\boldsymbol{c}' = [c_1 \quad c_2 \quad c_3]$ may be expressed as $\boldsymbol{c} = c_1\boldsymbol{e}_1 + c_2\boldsymbol{e}_2 + c_3\boldsymbol{e}_3$.

The angle θ between any two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ of the same order is defined by

$$\cos\theta = \frac{\boldsymbol{a}'\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\boldsymbol{a}}\sqrt{\boldsymbol{b}'\boldsymbol{b}}} \tag{A.18}$$

When $\theta = 90°$, $\cos\theta = 0$, and so $\boldsymbol{a}'\boldsymbol{b} = 0$. The two vectors intersect at a right angle, the inner product is zero, and the vectors are said to be **orthogonal.** Clearly the unit vectors are mutually orthogonal, and thus constitute an orthogonal basis for the space.

A.2.7 The Inverse Matrix

A related approach to the solution of Eq. (A.17) is via the inverse matrix. In scalar algebra the relation $ab = 1$ gives $b = a^{-1}$. In matrix algebra the question arises whether a similar type of relation exists. Specifically, if $\boldsymbol{A}$ is a square matrix, does there exist another square matrix, $\boldsymbol{B}$, such that $\boldsymbol{AB} = \boldsymbol{I}$? If the columns of $\boldsymbol{A}$ are linearly independent, the answer is yes. Letting $\boldsymbol{b}_1$ denote the first column of $\boldsymbol{B}$ gives the equation

$$\boldsymbol{A}\boldsymbol{b}_1 = \boldsymbol{e}_1 \tag{A.19}$$

where $\boldsymbol{e}_1' = [1 \quad 0 \quad 0 \quad \cdots \quad 0]$. Given linear independence of the columns of $\boldsymbol{A}$, $\boldsymbol{b}_1$ is uniquely determined. By the same argument each column of $\boldsymbol{B}$ is uniquely determined, and a matrix $\boldsymbol{B}$ exists, satisfying

$$\boldsymbol{AB} = \boldsymbol{I} \tag{A.20}$$

We state without proof the result that if the columns of the square matrix $\boldsymbol{A}$ are linearly independent, then so are the rows. Thus a similar argument shows that a square matrix $\boldsymbol{C}$ exists such that

$$\boldsymbol{CA} = \boldsymbol{I} \tag{A.21}$$

for each row of $\boldsymbol{C}$ is uniquely determined as the coefficients of a linear combination of the rows of $\boldsymbol{A}$. Putting Eqs. (A.20) and (A.21) together gives

$$\boldsymbol{C} = \boldsymbol{CI} = \boldsymbol{CAB} = \boldsymbol{IB} = \boldsymbol{B}$$

Thus, if the k columns (and rows) of $\boldsymbol{A}$ are linearly independent, a *unique* square matrix of order k exists, called the "inverse of $\boldsymbol{A}$" and denoted by $\boldsymbol{A}^{-1}$, with the property

$$\boldsymbol{AA}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I} \tag{A.22}$$

Premultiplying Eq. (A.17) by $\boldsymbol{A}^{-1}$ gives $\boldsymbol{b} = \boldsymbol{A}^{-1}\boldsymbol{c}$, which expresses the solution vector in terms of the known data. A matrix that has an inverse is said to be **nonsingular.** A matrix that has no inverse is **singular.**

A.2.8 The Rank of a Matrix

The rank of a matrix is defined as the maximum number of linearly independent columns (or rows) in the matrix. We will use the notation $\rho(\boldsymbol{A})$ to indicate the rank of $\boldsymbol{A}$. For any matrix the maximum number of linearly independent columns is always the same as the maximum number of linearly independent rows, so rank is a unique, unambiguous number. The rank of an $m \times n$ matrix must obviously satisfy

$$\text{Rank} \leq \min(m, n) \tag{A.23}$$

When the rank equals m ($< n$) the matrix is said to have **full row rank** and when the rank is n ($< m$) the matrix has **full column rank.** If all the columns (rows) of a square matrix are linearly independent, the matrix is said to have full rank.

EXAMPLE. Consider

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 1 \\ 2 & 2 & 4 & 5 \end{bmatrix}$$

The rank cannot be greater than three. Inspection, however, shows that the rows obey the relation, row 1 + row 2 − row 3 = zero vector. The rank therefore must be less than three. No row is a scalar multiple of another row, so the row rank of the matrix is two. Turning to the columns, we see there are four possible sets of three vectors, but in no set are the three columns linearly independent. The relations between them are

$$\begin{aligned} \text{col } 1 &= \text{col } 3 - \text{col } 2 \\ \text{col } 1 &= \text{col } 4 - 1.5 \text{ col } 2 \\ \text{col } 1 &= 3 \text{ col } 3 - 2 \text{ col } 4 \\ \text{col } 2 &= 2 \text{ col } 4 - 2 \text{ col } 3 \end{aligned}$$

No column is a scalar multiple of any other column, so the column rank of the matrix is two, as is the row rank.

Returning to the inverse matrix, we need to see more explicitly how such matrices are constructed and determine their properties. For a square matrix of order two,

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad A^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \tag{A.24}$$

Multiplication confirms that $AA^{-1} = A^{-1}A = I$. Since an inverse, if it exists, is unique, one can sometimes hazard a guess at an inverse and check by multiplication to see if it works. The common divisor in A^{-1} is a function of all the elements in A and is known as the **determinant** of A. It is a scalar quantity and denoted by det A or, alternatively, $|A|$. Sometimes one needs to take the absolute value of a determinant. This is written $|\det A|$. The determinant of the second-order matrix may be written

$$|A| = a_{11}a_{22} - a_{12}a_{21} = \sum_{\alpha,\beta} \pm a_{1\alpha}a_{2\beta} \tag{A.25}$$

The summation term gives the sum of all possible products of the elements of A, taken two at a time, with the first subscript in natural order 1, 2 and α, β indicating all possible permutations of 1, 2 for the second subscript, each product term being affixed with a positive (negative) sign as the number of inversions of the natural order in the second subscript is even (odd). There are only two possible permutations of 1, 2, namely 1, 2 itself and 2, 1. The latter permutation has only one inversion of the natural order, since 2 comes ahead of 1, which gives the explicit expression in Eq. (A.25).

The matrix in A^{-1} in Eq. (A.24) is a rearrangement of the elements of A. It is known as the **adjugate** or **adjoint matrix**, written (adj$A$). It is produced by the following two rules:

1. For each element in A, strike out the row and column containing that element and write down the remaining element prefixed with a positive or negative sign in the pattern

$$\begin{bmatrix} + & - \\ - & + \end{bmatrix}$$

   This gives the matrix $\begin{bmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{bmatrix}$

2. Transpose the latter matrix to get

$$\text{adj}A = \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

   Multiplication then gives

$$A(\text{adj}A) = (\text{adj}A)A = \begin{bmatrix} (a_{11}a_{22} - a_{12}a_{21}) & 0 \\ 0 & (a_{11}a_{22} - a_{12}a_{21}) \end{bmatrix} = \begin{bmatrix} |A| & 0 \\ 0 & |A| \end{bmatrix}$$

   and so
$$A^{-1} = \frac{1}{|A|}\text{adj}A \tag{A.26}$$

Turning to the 3 × 3 case, we have

$$|A| = \sum_{\alpha,\beta,\gamma} \pm a_{1\alpha}a_{2\beta}a_{3\gamma}$$
$$= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \tag{A.27}$$

The rules given for the determination of the adjoint matrix in the second-order case need modification for the third- and higher-order cases. Striking out the row and column containing, say, a_{11}, leaves the 2 × 2 submatrix,

$$\begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix}$$

rather than a scalar element. In fact, we replace a_{11} with the *determinant* of this submatrix. The amended rules are these:

1. Replace a_{ij} by $(-1)^{i+j}M_{ij}$ where M_{ij} is the determinant of the 2 × 2 submatrix obtained when row i and column j are deleted from $\boldsymbol{A}$. M_{ij} is termed a **minor,** and the signed minor is a **cofactor,** C_{ij}, that is

$$C_{ij} = (-1)^{i+j}M_{ij} \tag{A.28}$$

   The sign of M_{ij} does not change if $i + j$ is an even number and does change if that sum is odd.
2. Transpose this matrix of cofactors to obtain the adjoint matrix and divide each element by $|A|$ to produce the inverse matrix,

$$A^{-1} = \frac{1}{|A|}(\text{adj}\,A) = \frac{1}{|A|}\begin{bmatrix} C_{11} & C_{21} & C_{31} \\ C_{12} & C_{22} & C_{32} \\ C_{13} & C_{23} & C_{33} \end{bmatrix} \tag{A.29}$$

   It follows from Eq. (A.29) that

$$A(\text{adj}\,A) = (\text{adj}\,A)A = |A|I = \begin{bmatrix} |A| & 0 & 0 \\ 0 & |A| & 0 \\ 0 & 0 & |A| \end{bmatrix} \tag{A.30}$$

   This shows that there are various ways of expressing the determinant other than the general definition in Eq. (A.27). Equating the top left element in the matrices on the left and right of Eq. (A.30) gives

$$|A| = a_{11}C_{11} + a_{12}C_{12} + a_{13}C_{13} \tag{A.31}$$

   which defines $|A|$ as a linear combination of the elements in the first row, each element being multiplied by its cofactor. To illustrate this result, collect terms in Eq. (A.27) by the elements in the first row of A to obtain

$$|A| = a_{11}(a_{22}a_{33} - a_{23}a_{32}) + a_{12}(-a_{21}a_{33} + a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}) \tag{A.32}$$

   One may easily check that the terms in parentheses in Eq. (A.32) are the cofactors in Eq. (A.31). It is clear from Eq. (A.30) that the determinant may be expressed as a linear combination of the elements in any row (or column), provided the elements are multiplied by the corresponding cofactors. Note, however, that if the elements of any row (or column) are multiplied by the cofactors of a *different* row (or column), the result is zero. **Expansions in terms of alien cofactors vanish identically.**

EXAMPLE OF AN INVERSE MATRIX.

$$A = \begin{bmatrix} 1 & 3 & 4 \\ 1 & 2 & 1 \\ 2 & 4 & 5 \end{bmatrix}$$

The matrix of minors is

$$\begin{bmatrix} 6 & 3 & 0 \\ -1 & -3 & -2 \\ -5 & -3 & -1 \end{bmatrix}$$

Signing the minors and transposing gives the adjoint matrix

$$\operatorname{adj} A = \begin{bmatrix} 6 & 1 & -5 \\ -3 & -3 & 3 \\ 0 & 2 & -1 \end{bmatrix}$$

Expressing the determinant in terms of the elements of the first row of A gives

$$|A| = 1(6) + 3(-3) + 4(0) = -3$$

and so the inverse is

$$A^{-1} = \begin{bmatrix} -2 & -\frac{1}{3} & 1\frac{2}{3} \\ 1 & 1 & -1 \\ 0 & -\frac{2}{3} & \frac{1}{3} \end{bmatrix}$$

For the nth-order case the rules for obtaining the inverse are essentially those already stated for the third-order case. The determinant is

$$|A| = \sum_{\alpha,\beta,\ldots,\nu} \pm a_{1\alpha}a_{2\beta}\cdots a_{n\nu} \tag{A.33}$$

Alternatively, the expansion in terms of the ith row of A is

$$|A| = a_{i1}C_{i1} + a_{i2}C_{i2} + \cdots + a_{in}C_{in} \qquad i = 1, 2, \ldots, n \tag{A.34}$$

or, in terms of the jth column of A,

$$|A| = a_{1j}C_{1j} + a_{2j}C_{2j} + \cdots + a_{nj}C_{nj} \qquad j = 1, 2, \ldots, n \tag{A.35}$$

The cofactors are now the signed minors of matrices of order $n - 1$, and the inverse matrix is

$$A^{-1} = \frac{1}{|A|}\begin{bmatrix} C_{11} & C_{21} & \cdots & C_{n1} \\ C_{12} & C_{22} & \cdots & C_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ C_{1n} & C_{2n} & \cdots & C_{nn} \end{bmatrix} \tag{A.36}$$

A.2.9 Some Properties of Determinants

(*i*) *If a matrix **B** is formed from **A** by adding a multiple of one row (or column) to another row (or column), the determinant is unchanged.* Suppose the *i*th row of B is defined as the sum of the *i*th row of A and a multiple of the *j*th row of A; that is,

$$b_{ik} = a_{ik} + \lambda a_{jk} \qquad k = 1, 2, \ldots, n$$

Expanding in terms of the *i*th row,

$$|B| = \sum_k (a_{ik} + \lambda a_{jk})C_{ik} = \sum_k a_{ik}C_{ik} = |A|$$

where the cofactors of the *i*th row are obviously the same for each matrix. The result then follows since expansions in terms of alien cofactors vanish.

(*ii*) *If the rows (columns) of **A** are linearly dependent,* $|A| = 0$*, and if they are linearly independent,* $|A| \neq 0$. If row *i*, say, can be expressed as a linear combination of certain other rows, the rows of A are linearly dependent. Subtracting that linear combination from row *i* is simply a repeated application of Property (*i*), and so will leave the determinant unchanged. However, it produces a matrix with a row of zeros. Since each term in the determinantal expansion contains one element from any specific row, the determinant is zero. If the rows (columns) of A are linearly independent, there is no way to produce a zero row (column) and so $|A| \neq 0$. Thus, **nonsingular matrices have nonzero determinants** and **singular matrices have zero determinants.**

(*iii*) *The determinant of a triangular matrix is equal to the product of the diagonal elements.* A lower triangular matrix has zeros everywhere above the diagonal, as in

$$A = \begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & 0 & \cdots & 0 \\ a_{31} & a_{32} & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix}$$

Expanding in terms of the elements in the first row gives the determinant as the product of a_{11} and the determinant of the matrix of order $n - 1$ obtained by deleting the first row and column of A. Proceeding in this fashion gives

$$|A| = a_{11}a_{22}a_{33}\cdots a_{nn}$$

An upper triangular matrix has zeros everywhere below the diagonal and the same result obviously holds. Two special cases of this property follow directly:

The determinant of a diagonal matrix is the product of the diagonal elements.

The determinant of the unit (identity) matrix is one.

(*iv*) *Multiplying any row (column) of a matrix by a constant multiplies the determinant by the same constant. Multiplying a matrix of order n by a constant* *multiplies the determinant by that constant raised to the nth power.* This result follows from the determinantal expansion where each term is the product of n elements, one and only one from each row and column of the matrix.

(*v*) *The determinant of the product of two square matrices is the product of the determinants.*

$$|AB| = |A| \cdot |B|$$

A useful corollary is

$$|A^{-1}| = \frac{1}{|A|}$$

A.2.10 Properties of Inverse Matrices

We now state, mostly without proof, some of the main properties of inverse matrices.

(*i*) *The inverse of the inverse reproduces the original matrix.*

$$(A^{-1})^{-1} = A$$

From the definition of an inverse, $(A^{-1})(A^{-1})^{-1} = I$. Premultiplying by A gives the result.

(*ii*) *The inverse of the transpose equals the transpose of the inverse.*

$$(A')^{-1} = (A^{-1})'$$

Transposing $AA^{-1} = I$ gives $(A^{-1})'A' = I$. Postmultiplication by $(A')^{-1}$ yields the result.

(*iii*) *The inverse of an upper (lower) triangular matrix is also an upper (lower) triangular matrix.* We illustrate this result for a lower triangular 3 × 3 matrix:

$$A = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

By inspection one can see that three cofactors are zero, namely,

$$C_{21} = -\begin{vmatrix} 0 & 0 \\ a_{32} & a_{33} \end{vmatrix} \quad C_{31} = \begin{vmatrix} 0 & 0 \\ a_{22} & 0 \end{vmatrix} \quad C_{32} = -\begin{vmatrix} a_{11} & 0 \\ a_{21} & 0 \end{vmatrix}$$

Thus,

$$A^{-1} = \frac{1}{|A|}\begin{bmatrix} C_{11} & 0 & 0 \\ C_{12} & C_{22} & 0 \\ C_{13} & C_{23} & C_{33} \end{bmatrix}$$

(*iv*) *The inverse of a partitioned matrix may also be expressed in partitioned form.* If

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where A_{11} and A_{22} are square nonsingular matrices, then

$$A^{-1} = \begin{bmatrix} B_{11} & -B_{11}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}B_{11} & A_{22}^{-1} + A_{22}^{-1}A_{21}B_{11}A_{12}A_{22}^{-1} \end{bmatrix} \tag{A.37}$$

where $B_{11} = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$ or, alternatively,

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}B_{22}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}B_{22} \\ -B_{22}A_{21}A_{11}^{-1} & B_{22} \end{bmatrix} \tag{A.38}$$

where $B_{22} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$. The correctness of these results may be checked by multiplying out. These formulas are frequently used. The first form, Eq. (A.37), is the simpler for expressions involving the first row of the inverse. Conversely, the second form, Eq. (A.38), is more convenient for expressions involving the second row.

A very important special case of these results occurs when a data matrix is partitioned as $X = [X_1 \quad X_2]$. Then

$$X'X = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}$$

Substitution in the preceding formulae gives

$$B_{11} = \left[X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1\right]^{-1} = (X_1'M_2X_1)^{-1} \tag{A.39}$$

with
$$M_2 = I - X_2(X_2'X_2)^{-1}X_2' \tag{A.40}$$

A similar substitution, or simply interchanging the 1, 2 subscripts, gives

$$B_{22} = (X_2'M_1X_2)^{-1} \tag{A.41}$$

with
$$M_1 = I - X_1(X_1'X_1)^{-1}X_1' \tag{A.42}$$

The M_i are symmetric idempotent matrices. Premultiplication of any vector by M_i gives the residuals from the regression of that vector on X_i. Thus M_2X_1 gives the matrix of residuals when each of the variables in X_1 is regressed on X_2, and so forth. The OLS equations for y on X in partitioned form are

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}$$

Taking the first row, we have

$$\begin{aligned} b_1 &= \left[\,(X_1'M_2X_1)^{-1} - (X_1'M_2X_1)^{-1}X_1'X_2(X_2'X_2)^{-1}\,\right]\begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix} \\ &= (X_1'M_2X_1)^{-1}X_1'M_2y \end{aligned} \tag{A.43}$$

Similarly,
$$b_2 = (X_2'M_1X_2)^{-1}X_2'M_1y \tag{A.44}$$

These results provide an alternative look at OLS regression. Regressing y and X_1 on X_2 yields a vector of residuals, M_2y. and a matrix of residuals. M_2X_1. Regressing the former on the latter gives the b_1 coefficient vector in Eq. (A.43). There is a similar interpretation for the b_2 vector in Eq. (A.44).

A.2.11 More on Rank and the Solution of Equations

Consider the **homogeneous** equations

$$Ab = 0 \tag{A.45}$$

The elements of A are known constants, and $\boldsymbol{b}$ is an unknown solution vector. Clearly if A is square and nonsingular, the only solution is the zero vector, $\boldsymbol{b} = A^{-1}\mathbf{0} = \mathbf{0}$. A nonzero solution requires A to be singular. As an illustration, consider

$$\begin{aligned} a_{11}b_1 + a_{12}b_2 &= 0 \\ a_{21}b_1 + a_{22}b_2 &= 0 \end{aligned}$$

These equations give $\quad b_1 = -\dfrac{a_{12}}{a_{11}}b_2 \qquad b_1 = -\dfrac{a_{22}}{a_{21}}b_2$

For a nonzero solution we must have

$$\frac{a_{12}}{a_{11}} = \frac{a_{22}}{a_{21}}$$

that is, the determinant of A must be zero. A is then a singular matrix with rank of one. One row (column) is a multiple of the other row (column). The solution vector is a ray through the origin.

Now consider a rectangular system,

$$\begin{aligned} a_{11}b_1 + a_{12}b_2 + a_{13}b_3 &= 0 \\ a_{21}b_1 + a_{22}b_2 + a_{23}b_3 &= 0 \end{aligned}$$

The rank of A is at most two. If it is two, then A will have at least two linearly independent columns. Suppose that the first two columns are linearly independent. The equations then solve for b_1 and b_2 in terms of b_3, say, $b_1 = \lambda_1 b_3$, $b_2 = \lambda_2 b_3$. The solution vector may be written

$$\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ 1 \end{bmatrix} b_3$$

The scalar b_3 is arbitrary. Thus all solution vectors lie on a ray through the origin.

If the rank of A is one, then one row must be a multiple of the other. The equations then solve for, say, b_1 as a linear function of b_2 and b_3, which are arbitrary. Writing this as $b_1 = \lambda_2 b_2 + \lambda_3 b_3$, the solution vector is

$$\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} \lambda_2 \\ 1 \\ 0 \end{bmatrix} b_2 + \begin{bmatrix} \lambda_3 \\ 0 \\ 1 \end{bmatrix} b_3$$

All solution vectors thus lie in a two-dimensional **subspace** of $\mathbf{E}^3$.

The set of solutions to Eq. (A.45) constitutes a vector space called the **nullspace** of A. The dimension of this nullspace (the number of linearly independent vectors spanning the subspace) is called the **nullity.** All three examples satisfy the equation,

$$\text{Number of columns in } A = \text{rank of } A + \text{nullity} \tag{A.46}$$

This equation holds generally. Let A be of order $m \times n$ with rank r. Thus there is at least one set of r linearly independent rows and at least one set of r linearly independent columns. If necessary, rows and columns may be interchanged so that the first r rows and the first r columns are linearly independent. Partition A by the first r rows and columns as in

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where A_{11} is a square nonsingular matrix of order r, and A_{12} is of order $r \times (n - r)$. Dropping the last $m - r$ rows from Eq. (A.45) leaves

$$[A_{11} \quad A_{12}] \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \mathbf{0} \tag{A.47}$$

where b_1 contains r elements and b_2 the remaining $n - r$ elements. This is a set of r linearly independent equations in $n \geq r$ unknowns. Solving for b_1 gives

$$b_1 = -A_{11}^{-1} A_{12} b_2 \tag{A.48}$$

The b_2 subvector is arbitrary or "free" in the sense that the $n - r$ elements can be specified at will. For any such specification the subvector b_1 is determined by Eq. (A.48). The general solution vector to Eq. (A.47) is thus

$$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -A_{11}^{-1} A_{12} \\ I_{n-r} \end{bmatrix} b_2 \tag{A.49}$$

But any solution to Eq. (A.47) is also a solution to Eq. (A.45) since the rows discarded from Eq. (A.45) to reach Eq. (A.47) can all be expressed as linear combinations of the r independent rows in Eq. (A.47). Hence any solution that holds for the included rows also holds for the discarded rows. Thus Eq. (A.49) defines the general solution vector for Eq. (A.45). If b is a solution, then clearly λb is also a solution for arbitrary λ. If b_i and b_j are distinct solutions then $\lambda_i b_i + \lambda_j b_j$ is also a solution. Thus the solutions to Eq. (A.45) constitute a vector space, the nullspace of A. The dimension of the nullspace is determined from Eq. (A.49). The $n - r$ columns of the matrix in Eq. (A.49) are linearly independent since the columns of the submatrix I_{n-r} are necessarily independent. Thus, the dimension of the nullspace is $n - r$, which proves the relation in Eq. (A.46). One important application of this result occurs in the discussion of the identification of simultaneous equation models, namely, that if the rank of A is one less than the number of columns, the solution space is simply a ray through the origin.

Result (A.46) also yields simple proofs of some important theorems on the ranks of various matrices. In Chapter 3 we saw that a crucial matrix in the determination of the OLS vector is $X'X$ where X is the $n \times k$ data matrix. It is assumed that $n > k$. Suppose $\rho(X) = r$. The nullspace of X then has dimension $k - r$. If m denotes any vector in this nullspace,

$$Xm = \mathbf{0}$$

Premultiplying by X' gives $\qquad X'Xm = \mathbf{0}$

Thus, m also lies in the nullspace of $X'X$. Next let s be any vector in the nullspace of $X'X$ so that

$$X'Xs = \mathbf{0}$$

Premultiplying by s', we find

$$s'X'Xs = (Xs)'(Xs) = \mathbf{0}$$

Thus $\boldsymbol{Xs}$ is a vector with zero length and so must be the null vector; that is,

$$\boldsymbol{Xs} = \boldsymbol{0}$$

Thus, s lies in the nullspace of $\boldsymbol{X}$. Consequently, $\boldsymbol{X}$ and $\boldsymbol{X'X}$ have the same nullspace and hence the same nullity, $(k - r)$. They also have the same number of columns (k), and so by Eq. (A.46) have the same rank $r = k - (k - r)$. Thus

$$\rho(\boldsymbol{X}) = \rho(\boldsymbol{X'X}) \tag{A.50}$$

When $\boldsymbol{X}$ has linearly independent columns, its rank is k. Then $(\boldsymbol{X'X})$ has rank k and so is nonsingular with inverse $(\boldsymbol{X'X})^{-1}$, guaranteeing the uniqueness of the OLS vector.

Transposing a matrix does not change its rank; that is, $\rho(\boldsymbol{X}) = \rho(\boldsymbol{X'})$. Applying Eq. (A.50) gives

$$\rho(\boldsymbol{XX'}) = \rho(\boldsymbol{X'})$$

The general result is then

$$\rho(\boldsymbol{X}) = \rho(\boldsymbol{X'X}) = \rho(\boldsymbol{XX'}) \tag{A.51}$$

Notice that $\boldsymbol{XX'}$ is a square matrix of order n $(> k)$, so that even if $\boldsymbol{X}$ has full column rank, $\boldsymbol{XX'}$ is still singular.

Another important theorem on rank may be stated as follows. If $\boldsymbol{A}$ is a matrix of order $m \times n$ with rank r, and $\boldsymbol{P}$ and $\boldsymbol{Q}$ are square nonsingular matrices of order m and n, respectively, then

$$\rho(\boldsymbol{PA}) = \rho(\boldsymbol{AQ}) = \rho(\boldsymbol{PAQ}) = \rho(\boldsymbol{A}) \tag{A.52}$$

that is, premultiplication and/or postmultiplication of $\boldsymbol{A}$ by nonsingular matrices yields a matrix with the same rank as $\boldsymbol{A}$. This result may be established by the same methods as used for Eq. (A.51). Finally we state without proof a theorem for the general case of the multiplication of one rectangular matrix by another conformable rectangular matrix. Let $\boldsymbol{A}$ be $m \times n$ and $\boldsymbol{B}$ be $n \times s$. Then

$$\rho(\boldsymbol{AB}) \leq \min[\rho(\boldsymbol{A}), \rho(\boldsymbol{B})] \tag{A.53}$$

that is, the rank of the product is less than or equal to the smaller of the ranks of the constituent matrices. Again, a similar method of proof applies as in the previous two theorems.

A.2.12 Eigenvalues and Eigenvectors

Eigenvalues and eigenvectors occur in the solution of a special set of equations. Consider the set of first-order difference equations that appears in the discussion of VARs in Chapter 9, namely,

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{x}_{t-1} \tag{A.54}$$

where $\boldsymbol{x}_t$ is a $k \times 1$ vector of observations on a set of x variables at time t, and $\boldsymbol{A}$ is a $k \times k$ matrix of known numbers. By analogy with the treatment of the univariate case in Chapter 7 we postulate a solution vector for the multivariate case as

$$\boldsymbol{x}_t = \lambda^t \boldsymbol{c} \tag{A.55}$$

where λ is an unknown scalar and $\boldsymbol{c}$ is an unknown $k \times 1$ vector. If Eq. (A.55) is to be a solution for Eq. (A.54), substitution in Eq. (A.54) should give equality of the two sides. Making the substitution and dividing through by λ^{t-1} gives

$$\lambda \boldsymbol{c} = \boldsymbol{A}\boldsymbol{c}$$

or,

$$(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{c} = \boldsymbol{0} \tag{A.56}$$

The $\boldsymbol{c}$ vector thus lies in the nullspace of the matrix $\boldsymbol{A} - \lambda \boldsymbol{I}$. If this matrix is nonsingular, the *only* solution to Eq. (A.56) is the trivial $\boldsymbol{x} = \boldsymbol{0}$. A nontrivial solution requires the matrix to be singular or, in other words, to have a zero determinant, which gives

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0 \tag{A.57}$$

This condition gives the **characteristic equation** of the matrix $\boldsymbol{A}$. It is a polynomial of degree k in the unknown λ, which can be solved for the k roots. These λ's are the **eigenvalues** of $\boldsymbol{A}$. They are also known as **latent roots** or **characteristic roots**. Each λ_i may be substituted back in Eq. (A.56) and the corresponding $\boldsymbol{c}$ vector obtained. The $\boldsymbol{c}$ vectors are known as the **eigenvectors** of A. They are also known as **latent vectors** or **characteristic vectors.** Assembling all k solutions produces the matrix equation.

$$\boldsymbol{A}\begin{bmatrix} \vdots & \vdots & & \vdots \\ \boldsymbol{c}_1 & \boldsymbol{c}_2 & \cdots & \boldsymbol{c}_k \\ \vdots & \vdots & & \vdots \end{bmatrix} = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \lambda_1\boldsymbol{c}_1 & \lambda_2\boldsymbol{c}_2 & \cdots & \lambda_k\boldsymbol{c}_k \\ \vdots & \vdots & & \vdots \end{bmatrix}$$

$$= \begin{bmatrix} \vdots & \vdots & & \vdots \\ \boldsymbol{c}_1 & \boldsymbol{c}_2 & \cdots & \boldsymbol{c}_k \\ \vdots & \vdots & & \vdots \end{bmatrix}\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix}$$

which is written more compactly as

$$\boldsymbol{A}\boldsymbol{C} = \boldsymbol{C}\boldsymbol{\Lambda} \tag{A.58}$$

where $\boldsymbol{C}$ is the square matrix of eigenvectors and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. If we assume for the moment that $\boldsymbol{C}$ is nonsingular, it follows that

$$\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{C} = \boldsymbol{\Lambda} \tag{A.59}$$

and the matrix of eigenvectors serves to *diagonalize* the $\boldsymbol{A}$ matrix.

EXAMPLE. As a simple illustration consider

$$A = \begin{bmatrix} 1.3 & -0.1 \\ 0.8 & 0.4 \end{bmatrix}$$

The characteristic equation (A.57) is then

$$\begin{aligned} \begin{vmatrix} 1.3-\lambda & -0.1 \\ 0.8 & 0.4-\lambda \end{vmatrix} &= (1.3-\lambda)(0.4-\lambda) + 0.08 \\ &= \lambda^2 - 1.7\lambda + 0.6 \\ &= (\lambda - 1.2)(\lambda - 0.5) \\ &= 0 \end{aligned}$$

The eigenvalues are $\lambda_1 = 1.2$ and $\lambda_2 = 0.5$. Substituting the first eigenvalue in Eq. (A.56) gives

$$(A - \lambda_1 I)c_1 = \begin{bmatrix} 0.1 & -0.1 \\ 0.8 & -0.8 \end{bmatrix} \begin{bmatrix} c_{11} \\ c_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Thus, $c_{11} = c_{21}$. The eigenvector is determined up to a scale factor, and is any nonzero multiple of $c_1' = [1 \quad 1]$. Similarly, one can see that substituting the second eigenvalue in Eq. (A.56) gives $0.8c_{12} = 0.1c_{22}$. The second eigenvector is thus any nonzero multiple of $c_2' = [1 \quad 8]$. The matrix of eigenvectors may be written

$$C = \begin{bmatrix} 1 & 1 \\ 1 & 8 \end{bmatrix} \quad \text{with} \quad C^{-1} = \frac{1}{7}\begin{bmatrix} 8 & -1 \\ -1 & 1 \end{bmatrix}$$

and it is easy to check that

$$C^{-1}AC = \frac{1}{7}\begin{bmatrix} 8 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1.3 & -0.1 \\ 0.8 & 0.4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 1.2 & 0 \\ 0 & 0.5 \end{bmatrix}$$

which illustrates the diagonalization of A. The reader should check that any other arbitrary normalization of the eigenvectors leaves the eigenvalues unchanged.

A.2.13 Properties of Eigenvalues and Eigenvectors

In the properties to follow, A is a $k \times k$ matrix of real elements and rank k, Λ is a diagonal matrix of k eigenvalues, not necessarily all distinct, and C is a $k \times s$ ($s \leq k$) matrix, whose columns are the eigenvectors of A. Some properties apply generally to any real square matrix. Others depend on whether the matrix is symmetric or not. For such results we use (a) to refer to the nonsymmetric case and (b) to refer to the symmetric case. Some results are stated without proof. For others an outline of a proof is provided.

1(a). *The eigenvalues of a nonsymmetric matrix may be real or complex.*
1(b). *The eigenvalues of a symmetric matrix are all real.*

As an illustration, the matrix A, shown below, has characteristic equation $\lambda^2 + 1 = 0$, giving $\lambda = \pm i$, where $i = \sqrt{-1}$, and B has eigenvalues $\pm\sqrt{5}$:

$$A = \begin{bmatrix} 1 & -2 \\ 1 & -1 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & -2 \\ -2 & -1 \end{bmatrix}$$

2(a). *If all k eigenvalues are distinct,* **C** *will have k linearly independent columns and so, as just shown,*

$$C^{-1}AC = \Lambda \qquad \text{or} \qquad A = C\Lambda C^{-1} \tag{A.60}$$

The method of proof may be sketched for the $k = 2$ case. Assume the contrary result, that is, that the two eigenvectors are linearly dependent, so that one may write

$$b_1c_1 + b_2c_2 = \mathbf{0}$$

for some scalars b_1 and b_2, of which at least one is nonzero. Premultiply this linear combination by A to obtain

$$b_1Ac_1 + b_2Ac_2 = (b_1\lambda_1)c_1 + (b_2\lambda_2)c_2 = \mathbf{0}$$

Multiplying the linear combination by λ_1 gives

$$(b_1\lambda_1)\boldsymbol{c}_1 + (b_2\lambda_1)\boldsymbol{c}_2 = \mathbf{0}$$

Subtracting from the previous equation, we find

$$(\lambda_2 - \lambda_1)b_2\boldsymbol{c}_2 = \mathbf{0}$$

The eigenvalues are different by assumption, and $\boldsymbol{c}_2$, being an eigenvector, is not the null vector. Thus, $b_2 = 0$. Similarly, it may be shown that $b_1 = 0$, and so a contradiction is forced. Thus distinct eigenvalues generate linearly independent eigenvectors.

2(*b*). The proof in 2(*a*) did not involve the symmetry of A or the lack of it. Thus the diagonalization in Eq. (A.60) applies equally well to symmetric matrices. However, *when A is symmetric, the eigenvectors are not just linearly independent; they are also* ***pairwise orthogonal.***

Consider the first two eigenvectors as in

$$\boldsymbol{Ac}_1 = \lambda_1\boldsymbol{c}_1 \qquad \text{and} \qquad \boldsymbol{Ac}_2 = \lambda_2\boldsymbol{c}_2$$

Premultiplying the first equation by $\boldsymbol{c}_2'$ and the second by $\boldsymbol{c}_1'$ gives

$$\boldsymbol{c}_2'\boldsymbol{Ac}_1 = \lambda_1\boldsymbol{c}_2'\boldsymbol{c}_1 \qquad \text{and} \qquad \boldsymbol{c}_1'\boldsymbol{Ac}_2 = \lambda_2\boldsymbol{c}_1'\boldsymbol{c}_2$$

Transposing the second equation gives $\boldsymbol{c}_2'\boldsymbol{Ac}_1 = \lambda_2\boldsymbol{c}_2'\boldsymbol{c}_1$, provided A is symmetric. Thus,

$$\lambda_1\boldsymbol{c}_2'\boldsymbol{c}_1 = \lambda_2\boldsymbol{c}_2'\boldsymbol{c}_1$$

Since the eigenvalues are distinct, $\boldsymbol{c}_2'\boldsymbol{c}_1 = 0$. This result holds for any pair of eigenvectors and so they are pairwise orthogonal when $\boldsymbol{A}$ is symmetric. It is also customary in this case to normalize the eigenvectors to have unit length, $\|\boldsymbol{c}_i\| = 1$, for $i = 1, 2, \ldots, k$. Let $\boldsymbol{Q}$ denote the matrix whose columns are these normalized orthogonal eigenvectors. Then

$$\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I} \tag{A.61}$$

From the definition and uniqueness of the inverse matrix it follows that

$$\boldsymbol{Q}' = \boldsymbol{Q}^{-1} \tag{A.62}$$

The matrix $\boldsymbol{Q}$ is called an **orthogonal matrix,** that is, a matrix such that *its inverse is simply its transpose.* It follows directly from (A.62) that

$$\boldsymbol{QQ}' = \boldsymbol{I} \tag{A.63}$$

that is, although $\boldsymbol{Q}$ was constructed as a matrix with orthogonal columns, its row vectors are also orthogonal. An orthogonal matrix is thus defined by

$$\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{QQ}' = \boldsymbol{I} \tag{A.64}$$

For symmetric matrices the diagonalization may be written

$$\boldsymbol{Q}'\boldsymbol{AQ} = \boldsymbol{\Lambda} \qquad \text{or} \qquad \boldsymbol{A} = \boldsymbol{Q\Lambda Q}' \tag{A.65}$$

3(*a*). *When the eigenvalues are not all distinct there are usually fewer than k linearly independent eigenvectors.*

As an example, consider

$$A = \begin{bmatrix} 3 & -2 \\ 0.5 & 1 \end{bmatrix}$$

The eigenvalues are $\lambda_1 = \lambda_2 = 2$, that is, a single root with multiplicity two. Substitution in Eq. (A.56) gives

$$\begin{bmatrix} 1 & -2 \\ 0.5 & -1 \end{bmatrix} \begin{bmatrix} c_{11} \\ c_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This yields just a single eigenvector, $\boldsymbol{c}_1' = [2 \quad 1]$. The diagonalization in Eq. (A.60) is then impossible. However, it is possible to get close to diagonalization in the form of a **Jordan matrix.** In this example the Jordan matrix is

$$\boldsymbol{J} = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$$

It is seen to be upper triangular with the (repeated) eigenvalue displayed on the principal diagonal and the number 1 above the principal diagonal. There exists a nonsingular matrix $\boldsymbol{P}$ such that

$$\boldsymbol{P}^{-1}\boldsymbol{AP} = \boldsymbol{J} \qquad \text{or} \qquad \boldsymbol{A} = \boldsymbol{PJP}^{-1} \tag{A.66}$$

To find the $\boldsymbol{P}$ matrix in this example, rewrite Eq. (A.66) as $\boldsymbol{AP} = \boldsymbol{PJ}$; that is,

$$\boldsymbol{A}\begin{bmatrix} \vdots & \vdots \\ \boldsymbol{p}_1 & \boldsymbol{p}_2 \\ \vdots & \vdots \end{bmatrix} = \begin{bmatrix} \vdots & \vdots \\ \boldsymbol{p}_1 & \boldsymbol{p}_2 \\ \vdots & \vdots \end{bmatrix}\begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix} = [\lambda\boldsymbol{p}_1 \quad \boldsymbol{p}_1 + \lambda\boldsymbol{p}_2]$$

Thus,

$$\begin{aligned} \boldsymbol{Ap}_1 &= \lambda\boldsymbol{p}_1 \\ \boldsymbol{Ap}_2 &= \boldsymbol{p}_1 + \lambda\boldsymbol{p}_2 \end{aligned}$$

The first equation shows that $\boldsymbol{p}_1$ is the eigenvector $\boldsymbol{c}_1$, which has already been obtained. Substituting $\lambda = 2$ and $\boldsymbol{p}_1' = [2 \quad 1]$ gives $\boldsymbol{p}_2' = [4 \quad 1]$. The $\boldsymbol{P}$ matrix is then

$$\boldsymbol{P} = \begin{bmatrix} 2 & 4 \\ 1 & 1 \end{bmatrix}$$

where each column has been normalized by setting the second element at 1. Some arithmetic shows that these matrices satisfy Eq. (A.66).

In the general case where $\boldsymbol{A}$ has s ($\leq k$) independent eigenvectors, the Jordan matrix is block diagonal

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{J}_1 & & \\ & \ddots & \\ & & \boldsymbol{J}_s \end{bmatrix}$$

Each block relates to a single eigenvalue and the associated eigenvector. If an eigenvalue has multiplicity m, the corresponding block has that eigenvalue repeated m times on the principal diagonal and a series of 1s on the diagonal above the principal diagonal. All other elements are zero. If the eigenvalue is distinct, the block reduces to a scalar showing the eigenvalue. For example, if $k = 4$ and there are just two eigenvalues, one with multiplicity three, the Jordan matrix is

$$J = \begin{bmatrix} J_1 & \\ & J_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_2 \end{bmatrix}$$

When one or more roots are repeated, a nonsingular matrix P can always be found to satisfy Eq. (A.66). The equations in (A.66) are perfectly general and not just applicable to the $k = 2$ case. If Λ is a diagonal matrix of order k containing all the eigenvalues, including repeats, one may easily see that

$$\text{tr}(\Lambda) = \text{tr}(J) \qquad \text{and} \qquad |\Lambda| = |J| \tag{A.67}$$

3(b). *When A is symmetric, the same result, Eq. (A.65), holds for repeated eigenvalues as for distinct eigenvalues.*

The reason is that a root with multiplicity m has m orthogonal vectors associated with it.[1] As an illustration consider the matrix

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

The characteristic equation is $(1 - \lambda)^2(2 - \lambda) = 0$, with eigenvalues $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = 2$. For λ_3, $(A - \lambda I)c = \mathbf{0}$ gives

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} c_{13} \\ c_{23} \\ c_{33} \end{bmatrix} = \mathbf{0}$$

The first two elements in c_3 are thus zero, and the third element is arbitrary. So the eigenvector is any nonzero multiple of $e_3' = [0 \quad 0 \quad 1]$. The multiple eigenvalue gives

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} c = \mathbf{0}$$

The third element in c must be zero, but the other two elements are arbitrary. Denoting the arbitrary scalars by b_1 and b_2, we may write

$$\begin{bmatrix} b_1 \\ b_2 \\ 0 \end{bmatrix} = b_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + b_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

The eigenvalue with multiplicity 2 thus yields two orthogonal eigenvectors, e_1 and e_2. It is also seen that all three eigenvectors are mutually orthogonal. The diagonalization in Eq. (A.65) thus holds for all real symmetric matrices, whether or not the eigenvalues are distinct.

[1] For a proof see G. Hadley, *Linear Algebra*, Addison-Wesley, 1961, 243–245.

4. *The sum of all the eigenvalues is equal to the trace of* ***A***.

From Eq. (A.59) we may write

$$\text{tr}(\Lambda) = \text{tr}(C^{-1}AC) = \text{tr}(ACC^{-1}) = \text{tr}(A) \tag{A.68}$$

The same method of proof works directly for symmetric matrices in Eq. (A.65), and also for Eq. (A.66) since we saw in Eq. (A.67) that $\text{tr}(J) = \text{tr}(\Lambda)$.

5. *The product of the eigenvalues is equal to the determinant of* ***A***.

From property (v) of determinants,

$$|\Lambda| = |C^{-1}AC| = |C^{-1}||A||C| = |A| \tag{A.69}$$

The same method of proof works for the other two cases in Eqs. (A.65) and (A.66), noting that $|\Lambda| = |J|$, as shown in Eq. (A.67).

6. *The rank of* ***A*** *is equal to the number of nonzero eigenvalues.*

It was established in Eq. (A.52) that premultiplication and/or postmultiplication of a matrix by nonsingular matrices leaves the rank of the matrix unchanged. Thus, in the first two diagonalizations, Eqs. (A.59) and (A.65),

$$\rho(A) = \rho(\Lambda) \tag{A.70}$$

The rank of Λ is the order of the largest nonvanishing determinant that can be formed from its diagonal elements. This is simply equal to the number of nonvanishing eigenvalues. It also follows that the rank of $\boldsymbol{J}$ is equal to the rank of Λ, and so the result holds for all three cases.

7. *The eigenvalues of* $\boldsymbol{\Pi} = \boldsymbol{I} - \boldsymbol{A}$ *are the complements of the eigenvalues of* ***A***, *but the eigenvectors of the two matrices are the same.*

An eigenvalue and associated eigenvector for $\boldsymbol{A}$ are given by

$$\boldsymbol{Ac} = \lambda\boldsymbol{c}$$

Subtracting each side from $\boldsymbol{c}$ gives

$$\boldsymbol{c} - \boldsymbol{Ac} = \boldsymbol{c} - \lambda\boldsymbol{c}$$

that is,

$$(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{c} = (1 - \lambda)\boldsymbol{c} \tag{A.71}$$

which establishes the result.

8. *The eigenvalues of* $\boldsymbol{A}^2$ *are the squares of the eigenvalues of* ***A***, *but the eigenvectors of both matrices are the same.*

Premultiplying $\boldsymbol{Ac} = \lambda\boldsymbol{c}$ by $\boldsymbol{A}$ gives

$$\boldsymbol{A}^2\boldsymbol{c} = \lambda\boldsymbol{Ac} = \lambda^2\boldsymbol{c} \tag{A.72}$$

which establishes the result.

9. *The eigenvalues of* $\boldsymbol{A}^{-1}$ *are the reciprocals of the eigenvalues of* ***A***, *but the eigenvectors of both matrices are the same.*

From $\boldsymbol{Ac} = \lambda\boldsymbol{c}$, we write

$$\boldsymbol{c} = \lambda\boldsymbol{A}^{-1}\boldsymbol{c}$$

giving
$$\boldsymbol{A}^{-1}\boldsymbol{c} = \left(\frac{1}{\lambda}\right)\boldsymbol{c} \tag{A.73}$$

10. *Each eigenvalue of an idempotent matrix is either zero or one.*

From Eq. (A.72)

$$\boldsymbol{A}^2\boldsymbol{c} = \lambda^2\boldsymbol{c}$$

When $\boldsymbol{A}$ is idempotent, $\quad \boldsymbol{A}^2\boldsymbol{c} = \boldsymbol{Ac} = \lambda\boldsymbol{c}$

Thus, $\quad \lambda(\lambda - 1)\boldsymbol{c} = \boldsymbol{0}$

and since any eigenvector $\boldsymbol{c}$ is not the null vector,

$$\lambda = 0 \qquad \text{or} \qquad \lambda = 1$$

11. *The rank of an idempotent matrix is equal to its trace.*

$$\begin{aligned}\rho(\boldsymbol{A}) &= \rho(\boldsymbol{\Lambda}) \\ &= \text{number of nonzero eigenvalues} \\ &= \text{tr}(\boldsymbol{\Lambda}) \\ &= \text{tr}(\boldsymbol{A})\end{aligned}$$

The first step comes from Eq. (A.70), the second from property 6, the third from property 10, and the last from Eq. (A.68).

A.2.14 Quadratic Forms and Positive Definite Matrices

A simple example of a quadratic form was given in the treatment of vector differentiation earlier in this appendix. In this section $\boldsymbol{A}$ denotes a real symmetric matrix of order $k \times k$. A quadratic form is defined as

$$q = \boldsymbol{b}'\boldsymbol{A}\boldsymbol{b}$$

where q is a scalar and $\boldsymbol{b}$ is a nonnull $k \times 1$ vector. The quadratic form and matrix are said to be **positive definite** if q is strictly positive for any nonzero $\boldsymbol{b}$. The form and matrix are **positive semidefinite** if q is nonnegative. There is an intimate link between the nature of the quadratic form and the eigenvalues of $\boldsymbol{A}$.

1. *A necessary and sufficient condition for the real symmetric matrix $\boldsymbol{A}$ to be positive definite is that all the eigenvalues of $\boldsymbol{A}$ be positive.* To prove the necessary condition, assume $\boldsymbol{b}'\boldsymbol{A}\boldsymbol{b} > 0$. For any eigenvalue and corresponding eigenvector, $\boldsymbol{Ac} = \lambda\boldsymbol{c}$. Premultiplying by $\boldsymbol{c}'$ gives

$$\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} = \lambda\boldsymbol{c}'\boldsymbol{c} = \lambda$$

since the eigenvectors can be given unit length. Positive definiteness thus implies positive eigenvalues. To prove sufficiency, assume all eigenvalues to be positive.

From Eq. (A.65),

$$A = C\Lambda C'$$

where C is an orthogonal matrix of eigenvectors. For any nonnull vector b,

$$\begin{aligned} b'Ab &= b'C\Lambda C'b \\ &= d'\Lambda d \\ &= \sum \lambda_i d_i^2 \end{aligned}$$

where $d = C'b$. Because C is nonsingular, the d vector is nonnull. Thus $b'Ab > 0$, which proves the result.

2. *If A is symmetric and positive definite, a nonsingular matrix P can be found such that $A = PP'$.* When all eigenvalues are positive, Λ may be factored into

$$\Lambda = \Lambda^{1/2}\Lambda^{1/2}$$

where $$\Lambda^{1/2} = \begin{bmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_k} \end{bmatrix}$$

Then $$A = C\Lambda C' = C\Lambda^{1/2}\Lambda^{1/2}C' = (C\Lambda^{1/2})(C\Lambda^{1/2})'$$

which gives the result with $P = C\Lambda^{1/2}$.

3. *If A is positive definite and* B *is $s \times k$ with $\rho(B) = k$, then $B'AB$ is positive definite.* For any nonnull vector d

$$d'(B'AB)d = (Bd)'A(Bd)$$

The vector Bd is a linear combination of the columns of B and cannot be null since the columns of B are linearly independent. Setting $A = I$ gives $B'B$ as a positive definite matrix. In least-squares analysis the data matrix X is conventionally of order $n \times k$ with rank k. Thus $X'X$, the matrix of sums of squares and cross products, is positive definite. Dividing by n gives the sample variance-covariance matrix, which is thus positive definite. This result also holds for population or theoretical variance-covariance matrices provided there is no linear dependence between the variables.

APPENDIX B

Statistics

It is assumed that the reader has had at least an elementary course in statistics, covering the basic principles of estimation and hypothesis testing. The purpose of this appendix is to highlight some of the more important theoretical results and, in particular, to provide a matrix treatment of relevant concepts and theorems.

B.1 RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

We begin with the *univariate* case. A random variable has a designated set of possible values and associated probabilities. A discrete random variable X consists of a set of possible values $x_1, x_2, \ldots, x_k$ and associated nonnegative fractions (probabilities) $p_1, p_2, \ldots, p_k$ such that

$$\sum_{i=1}^{k} p_i = 1$$

The two most important features of the probability distribution are the mean and variance. The mean, or expected value, is usually denoted by μ and is defined as

$$\mu = E(X) = \sum_{i=1}^{k} x_i p_i \tag{B.1}$$

which is just a weighted average of the x values, the weights being the respective probabilities. E is the **expectation operator,** and it may also be applied to various functions of X. For example, $E(X^2)$ indicates the expected value of X^2. The possible values of X^2 are $x_1^2, x_2^2, \ldots, x_k^2$, which occur with probabilities $p_1, p_2, \ldots, p_k$. Thus

$$E(X^2) = \sum_{i=1}^{k} x_i^2 p_i$$

The variance, usually indicated by σ^2, is the *expected squared deviation about the mean*. Thus

$$\sigma^2 = E[(X - \mu)^2] \tag{B.2}$$

Evaluating this from first principles, we have

$$\begin{aligned} E[(X - \mu)^2] &= \sum (x_i - \mu)^2 p_i \\ &= \sum x_i^2 p_i - 2\mu \sum x_i p_i + \mu^2 \sum p_i \\ &= \sum x_i^2 p_i - \left(\sum x_i p_i\right)^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

This result may also be obtained by first expanding the squared term in Eq. (B.2) and then applying the expectation operator to each term in turn. Thus

$$\begin{aligned} E[(X - \mu)^2] &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + E(\mu^2) \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

since $E(\mu^2)$ indicates the expectation of a constant, which is simply the constant.

When the random variable is continuous, the discrete probabilities are replaced by a continuous **probability density function** (pdf), usually denoted by $p(x)$ or $f(x)$. The pdf has the properties that

$$f(x) \geq 0 \qquad \text{for all } x$$

$$\int f(x)\,dx = 1$$

and

$$\int_a^b f(x)\,dx = \text{prob}[a < x < b]$$

This probability is shown in Fig. B.1. The mean and variance are defined as before, but integrals now replace summation signs. Thus

$$\mu = \int x f(x)\,dx \tag{B.3}$$

and

$$\sigma^2 = \int (x - \mu)^2 f(x)\,dx \tag{B.4}$$

B.2 THE UNIVARIATE NORMAL PROBABILITY DISTRIBUTION

The pdf for the univariate normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x - \mu)^2\right] \tag{B.5}$$

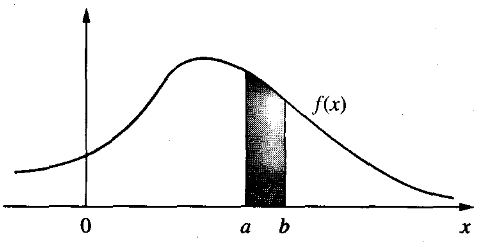


FIGURE B.1

This defines a two-parameter family of distributions, the parameters being the mean μ and the variance σ^2. The bell-shaped curve reaches its maximum at $x = \mu$ and is symmetrical around that point. A special member of the family is the **standard normal distribution,** which has zero mean and unit variance. An area under any specific normal distribution may be expressed as an equivalent area under the standard distribution by defining

$$z = \frac{x - \mu}{\sigma}$$

Clearly $E(z) = 0$ and $\mathrm{var}(z) = 1$, so that

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \tag{B.6}$$

Then

$$\int_{x_1}^{x_2} f(x)\,dx = \int_{z_1}^{z_2} f(z)\,dz$$

where $z_i = (x_i - \mu)/\sigma$. The areas under the standard normal distribution are tabulated in Appendix D.

B.3 BIVARIATE DISTRIBUTIONS

We are often interested in the *joint variation* of a pair of random variables. Let the variables X, Y have a bivariate pdf denoted by $f(x, y)$. Then

$$f(x, y) \geq 0 \qquad \text{for all } x, y$$

$$\iint f(x, y)\,dx\,dy = 1$$

and

$$\int_c^d \int_a^b f(x, y)\,dx\,dy = \mathrm{prob}[a < x < b, c < y < d]$$

Given the joint density, a **marginal density** is obtained for each variable by integrating over the range of the other variable. Thus

$$\text{Marginal pdf for } X = \int_{-\infty}^{\infty} f(x, y)\,dy = f(x) \tag{B.7}$$

and $$\text{Marginal pdf for } Y = \int_{-\infty}^{\infty} f(x, y)\,dx = f(y) \tag{B.8}$$

A **conditional density** for Y, given X, is defined as

$$f(y \mid x) = \frac{f(x, y)}{f(x)} \tag{B.9}$$

and similarly, a conditional pdf for X, given Y, is defined as

$$f(x \mid y) = \frac{f(x, y)}{f(y)} \tag{B.10}$$

Two variables are said to be *statistically independent*, or *independently distributed*, if the marginal and conditional densities are the same. In this case the joint density can be written as the product of the marginal densities,

$$f(x, y) = f(x) \cdot f(y) \tag{B.11}$$

Returning to the general bivariate distribution, we may obtain the mean and variance for each variable from the marginal densities. Thus,

$$\begin{aligned}
\mu_x &= E(X) \\
&= \int\!\!\int x\, f(x, y)\,dx\,dy \\
&= \int\!\!\int x\, f(y \mid x)\, f(x)\,dx\,dy \\
&= \int x \left[\int f(y \mid x)\,dy\right] f(x)\,dx \\
&= \int x\, f(x)\,dx
\end{aligned}$$

The term in brackets in the fourth line is the sum of the conditional Y probabilities and is equal to one for any X. By the same method

$$\sigma^2 = \operatorname{var}(X) = \int (x - \mu_x)^2 f(x)\,dx$$

and similarly for the mean and variance of Y.

A new statistic in the bivariate case is the **covariance**. It is defined as

$$\sigma_{xy} = \operatorname{cov}(X, Y) = E\big[(x - \mu_x)(y - \mu_y)\big] = \int\!\!\int (x - \mu_x)(y - \mu_y) f(x, y)\,dx\,dy$$

and measures the linear association between the two variables. A related concept is the **correlation coefficient,** $\rho = \sigma_{xy}/\sigma_x\sigma_y$. For independently distributed variables the covariance is zero because Eq. (B.11) gives

$$\operatorname{cov}(X, Y) = \int (x - \mu_x) f(x)\,dx \int (y - \mu_y) f(y)\,dy = 0$$

In general the converse of this proposition is not true; that is, a zero covariance does not necessarily imply independence. An important exception, however, exists in the case of normally distributed variables. Here a zero covariance does imply independence. The bivariate normal density was introduced in Eq. (1.13) of Chapter 1. There we saw that if the correlation ρ is zero, the joint density factors into the product of two marginal, normal densities.

B.4 RELATIONS BETWEEN THE NORMAL, χ^2, t, AND F DISTRIBUTIONS

Let $z \sim N(0, 1)$ be a standard normal variable. If n values $z_1, z_2, \ldots, z_n$ are drawn at random from this distribution, squared, and summed, the resultant statistic is said to have a χ^2 distribution with n degrees of freedom:

$$(z_1^2 + z_2^2 + \cdots + z_n^2) \sim \chi^2(n)$$

The precise mathematical form of the χ^2 distribution need not concern us here. The important point is that it constitutes a *one-parameter* family of distributions. The parameter is conventionally labeled the *degrees of freedom* of the distribution. The mean and variance of the distribution are given by

$$E[\chi^2(n)] = n \qquad \text{and} \qquad \text{var}[\chi^2(n)] = 2n \tag{B.12}$$

The t distribution is defined in terms of a standard normal variable and an independent χ^2 variable. Let

$$z \sim N(0, 1) \qquad \text{and} \qquad y \sim \chi^2(n)$$

where z and y are independently distributed. Then

$$t = \frac{z\sqrt{n}}{\sqrt{y}} \tag{B.13}$$

has Student's t distribution with n degrees of freedom. The t distribution, like that of χ^2, is a one-parameter family. It is symmetrical about zero and tends asymptotically to the standard normal distribution. Its critical values are given in Appendix D.

The F distribution is defined in terms of two independent χ^2 variables. Let y_1 and y_2 be independently distributed χ^2 variables with n_1 and n_2 degrees of freedom, respectively. Then the statistic

$$F = \frac{y_1/n_1}{y_2/n_2} \tag{B.14}$$

has the F distribution with (n_1, n_2) degrees of freedom. Critical values are tabulated in Appendix D. In using the table note carefully that n_1 refers to the degrees of freedom of the variable in the numerator and n_2 to the variable in the denominator.

If the expression for t in Eq. (B.13) is squared, the result may be written

$$t^2 = \frac{z^2/1}{y/n}$$

where z^2, being the square of a standard normal variable, has the $\chi^2(1)$ distribution. Thus $t^2(n) = F(1, n)$; that is, the square of a t variable with n degrees of freedom is an F variable with $(1, n)$ degrees of freedom. It also follows from Eq. (B.12) that

$$E\left[\frac{\chi^2(n)}{n}\right] = 1 \qquad \text{and} \qquad \text{var}\left[\frac{\chi^2(n)}{n}\right] = \frac{2}{n}$$

This variance goes to zero with increasing n, and so

$$\text{plim}\left[\frac{\chi^2(n)}{n}\right] = 1$$

It also follows that the distribution of $n_1 \cdot F(n_1, n_2)$ goes asymptotically to $\chi^2(n_1)$ as n_2 goes to infinity.

B.5 EXPECTATIONS IN BIVARIATE DISTRIBUTIONS

We have already had an example of this process in the foregoing development of μ_x. More generally, let $g(x, y)$ be some well-defined scalar function of the two variables. Then

$$\begin{aligned} E[g(x, y)] &= \int\int g(x, y) f(x, y)\, dx\, dy \\ &= \int\left[\int g(x, y) f(y \mid x)\, dy\right] f(x)\, dx \end{aligned}$$

The term in brackets is the expected value of $g(x, y)$ in the conditional distribution $f(y \mid x)$. Let this conditional expectation be denoted by $E_{\cdot|x}$. Each conditional expectation is thus a function of x only, and they are then averaged over the marginal x distribution. an operation denoted by E_x. The process might just as well have been started with the alternative decomposition of the joint density. Thus

$$E[g(x, y)] = E_x[E_{\cdot|x} g(x, y)] = E_y[E_{\cdot|y} g(x, y)] \tag{B.15}$$

This result is often useful in evaluating complicated functions. To obtain an *unconditional* expectation one may take expectations conditional on one of the variables and then take expectations over that variable. The process is referred to as the **law of iterated expectations.**

B.6 MULTIVARIATE DENSITIES

Univariate and bivariate densities are simple cases of multivariate densities. In the general case we let $\boldsymbol{x}$ denote a vector of random variables $X_1, X_2, \ldots, X_k$. Now the same letter is used for all variables and the subscripts distinguish the variables. An observation at sample point t gives the k-vector $\boldsymbol{x}_t' = [x_{1t} \quad x_{2t} \quad \cdots, \quad x_{kt}]$. Each variable has an expected value

$$\mu_i = E(X_i) \qquad i = 1, 2, \ldots, k$$

Collecting these expected values in a vector $\boldsymbol{\mu}$ gives

$$\boldsymbol{\mu} = E(\boldsymbol{x}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_k) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}$$

The application of the operator E to a vector means that E is applied to each element in the vector. The variance of X_i, by definition, is

$$\text{var}(X_i) = E[(X_i - \mu_i)^2]$$

and the covariance between X_i and X_j is

$$\text{cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$$

If we form the vector $(\boldsymbol{x} - \boldsymbol{\mu})$ and then define the matrix

$$E\{(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})'\} = E\left\{\begin{bmatrix} (X_1 - \mu_1) \\ \vdots \\ (X_k - \mu_k) \end{bmatrix} [(X_1 - \mu_1) \quad \cdots \quad (X_k - \mu_k)]\right\}$$

$$= \begin{bmatrix} E(X_1 - \mu_1)^2 & E(X_1 - \mu_1)(X_2 - \mu_2) & \cdots & E(X_1 - \mu_1)(X_k - \mu_k) \\ E(X_2 - \mu_2)(X_1 - \mu_1) & E(X_2 - \mu_2)^2 & \cdots & E(X_2 - \mu_2)(X_k - \mu_k) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_k - \mu_k)(X_1 - \mu_1) & E(X_k - \mu_k)(X_2 - \mu_2) & \cdots & E(X_k - \mu_k)^2 \end{bmatrix}$$

we see that the elements of this matrix are the variances and covariances of the X variables, the variances being displayed on the principal diagonal and the covariances in the off-diagonal positions. The matrix is known as a variance-covariance matrix, or more simply as a variance matrix or covariance matrix. We will usually refer to it as a variance matrix and denote it by

$$\text{var}(\boldsymbol{x}) = E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})'] = \boldsymbol{\Sigma} \tag{B.16}$$

The variance matrix is clearly symmetric. To examine the conditions under which it is positive definite or not, define a scalar random variable Y as a linear combination of the Xs, that is,

$$y = (\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{c}$$

where $\boldsymbol{c}$ is any arbitrary nonnull column vector with k elements. Squaring both sides and taking expectations gives

$$\begin{aligned} E(Y^2) &= E[\boldsymbol{c}'(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{c}] \\ &= \boldsymbol{c}'E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})']\boldsymbol{c} \\ &= \boldsymbol{c}'\Sigma\boldsymbol{c} \end{aligned}$$

There are two useful points to notice about this development. First, $(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{c}$ is a scalar, and so its square may be found by multiplying it by its transpose. Second, whenever we take the expectation of a matrix product, the expectation operator may be moved to the right past any vectors or matrices consisting only of constants, but it must be stopped in front of any expression involving random variables. Since Y is a scalar random variable, $E(Y^2) \geq 0$. Thus

$$\boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c} \geq 0$$

and $\boldsymbol{\Sigma}$ is positive semidefinite. The quadratic form would only take on the zero value if the X deviations were linearly dependent. Thus, in the absence of linear dependence between the random variables the variance matrix is positive definite.

B.7 MULTIVARIATE NORMAL pdf

The random variables have some **multivariate pdf,** written $f(\boldsymbol{x}) = f(X_1, X_2, \ldots, X_k)$. The most important multivariate pdf is the multivariate normal. It is defined in terms of its mean vector $\boldsymbol{\mu}$ and its variance matrix $\boldsymbol{\Sigma}$. The equation is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{B.17}$$

A compact shorthand statement of Eq. (B.17) is

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

to be read "the variables in $\boldsymbol{x}$ are distributed according to the multivariate normal law with mean vector $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$." A useful exercise for the reader is to check that substitution of 1 and 2 for k gives the univariate and bivariate pdf's already noticed. Each variable in $\boldsymbol{x}$ has a marginal distribution that is univariate normal, that is, $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, k$, where σ_i^2 is the ith element on the principal diagonal of $\boldsymbol{\Sigma}$.

An important special case of Eq. (B.17) occurs when all of the Xs have the same variance σ^2 and are all pairwise uncorrelated.[1] Then

$$\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$$

giving $\qquad \left|\boldsymbol{\Sigma}\right| = \sigma^{2k} \qquad$ and $\qquad \boldsymbol{\Sigma}^{-1} = \dfrac{1}{\sigma^2}\boldsymbol{I}$

The multivariate pdf then simplifies to

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{1}{(2\pi\sigma^2)^{k/2}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu})'(\boldsymbol{x}-\boldsymbol{\mu})\right] \\ &= \prod_{i=1}^{k}\left\{\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(X_i-\mu_i)^2\right]\right\} \qquad \text{(B.18)} \\ &= f(X_1)f(X_2)\cdots f(X_k) \end{aligned}$$

so that the multivariate density is the product of the marginal densities, that is, the Xs are distributed independently of one another. This result is extremely important. *Zero correlations between normally distributed variables imply statistical independence.* The result does not necessarily hold for variables that are not normally distributed.

[1] The assumption of a common variance is only made for simplicity. All that is required for the result is that the variance matrix be diagonal.

A more general case of this property may be derived as follows. Suppose that $\boldsymbol{\Sigma}$ is block diagonal with the form

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \tag{B.19}$$

where $\boldsymbol{\Sigma}_{11}$ is square and nonsingular of order r and $\boldsymbol{\Sigma}_{22}$ is square and nonsingular of order $(k - r)$. The form of Eq. (B.19) means that each and every variable in the set $X_1, X_2, \ldots, X_r$ is uncorrelated with each and every variable in the set $X_{r+1}, X_{r+2}, \ldots, X_k$. If we apply a similar partitioning to $\boldsymbol{x}$ and $\boldsymbol{\mu}$ and use the result that in Eq. (B.19) $|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{11}|\,|\boldsymbol{\Sigma}_{22}|$, substitution in Eq. (B.17) gives

$$\begin{aligned} f(\boldsymbol{x}) = &\left\{\frac{1}{(2\pi)^{r/2}\left|\boldsymbol{\Sigma}_{11}\right|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)\right]\right\} \\ &\times\left\{\frac{1}{(2\pi)^{(k-r)/2}\left|\boldsymbol{\Sigma}_{22}\right|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)\right]\right\} \end{aligned} \tag{B.20}$$

that is,

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_1)\,f(\boldsymbol{x}_2)$$

The multivariate density for all k variables is the product of two separate multivariate densities, one for the first r variables, and the other for the remaining $(k - r)$ variables.

B.8 DISTRIBUTIONS OF QUADRATIC FORMS

Suppose

$$\boldsymbol{x} \sim N(\mathbf{0}, \boldsymbol{I})$$

that is, the k variables in $\boldsymbol{x}$ have independent (standard) normal distributions, each with zero mean and unit variance. The sum of squares $\boldsymbol{x}'\boldsymbol{x}$ is a particularly simple example of a quadratic form with matrix $\boldsymbol{I}$. From the definition of the χ^2 variable,

$$\boldsymbol{x}'\boldsymbol{x} \sim \chi^2(k)$$

Suppose now that

$$\boldsymbol{x} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$$

The variables are still independent and have zero means, but each X has to be divided by σ to yield a variable with unit variance. Thus

$$\frac{X_1^2}{\sigma^2} + \frac{X_2^2}{\sigma^2} + \cdots + \frac{X_k^2}{\sigma^2} \sim \chi^2(k)$$

which may be written

$$\frac{1}{\sigma^2}\boldsymbol{x}'\boldsymbol{x} \sim \chi^2(k)$$

or

$$\boldsymbol{x}'(\sigma^2\boldsymbol{I})^{-1}\boldsymbol{x} \sim \chi^2(k) \tag{B.21}$$

The expression in Eq. (B.21) shows that the matrix of the quadratic form is the inverse of the variance matrix.

Suppose now that

$$\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{B.22}$$

where $\boldsymbol{\Sigma}$ is a positive definite matrix. The variables are still normally distributed with zero means, but they are no longer independently distributed. The equivalent expression to Eq. (B.21) would be

$$\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} \sim \chi^2(k) \tag{B.23}$$

This result is in fact true, but the proof is no longer direct since the Xs are correlated. The trick is to transform the Xs into Ys, which will be independently distributed standard normal variables. As shown in Appendix A, a positive definite matrix $\boldsymbol{\Sigma}$ can be factorized as $\boldsymbol{\Sigma} = \boldsymbol{PP}'$ where $\boldsymbol{P}$ is a nonsingular $k \times k$ matrix. Thus

$$\boldsymbol{\Sigma}^{-1} = (\boldsymbol{P}^{-1})'\boldsymbol{P}^{-1} \qquad \text{and} \qquad \boldsymbol{P}^{-1}\boldsymbol{\Sigma}(\boldsymbol{P}^{-1})' = \boldsymbol{I} \tag{B.24}$$

Define a k-element $\boldsymbol{y}$ vector as

$$\boldsymbol{y} = \boldsymbol{P}^{-1}\boldsymbol{x}$$

The Ys are multivariate normal since they are linear combinations of the Xs. Clearly, $E(\boldsymbol{y}) = \boldsymbol{0}$ and

$$\begin{aligned} \text{var}(\boldsymbol{y}) &= E\{\boldsymbol{P}^{-1}\boldsymbol{xx}'(\boldsymbol{P}^{-1})'\} \\ &= \boldsymbol{P}^{-1}\boldsymbol{\Sigma}(\boldsymbol{P}^{-1})' \\ &= \boldsymbol{I} \end{aligned}$$

The Ys are independent, standard normal variables, and so

$$\boldsymbol{y}'\boldsymbol{y} \sim \chi^2(k)$$

However, $$\boldsymbol{y}'\boldsymbol{y} = \boldsymbol{x}'(\boldsymbol{P}^{-1})'\boldsymbol{P}^{-1}\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$$

and so $$\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \sim \chi^2(k)$$

which is the result anticipated in Eq. (B.23).

Finally consider the quadratic form $\boldsymbol{x}'\boldsymbol{Ax}$ where $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{A}$ is a symmetric, idempotent matrix of rank $r \leq k$. If we denote the orthogonal matrix of eigenvectors by $\boldsymbol{Q}$, then

$$\boldsymbol{Q}'\boldsymbol{AQ} = \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \tag{B.25}$$

that is, $\boldsymbol{\Lambda}$ will have r ones and $(k - r)$ zeros on the principal diagonal. Define

$$\boldsymbol{y} = \boldsymbol{Q}'\boldsymbol{x} \qquad \boldsymbol{x} = \boldsymbol{Qy}$$

Then $E(\boldsymbol{y}) = \boldsymbol{0}$ and $$\begin{aligned} \text{var}(\boldsymbol{y}) &= E(\boldsymbol{yy}') \\ &= E(\boldsymbol{Q}'\boldsymbol{xx}'\boldsymbol{Q}) \\ &= \boldsymbol{Q}'\boldsymbol{IQ} \\ &= \boldsymbol{I} \end{aligned}$$

The Ys are thus independent, standard normal variables. By using Eq. (B.25), the quadratic form may now be expressed as

$$\begin{aligned} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} &= \boldsymbol{y}'\boldsymbol{Q}'\boldsymbol{A}\boldsymbol{Q}\boldsymbol{y} \\ &= Y_1^2 + Y_2^2 + \cdots + Y_r^2 \end{aligned}$$

and so $$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \sim \chi^2(r)$$

The general result is this:

If $\boldsymbol{x} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{A}$ is symmetric idempotent of rank r,

$$\frac{1}{\sigma^2}\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \sim \chi^2(r)$$

This result can be used to derive the distribution of the residual sum of squares (RSS) in the linear regression model. It was shown in (3.17) that

$$\boldsymbol{e} = \boldsymbol{M}\boldsymbol{y} \qquad \text{where} \qquad \boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

$\boldsymbol{M}$ is symmetric idempotent and $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}$. Thus

$$\boldsymbol{e} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}) = \boldsymbol{M}\boldsymbol{u}$$

Therefore, $$\boldsymbol{e}'\boldsymbol{e} = \boldsymbol{u}'\boldsymbol{M}\boldsymbol{u}$$

The assumption that $\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ then gives

$$\frac{\boldsymbol{e}'\boldsymbol{e}}{\sigma^2} \sim \chi^2(r)$$

where r is the rank of $\boldsymbol{M}$. From the result in Appendix A that the rank of a symmetric idempotent matrix is equal to its trace, we have

$$\begin{aligned} \rho(\boldsymbol{M}) &= \operatorname{tr}\left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right] \\ &= n - \operatorname{tr}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\right] \\ &= n - k \end{aligned}$$

and so, finally, $$\frac{\boldsymbol{e}'\boldsymbol{e}}{\sigma^2} \sim \chi^2(n-k)$$

which is the result stated in Eq. (3.37).

B.9 INDEPENDENCE OF QUADRATIC FORMS

Suppose $\boldsymbol{x} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ and we have two quadratic forms $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric idempotent matrices. We seek the condition for the two forms to be independently distributed. Because the matrices are symmetric idempotent,

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})'(\boldsymbol{A}\boldsymbol{x}) \qquad \text{and} \qquad \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x} = (\boldsymbol{B}\boldsymbol{x})'(\boldsymbol{B}\boldsymbol{x})$$

If each of the variables in the vector $\boldsymbol{A}\boldsymbol{x}$ has zero correlation with each variable in $\boldsymbol{B}\boldsymbol{x}$, these variables will be distributed independently of one another, and hence any function of the one set of variables, such as $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$, will be distributed independently of any function of the other set, such as $\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}$. The covariances between the variables in $\boldsymbol{A}\boldsymbol{x}$ and those in $\boldsymbol{B}\boldsymbol{x}$ are given by

$$E\{(\boldsymbol{A}\boldsymbol{x})(\boldsymbol{B}\boldsymbol{x})'\} = E(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}'\boldsymbol{B}) = \sigma^2\boldsymbol{A}\boldsymbol{B}$$

These covariances (and hence the correlations) are all zero if and only if

$$\boldsymbol{A}\boldsymbol{B} = \mathbf{0} \tag{B.26}$$

Since both matrices are symmetric, the condition may be stated equivalently as $\boldsymbol{B}\boldsymbol{A} = \mathbf{0}$. Thus, *two quadratic forms in normal variables with symmetric idempotent matrices will be distributed independently if the product of the matrices is the null matrix.*

B.10 INDEPENDENCE OF A QUADRATIC FORM AND A LINEAR FUNCTION

Assume $\boldsymbol{x} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$. Let $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ be a quadratic form with A a symmetric idempotent matrix of order k, and let $\boldsymbol{L}\boldsymbol{x}$ be an r-element vector, each element being a linear combination of the Xs. Thus $\boldsymbol{L}$ is of order $r \times k$, and we note that it need not be square nor symmetric. If the variables in $\boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{L}\boldsymbol{x}$ are to have zero covariances, we require

$$E(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}'\boldsymbol{L}') = \sigma^2\boldsymbol{A}\boldsymbol{L}' = \mathbf{0}$$

giving the condition $$\boldsymbol{L}\boldsymbol{A} = \mathbf{0} \tag{B.27}$$

This result may be used to prove the independence of the coefficient vector and the residual sum of squares in the linear regression model. From Eq. (3.23), $\boldsymbol{b} - \boldsymbol{\beta} = (X'X)^{-1}X'\boldsymbol{u}$, giving the matrix of the linear form as $\boldsymbol{L} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The matrix of the quadratic form is $\boldsymbol{A} = \boldsymbol{M}$. It follows simply that Eq. (B.27) is satisfied and the independence is established.

Index