In **Chapter 15** we first discuss the nature of the simultaneity problem and show why the method of ordinary least squares (OLS), which we have used so extensively in the preceding chapters, may not be appropriate for estimating the parameters of the simultaneous equation models. Two alternative methods of estimation, namely, the method of *indirect least squares* (ILS) and the method of *two-stage least squares* (2SLS), are presented, along with discussion of their relative merits and demerits.

We also discuss in this chapter the so-called *identification problem*, which precedes estimation. The crux of the problem is that two or more equations in a simultaneous equation model may look so alike that it may not be easy to tell (identify) which is the one that we really want to estimate. Thus, in a two-equation model involving the supply and demand for a product, if we have data only on the price and quantity of the product, it may not be possible to tell which is the demand function and which is the supply function. To find this out, we need some additional information. In this chapter we show how, with the *order condition of identification*, one can tell whether a particular equation is identified.

The concepts and techniques used in this chapter, as usual, are amply illustrated with numerical and actual economic examples.

CHAPTER 15

SIMULTANEOUS EQUATION MODELS

All the regression models we have considered so far have been *single equation* regression models in that a single dependent variable (Y) was expressed as a function of one more or more explanatory variables (the Xs). The underlying economic theory determined why Y was treated as the dependent variable and the Xs as the determining or causal variables. In other words, in such single equation regression models the causality, if any, ran from the Xs to Y. Thus, in our mortgage debt illustrative example in Section 7.4, it was economic theory that suggested that income (X_2) and mortgage cost (X_3) determined the demand for mortgage loans (Y).

However, there are situations in which such a *unidirectional relationship* between Y and Xs cannot be maintained. It is quite possible that the Xs not only affect Y but Y can also affect one or more Xs. If that is the case, we have a *bilateral*, or *feedback, relationship* between Y and the Xs. Obviously, if this is the case, the single equation modeling strategy that we have discussed in the previous chapters will not suffice, and in some cases it may be quite inappropriate because it may lead to biased (in the statistical sense) results. To take into account the bilateral relationship between Y and the Xs, we will therefore need more than one regression equation. Regression models in which there is more than one equation and in which there are feedback relationships among variables are known as **simultaneous equation regression models.** In the rest of this chapter we will discuss the nature of such simultaneous equation models. Our treatment of the topic is heuristic. For a detailed treatment of this topic, the reader may consult the references.[1]

[1]An extended treatment of this subject can be found in Damodar N. Gujarati, *Basic Econometrics*, 3d ed., McGraw-Hill, New York, 1995, Chaps. 18–20.

15.1 THE NATURE OF SIMULTANEOUS EQUATION MODELS

The best way to proceed is to consider some examples from economics.

Example 15.1. The Keynesian model of income determination. A beginning student of economics is exposed to the simple Keynesian model of income determination. Using the standard macroeconomics textbook convention, let C stand for consumption (expenditure), Y for income, I for investment (expenditure), and S for savings. The simple Keynesian model of income determination consists of the following two equations:

$$\text{Consumption function:} \quad C_t = B_1 + B_2 Y_t + u_t \tag{15.1}$$

$$\text{Income identity:} \quad Y_t = C_t + I_t \tag{15.2}$$

where t is the time subscript, u is the stochastic error term, and $I_t = S_t$.

This simple Keynesian model assumes a *closed economy* (i.e., there is no foreign trade) and no government expenditure [recall that the income identity is generally written as $Y_t = C_t + I_t + G_t + NX_t$, where G is government expenditure and NX is net export (export − import)]. The model also assumes that I, investment expenditure, is determined *exogenously*, say, by the private sector.

The consumption function states that consumption expenditure is linearly related to income; the stochastic error term is added to the function to reflect the fact that in empirical analysis the relation between the two is only approximate. The (national income) identity says that total income is equal to the sum of consumption expenditure and investment expenditure, the latter is equal to total savings. As we know, the slope coefficient B_2 in the consumption function is the *marginal propensity to consume (MPC)*, the amount of extra consumption expenditure resulting from an extra dollar of income. Keynes assumed that MPC is positive but less than 1, which is reasonable because people may save part of their additional income.

Now we can see the feedback, or simultaneous, relationship between consumption expenditure and income. From Equation (15.1) we see that income affects consumption expenditure, but from Equation (15.2) we also see that consumption is a component of income. Thus, consumption expenditure and income are *interdependent*. The objective of analysis is to find out how consumption (expenditure) and income are determined simultaneously. Thus consumption and income are *jointly dependent* variables. In the language of simultaneous equation modeling, such jointly dependent variables are known as **endogenous variables.** In the simple Keynesian model, investment I is not an endogenous variable, for its value is determined independently; so it is called an **exogenous,** or **predetermined, variable.** In more refined Keynesian models, investment can also be made endogenous.

In general, an endogenous variable is a "variable that is an inherent part of the system being studied and that is determined within the system. In other words, a variable that is caused by other variables in a causal system," and an exogenous variable "is a variable entering from and determined from outside the system being studied. A causal system says nothing about its exogenous variables."[2]

Equations (15.1) and (15.2) represent a two-equation model involving two endogenous variables, C and Y. If there are more endogenous variables, there will be more equations, one for each of the endogenous variables. Some equations in the system are *structural*, or *behavioral, equations* and some are *identities*. Thus, in our simple Keynesian model, Eq. (15.1) is a **structural,** or **behavioral, equation,** for it depicts the structure or behavior of a particular sector of the economy, the consumption sector here. The coefficients (or parameters) of the structural equations, such as B_1 and B_2 are known as *structural coefficients*. Equation (15.2) is an **identity,** a relationship that is true by definition: total income is equal to total consumption plus total investment.

Example 15.2. Demand and supply model. As every student of economics knows, the price P of a commodity and the quantity Q sold are determined by the intersection of the demand and supply curves for that commodity. Thus, assuming for simplicity that the demand and supply curves are linearly related to price and adding the stochastic, or random error, terms u_1 and u_2, we may write the empirical demand and supply functions as:

$$\text{Demand function:} \quad Q_t^d = A_1 + A_2 P_t + u_{1t} \tag{15.3}$$

$$\text{Supply function:} \quad Q_t^s = B_1 + B_2 P_t + u_{2t} \tag{15.4}$$

$$\text{Equilibrium condition:} \quad Q_t^d = Q_t^s \tag{15.5}$$

where Q_t^d = quantity demanded, Q_t^s = quantity supplied, and t = time.

According to economic theory, A_2 is expected to be negative (downward-sloping demand curve) and B_2 is expected to be positive (upward-sloping supply curve). Equations (15.3) and (15.4) are both structural equations, the former representing the consumers and the latter the suppliers. The As and Bs are structural coefficients.

Now it is not too difficult to see why there is a simultaneous, or two-way, relationship between P and Q. If, for example, u_{1t} [in Eq. (15.3)] changes because of changes in other variables affecting demand (such as income, wealth, and tastes), the demand curve will shift upward if u_{1t} is positive and downward if u_{1t} is negative. As Figure 15-1 shows, a shift in the demand curve changes both P and Q. Similarly, a change in u_{2t} (because of strikes, weather, hurricanes) will shift the supply curve, again affecting both P and Q. Therefore, there is a bilateral, or simultaneous, relationship between the two variables; the P and Q variables are thus *jointly dependent* or *endogenous, variables.* This is known as the **simultaneity problem.**

[2] W. Paul Vogt, *Dictionary of Statistics and Methodology: A Nontechnical Guide for the Social Sciences,* Sage Publications, California, 1993, pp. 81, 85.

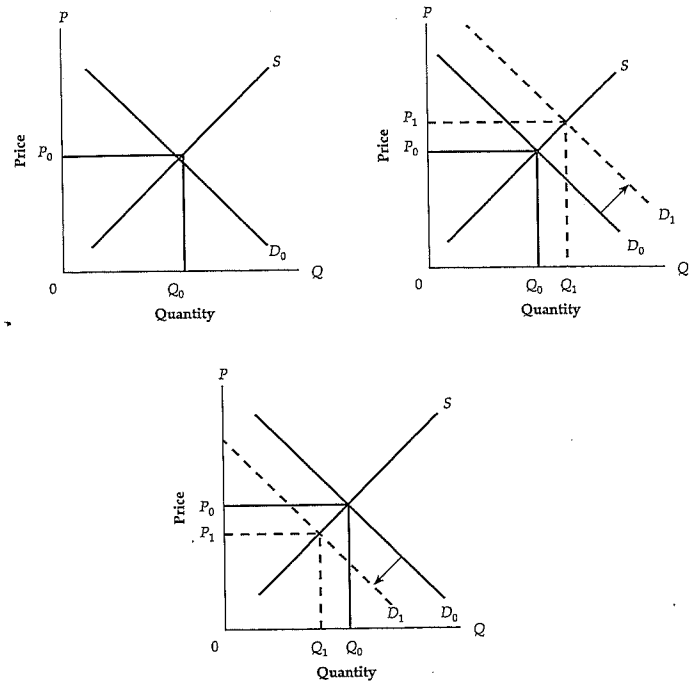

FIGURE 15-1
Interdependence of price and quantity.

15.2 THE SIMULTANEOUS EQUATION BIAS: INCONSISTENCY OF OLS ESTIMATORS

Why is simultaneity a problem? To understand the nature of this problem, return to Example 15.1, which discusses the simple Keynesian model of income determination. Assume for the moment that we neglect the simultaneity between consumption expenditure and income and just estimate the consumption function (15.1) by the usual OLS procedure. Using the OLS formula given in Eq. (5.17), we obtain

$$b_2 = \frac{\sum(C_t - \bar{C})(Y_t - \bar{Y})}{\sum(Y_t - \bar{Y})^2} \tag{15.6}$$

Now recall from Chapter 6 that if we work within the framework of the classical linear regression model (CLRM), which is the framework we have used thus far, the OLS estimators are BLUE (best linear unbiased estimator).[3] Is b_2 given in Equation (15.6) a BLUE estimator of the true marginal propensity to consume B_2? It can be shown that in the presence of the simultaneity problem the OLS estimators are generally not BLUE. In our case b_2 is not a BLUE estimator of B_2. In particular, b_2 is a *biased* estimator of B_2; on average, it underestimates or overestimates the true B_2. A formal proof of this statement is given in Appendix 15A. But intuitively it is easy to see why b_2 cannot be BLUE.

As discussed in Section 6.1, one of the assumptions of the CLRM is that the stochastic error term u and the explanatory variable(s) *are not correlated*. Thus, in the Keynesian consumption function Y (income) and error term u_t must not be correlated, if we want to use OLS to estimate the parameters of the consumption function (15.1). But that is not the case here. To see this, we proceed as follows:

$$\begin{aligned} Y_t &= C_t + I_t \qquad (15.2)\\ &= (B_0 + B_1 Y_t + u_t) + I_t \qquad \text{substituting for } C_t \text{ from (15.1)}\\ &= B_0 + B_1 Y_t + u_t + I_t \end{aligned}$$

Therefore, transferring the $B_1 Y_t$ term to the left-hand side and simplifying, we obtain

$$Y_t = \frac{B_0}{1 - B_1} + \frac{1}{1 - B_1} I_t + \frac{1}{1 - B_1} u_t \tag{15.7}$$

Notice an interesting feature of this equation. National income Y not only depends on investment I but also on the stochastic error term u! Recall that the error term u represents all kinds of influences not explicitly included in the model. Let us suppose that one of these influences is consumer confidence as measured by, say, the consumer confidence index developed by the University of Michigan. Suppose consumers feel upbeat about the economy because of a boom in the stock market (as happened in the United States in 1996 and 1997). Therefore, consumers increase their consumption expenditure, which affects income Y in view of the income identity (15.2). This increase in income will lead to another round of increase in consumption because of the presence of Y in the consumption function (15.1), which will lead to further increases in income, and so on. What will be the end result of this process? Students familiar with elementary macroeconomics will recognize that the end result will depend on the value of the *multiplier* $\frac{1}{1 - B_2}$. If, for example, the MPC (B_2) is 0.8 (i.e., 80 cents of every additional dollar's worth of income is spent on consumption), the multiplier will be 5.

[3] This is the essence of the *Gauss-Markov theorem* discussed in Sec. 6.3. It is this theorem that gives theoretical justification for the use of the method of ordinary least squares.

The point to note is that Y and u in Eq. (15.1) are correlated, and hence we cannot use OLS to estimate the parameters of the consumption function (15.1). If we persist in using it, the estimators will be biased. Not only that, as Appendix 15A shows, the estimators are not even *consistent*. As discussed in Section 4.4, roughly speaking, an estimator is said to be an inconsistent estimator if it does not approach the true parameter value even if the sample size increases indefinitely. *In sum, then, because of the correlation between Y and u, the estimator b_2 is biased (in small samples) as well as inconsistent (in large samples).* This just about destroys the usefulness of OLS as an estimating method in the context of simultaneous equation models. Obviously, we need to explore other estimating methods. One such method is discussed in the following section. In passing, note that if an explanatory variable in a regression equation is correlated with the error term in that equation, that variable essentially becomes a random, or stochastic, variable. In most of the regression models considered previously, we either assumed that the explanatory variables assume fixed values, or if they were random, they were uncorrelated with the error term. Such is not the case in the present instance.

Before proceeding further, notice an interesting feature of Equation (15.7): It expresses Y (income) as a function of I (investment), which is given exogenously, and error term u. Such an equation, which expresses an *endogenous* variable solely as a function of an exogenous variable(s) and the error term, is known as a **reduced form equation** (regression). We will see the utility of such reduced form equations shortly.

If we now substitute Y from Eq. (15.7) into the consumption function (15.1), we obtain the reduced form equation for C as

$$C_t = \frac{B_1}{1 - B_2} + \frac{B_2}{1 - B_2} I_t + \frac{1}{1 - B_2} u_t \tag{15.8}$$

As in Eq. (15.7), this equation expresses the endogenous variable C (consumption) solely as a function of the exogenous variable I and the error term.

15.3 THE METHOD OF INDIRECT LEAST SQUARES (ILS)

For reasons just stated, we should not use OLS to estimate the parameters B_1 and B_2 of the consumption function (15.1) because of correlation between Y and u. What is the alternative? The alternative can be found in Equation (15.8). Why not simply regress C on I, using the method of OLS? Yes, we could do that, because I, being exogenous by assumption, is uncorrelated with u; this was not the case with the original consumption function (15.1)

But how does the regression (15.8) enable us to estimate the parameters of the original consumption function (15.1), the object of our primary interest? This is easy enough. Let us write Eq. (15.8) as

$$C_t = A_1 + A_2 I_t + v_t \tag{15.9}$$

where $A_1 = B_1/(1 - B_2)$, $A_2 = B_2/(1 - B_2)$, and $v_t = u_t/(1 - B_2)$. Like u, v is also a stochastic error term; it is simply a rescaled u. The coefficients A_1 and A_2 are known as the *reduced form coefficients* because they are the coefficients attached to the reduced form (regression) equation. Observe that the reduced form coefficients are (nonlinear) combinations of the original structural coefficients of consumption function (15.1).

Now from the relationship between the A and B coefficients just given, it is easy to verify that

$$B_1 = \frac{A_1}{1 + A_2} \tag{15.10}$$

$$B_2 = \frac{A_2}{1 + A_2} \tag{15.11}$$

Therefore, once we estimate A_1 and A_2, we can easily "retrieve" B_1 and B_2 from them.

This method of obtaining the estimates of the parameters of the consumption function (15.1) is known as the method of **indirect least squares (ILS)**, for we obtain the estimates of the original parameters indirectly by first applying OLS to the reduced form regression (15.9). What are the statistical properties of ILS estimators? We state (without proof) that the ILS estimators are *consistent* estimators, that is, as the sample size increases indefinitely, these estimators converge to their true population values. However, in small, or finite, samples, the ILS estimators may be biased. In contrast, the OLS estimators are biased as well as inconsistent.[4]

15.4 INDIRECT LEAST SQUARES: AN ILLUSTRATIVE EXAMPLE

As an application of the ILS, consider the data given in Table 15-1. The data on consumption, income, and investment are for the United States for the years 1960 to 1994 and are given in billions of 1992 dollars; that is, the data are expressed in the purchasing power of the dollar in 1992. It should be noted that the data on income is simply the sum of consumption and investment expenditure, in keeping with our simple Keynesian model of income determination.

Following our discussion of ILS, we first estimate the reduced form regression (15.8). Using the data given in Table 15-1, we obtain the following results; the results are given in the standard format as per Eq. (6.48).

$$\widehat{C_t} = 191.8108 + 4.46407\ I_t \tag{15.12}$$

$$\text{se} = (130.0330)\ (0.2071) \qquad r^2 = 0.9336$$

$$t = (1.4750) \quad (21.5531)$$

[4]For a proof of these statements, you may consult Damodar N. Gujarati, *Basic Econometrics*, 3d ed., McGraw-Hill, New York, 1995, Chap. 18.

TABLE 15-1
Income, consumption, and investment for the United States, 1960–1994

Year	Income	Personal consumption expenditure	Gross private domestic investment
1960	1703.100	1432.600	270.5000
1961	1726.700	1461.500	265.2000
1962	1832.300	1533.800	298.5000
1963	1914.700	1596.600	318.1000
1964	2036.900	1692.300	344.6000
1965	2191.600	1799.100	392.5000
1966	2325.500	1902.000	423.5000
1967	2365.500	1958.600	406.9000
1968	2500.000	2070.200	429.8000
1969	2601.900	2147.500	454.4000
1970	2617.300	2197.800	419.5000
1971	2746.900	2279.500	467.4000
1972	2938.000	2415.900	522.1000
1973	3116.100	2532.600	583.5000
1974	3059.100	2514.700	544.4000
1975	3010.500	2570.000	440.5000
1976	3250.900	2714.300	536.6000
1977	3456.900	2829.800	627.1000
1978	3637.600	2951.600	686.0000
1979	3724.700	3020.200	704.5000
1980	3635.900	3009.700	626.2000
1981	3736.100	3046.400	689.7000
1982	3671.900	3081.500	590.4000
1983	3888.400	3240.600	647.8000
1984	4239.200	3407.600	831.6000
1985	4395.700	3566.500	829.2000
1986	4522.500	3708.700	813.8000
1987	4642.800	3822.300	820.5000
1988	4798.700	3972.700	826.0000
1989	4926.500	4064.600	861.9000
1990	4949.500	4132.200	817.3000
1991	4843.100	4105.800	737.3000
1992	5010.200	4219.800	790.4000
1993	5197.000	4339.700	857.3000
1994	5450.700	4471.100	979.6000

Source: *Economic Report of the President*, 1996 Table B-2, p. 282.

Note: Income = consumption expenditure + gross private domestic investment. All data are in billions of 1992 dollars.

Thus $a_1 = 191.8108$ and $a_2 = 4.4640$, which are respectively the estimates of A_1 and A_2, the parameters of the reduced form regression (15.8). Now we use Eqs. (15.10) and (15.11) to obtain the *estimates* of B_1 and B_2, the parameters of the consumption function (15.1):

$$b_1 = \frac{a_1}{1 + a_2} = \frac{191.8108}{1 + 4.4640}$$

$$= 35.1044 \tag{15.13}$$

$$b_2 = \frac{a_2}{1 + a_2} = \frac{4.4640}{1 + 4.4640}$$

$$= 0.8169 \tag{15.14}$$

These are the ILS estimates of the parameters of the consumption function. And the estimated consumption function now is

$$\widehat{C_t} = 35.1044 + 0.8169\, Y_t \tag{15.15}$$

Thus, the estimated MPC is about 0.82.

For comparison, we give the results based on OLS, that is, the results obtained by directly regressing C on Y without the intermediary of the reduced form:

$$\widehat{C_t} = 6.5415 + 0.8252\, Y_t \tag{15.16}$$

$$\text{se} = (23.9802) \quad (0.0066) \qquad r^2 = 0.997$$

$$t = (0.2727) \quad (124.6031)$$

See the difference between the ILS and OLS estimates of the parameters of the consumption function. Although the estimated marginal propensities to consume do not differ substantially, there is a difference in the estimated intercept values. Which results should we trust? The ones obtained from the method of ILS, for we know that in the presence of the simultaneity problem, the OLS results are not only biased but are inconsistent as well.[5]

It would seem that one can always use the method of indirect least squares to estimate the parameters of simultaneous equation models. The issue is whether we can retrieve the original structural parameters from these reduced form estimates: Sometimes we can, and sometimes we cannot. The answer depends on the so-called *identification problem*. In the following section we discuss this problem and then in the ensuing sections discuss other methods of estimating the parameters of the simultaneous equation models.

15.5 THE IDENTIFICATION PROBLEM: A ROSE BY ANY OTHER NAME MAY NOT BE A ROSE

Let us return to the supply and demand model of Example 15.2. Suppose we have data on P and Q only, and we want to estimate the demand function. Suppose we regress Q on P. How do we know that this regression in fact estimates a demand function? You might say that if the slope of the estimated

[5]Notice that we have given standard errors and t values for the OLS regression (15.16) but not for the ILS regression (15.15). This is because the coefficients of the latter, obtained from Eqs. (15.13) and (15.14), are nonlinear functions of a_1 and a_2, and there is no *simple* method of obtaining standard errors of nonlinear functions.

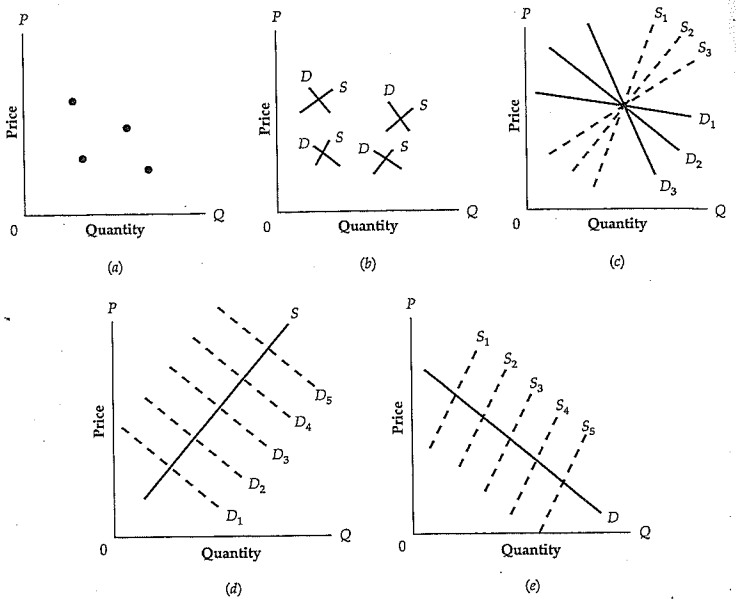

FIGURE 15-2
Hypothetical supply and demand functions and the identification problem.

regression is negative, it is a demand function because of the inverse relationship between price and quantity demanded. But suppose the slope coefficient turns out to be positive. What then? Do you then say that it must be a supply function because there is a positive relationship between price and quantity supplied?

You can see the potential problem involved in simply regressing quantity on price: A given P_t and Q_t combination represents simply the point of intersection of the appropriate supply and demand curves because of the equilibrium condition that demand is equal to supply. To see this more clearly, consider Figure 15-2. Figure 15-2(*a*) gives a few scatterpoints relating P to Q. Each scatterpoint represents the intersection of a demand and supply curve, as shown in Figure 15-2(*b*). Now consider a single point, such as that shown in Figure 15-2(*c*). There is no way we can be sure which demand and supply curve of a whole family of curves shown in that panel generated that particular point. Clearly, some additional information about the nature of the demand and supply curves is needed. For example, if the demand curve shifts over

time because of change in income, tastes, etc., but the supply curve remains relatively stable, as in Figure 15-2(*d*), the scatterpoints trace out a supply curve. In this situation, we say that *the supply curve is identified;* that is, we can uniquely estimate the parameters of the supply curve. By the same token, if the supply curve shifts over time because of weather factors (in the case of agricultural commodities) or other extraneous factors but the demand curve remains relatively stable, as in Figure 15-2(*e*), the scatterpoints trace out a demand curve. In this case, we say that the demand curve is identified; that is, we can uniquely estimate its parameters.

The **identification problem** therefore addresses whether we can estimate the parameters of the particular equation (be it a demand or a supply function) uniquely. If that is the case, we say that the particular equation is **exactly identified.** If we cannot estimate the parameters, we say that the equation is **unidentified** or **underidentified.** Sometimes it can happen that there is more than one numerical value for one or more parameters of that equation. In that case, we say that the equation is **overidentified.** We will not consider each of these cases briefly.

Underidentification

Consider once again the Example 15.2. By the equilibrium condition that supply equals demand, we obtain

$$A_1 + A_2P_t + u_{1t} = B_1 + B_2P_t + u_{2t} \tag{15.17}$$

Solving Equation (15.17), we obtain the equilibrium price

$$P_t = \Pi_1 + v_{1t} \tag{15.18}$$

where

$$\Pi_1 = \frac{B_1 - A_1}{A_2 - B_2} \tag{15.19}$$

$$v_{1t} = \frac{u_{2t} - u_{1t}}{A_2 - B_2} \tag{15.20}$$

where v_1 is a stochastic error term, which is a linear combination of the us. The symbol Π is read as pi and is used here to represent a reduced form regression coefficient.

Substituting P_t from Equation (15.18) into either the supply or demand function of Example 15.2, we obtain the following equilibrium quantity:

$$Q_t = \Pi_2 + u_{2t} \tag{15.21}$$

where

$$\Pi_2 = \frac{A_2B_1 - A_1B_2}{A_2 - B_2} \tag{15.22}$$

$$v_{2t} = \frac{A_2u_{2t} - B_2u_{1t}}{A_2 - B_2} \tag{15.23}$$

where v_2 is also a stochastic, or random, error term.

Equations (15.19) and (15.21) are reduced form regressions. Now our demand and supply model has four structural coefficients, A_1, A_2, B_1, and B_2, but there is no unique way of estimating them from the two reduced form coefficients, Π_1 and Π_2. As elementary algebra teaches us, to estimate four unknowns we must have four (independent) equations. Incidentally, if we run the reduced form regressions (15.19) and (15.21) we see that there are no explanatory variables, only the constants, the Π's, and these constants will simply give the mean values of P and Q (why?). There is no way of estimating the four structural coefficients from the two mean values. In short, both the demand and supply functions are *unidentified.*

Just or Exact Identification

We have already considered this case in the previous section where we discussed the estimation of the Keynesian consumption function using the method of indirect least squares. As shown there, from the reduced form regression (15.12), we were able to obtain unique values of the parameters of the consumption function, as can be seen from Eqs. (15.13) and (15.14).

To further illustrate exact identification, let us continue with our demand and supply example, but now we modify the model as follows:

$$\text{Demand function:} \quad Q_t^d = A_1 + A_2 P_t + A_3 X_t + u_{1t} \tag{15.24}$$

$$\text{Supply function:} \quad Q_t^d = B_1 + B_2 P_t + u_{2t} \tag{15.25}$$

where in addition to the variables already defined, X = income of the consumer. Thus, the demand function states that the quantity demanded is a function of its price as well as income of the consumer; economic theory of demand generally has price and income as its two main determinants. The inclusion of the income variable in the model will give us some additional information about consumer behavior. It is assumed that the income of the consumer is determined exogenously.

Using the market-clearing mechanism, quantity demanded = quantity supplied, we obtain

$$A_1 + A_2 P_t + A_3 X_t + u_{1t} = B_1 + B_2 P_t + u_{2t} \tag{15.26}$$

Solving Equation (15.26) provides the following equilibrium value of P_t:

$$P_t = \Pi_1 + \Pi_2 X_t + v_{1t} \tag{15.27}$$

where the reduced form coefficients are

$$\Pi_1 = \frac{B_1 - A_1}{A_2 - B_2} \tag{15.28}$$

$$\Pi_2 = -\frac{A_3}{A_2 - B_2} \tag{15.29}$$

$$v_{1t} = \frac{u_{2t} - u_{1t}}{A_2 - B_2} \tag{15.30}$$

Substituting the equilibrium value of P_t into the preceding demand or supply function, we obtain the following equilibrium, or market clearing, quantity:

$$Q_t = \Pi_3 + \Pi_4 X_t + v_{2t} \tag{15.31}$$

where

$$\Pi_3 = \frac{A_2 B_1 - A_1 B_2}{A_2 - B_2} \tag{15.32}$$

$$\Pi_4 = -\frac{A_3 B_2}{A_2 - B_2} \tag{15.33}$$

$$v_{2t} = \frac{A_2 u_{2t} - B_2 u_{1t}}{A_2 - B_2} \tag{15.34}$$

Since Equations (15.27) and (15.31) are both reduced form regressions, as noted before, OLS can always be applied to estimate their parameters. The question that remains is whether we can uniquely estimate the parameters of the structural equations from the reduced form coefficients.

Observe that the demand and supply models (15.24) and (15.25) contain five structural coefficients, A_1, A_2, A_3, B_1, and B_2. But we have only four equations to estimate them, the four reduced form coefficients, the four Πs. So, we cannot obtain unique values of all the five structural coefficients. But which of these coefficients can be uniquely estimated? The reader can verify that the parameters of the supply function can be uniquely estimated, for

$$B_1 = \Pi_3 - B_2 \Pi_1 \tag{15.35}$$

$$B_2 = \frac{\Pi_4}{\Pi_2} \tag{15.36}$$

Therefore, the supply function is exactly identified. But the demand function is unidentified because there is no unique way of estimating its parameters, the A coefficients.

Observe an interesting fact: *It is the presence of an additional variable in the demand function that enables us to identify the supply function.* Why? The inclusion of the income variable in the demand equation provides us some additional information about the variability of the function, as indicated in Figure 15-2(*d*). The figure shows how the intersection of the stable supply curve with the shifting demand curve (on account of changes in income) enables us to trace (identify) the supply curve.

How can the demand function be identified? Suppose we include P_{t-1}, the one-period lagged value of price as an additional variable in the supply function (15.25). This amounts to saying that the supply depends not only on the current price but also on the price prevailing in the previous period, not an unreasonable assumption for many agricultural commodities. Since at time t the value of P_{t-1} is already known, we can treat it as an exogenous, or predetermined, variable. Thus the new model is

$$\text{Demand function:} \quad Q_t^d = A_1 + A_2 P_t + A_3 X_t + u_{1t} \tag{15.37}$$

$$\text{Supply function:} \quad Q_t^s = B_1 + B_2 P_t + B_3 P_{t-1} + u_{2t} \tag{15.38}$$

Using Equations (15.37) and (15.38) and the market-clearing condition, the reader is invited to obtain the reduced form regressions and verify that now both the demand and supply functions are identified; each reduced form regression will have X_t and P_{t-1} as explanatory variables, and since the values of these variables are determined outside the model, they are uncorrelated with the error terms. Once again notice how the inclusion or exclusion of a variable(s) from an equation helps us to identify that equation, that is, to obtain unique values of the parameters of that equation. Thus it is the exclusion of the P_{t-1} variable from the demand function that helps us to identify it, just as the exclusion of the income variable (X_t) from the supply function helps us to identify it. One implication is that an equation in a simultaneous equation system cannot be identified if it includes all the variables (endogenous as well as exogenous) in the system. Later we provide a simple rule of identification that generalizes this idea (see Section 15.6).

Overidentification

Although the exclusion of certain variables from an equation may enable us to identify it as we just showed, sometimes we can overdo it. This leads to the problem of **overidentification,** a situation in which there is more than one value for one or more parameters of an equation in the model. Let us see how this can happen.

Once again return to the demand-supply model and write it as

$$\text{Demand function:} \quad Q_t = A_1 + A_2P_t + A_3X_T + A_4W_t + u_{1t} \tag{15.39}$$

$$\text{Supply function:} \quad Q_t = B_1 + B_2P_t + B_3P_{t-1} + u_{2t} \tag{15.40}$$

where in addition to the variables introduced previously, W_t stands for the wealth of the consumer. For many commodities, income as well as wealth are important determinants of demand. Compare the demand and supply models (15.37) and (15.38) with the models (15.39) and (15.40). Whereas originally the supply function excluded only the income variable, in the new model it excludes both the income and wealth variables. Before, the exclusion of the income variable from the supply function enabled us to identify it; now the exclusion of both the income and wealth variables from the supply function *overidentifies* it in the sense we have two estimates of the supply parameter B_2, as we show below.

Equating models (15.39) and (15.40), we now obtain the following reduced form regressions:

$$P_t = \Pi_1 + \Pi_2 X_T + \Pi_3 W_t + \Pi_4 P_{t-1} + v_{1t} \tag{15.41}$$

$$Q_t = \Pi_5 + \Pi_6 X_T + \Pi_7 W_t + \Pi_8 P_{t-1} + v_{2t} \tag{15.42}$$

where

$$\Pi_1 = \frac{B_1 - A_1}{A_2 - B_2} \qquad \Pi_2 = -\frac{A_3}{A_2 - B_2}$$

$$\Pi_3 = -\frac{A_4}{A_2 - B_2} \qquad \Pi_4 = \frac{B_3}{A_2 - B_2}$$

$$\Pi_5 = \frac{A_2B_1 - A_1B_2}{A_2 - B_2} \qquad \Pi_6 = -\frac{A_3B_2}{A_2 - B_2} \tag{15.43}$$

$$\Pi_7 = -\frac{A_4B_2}{A_2 - B_2} \qquad \Pi_8 = \frac{A_2B_3}{A_2 - B_2}$$

$$v_{1t} = \frac{u_{2t} - u_{1t}}{A_2 - B_2} \qquad v_{2t} = \frac{A_2u_{2t} - B_2u_{1t}}{A_2 - B_2}$$

Remember that the supply and demand models we are considering have in all seven structural coefficients, the four As and three Bs. But there are eight reduced form coefficients in Eq. (15.43). We have more equations than unknowns. Clearly, there is more than one solution to a parameter. As the reader can readily verify, we have in fact two values for B_2:

$$B_2 = \frac{\Pi_7}{\Pi_3} \quad \text{or} \quad B_2 = \frac{\Pi_6}{\Pi_2} \tag{15.44}$$

And there is no reason to believe that these two estimates will be the same.

Since B_2 appears in the denominators of all the reduced form coefficients given in Eq. (15.43), the ambiguity in the estimation of B_2 will be transmitted to other structural coefficients also. Why do we obtain such a result? It seems that we have *too much information*—exclusion of either the income or wealth variable would have sufficed to identify the supply function. This is the opposite of the case of underidentification, where there was too little information. The point here is that more information may not be always better! It should be noted, though, that the problem of overidentification occurs not because we are deliberately adding more variables. It is simply that sometimes theory tells us what variables to include or exclude from an equation, and the equation then ends up either unidentified or identified (either exactly or over).

In summary, an equation in a simultaneous equation model may be unidentified, exactly identified, or overidentified. There is nothing we can do about underidentification, assuming the model is correct. Underidentification is not a statistical problem that can be solved with a larger sample size. You can look at those four dots in Figure 15-2(*a*) all year long, but they will never tell you the slope of the supply and demand curves that generated them. If an equation is exactly identified, we can use the method of indirect least squares (ILS) to estimate its parameters. If an equation is overidentified, ILS will not provide unique estimates of the parameters. Fortunately, we can use the method of **two-stage least-squares (2SLS)** to estimate the parameters of an overidentified equation. But before we turn to 2SLS, we would like to find out

if there is a systematic way to determine whether an equation is underidentified, exactly identified, or overidentified; the method of reduced form regression to determine identification is rather cumbersome, especially if the model contains several equations.

15.6 RULES FOR IDENTIFICATION: THE ORDER CONDITION OF IDENTIFICATION

To understand the so-called **order condition of identification,** we introduce the following notations:

m = number of endogenous (or jointly dependent) variables in the model
k = total number of variables (endogenous and exogenous) *excluded* from the equation under consideration

Then,

1. If $k = m - 1$, the equation is exactly identified.
2. If $k > m - 1$, the equation is overidentified.
3. If $k < m - 1$, the equation is underidentified

To apply the order condition, all one has to do is to count the number of endogenous variables (= number of equations in the model) and the total number of variables (endogenous as well as exogenous) excluded from the particular equation under consideration. Although the order condition of identification is only *necessary and not sufficient,* in most practical applications is has been found to be very helpful.

Thus, applying the order condition to the supply and demand models (15.39) and (15.40), we see that $m = 2$ and that the supply function excludes the variables X_t and W_t (that is, $k = 2$). Since $k > m - 1$, the supply equation is overidentified. As for the demand function, it excludes P_{t-1}. Since $k = m - 1$, the demand function is identified. But we now have a slight complication. If we try to estimate the parameters of the demand function from the reduced form coefficients given in Eq. (15.43), the estimates will not be unique because B_2, which enters into the computations, takes two values, as shown in Eq. (15.44). This complication can, however, be avoided if we use the method of 2SLS, to which we now turn.

15.7 ESTIMATION OF AN OVERIDENTIFIED EQUATION: THE METHOD OF TWO-STAGE LEAST SQUARES (2SLS)

To illustrate the method of *two-stage least squares* **(2SLS),** consider the following model:

Income function: $$Y_t = A_1 + A_2M_t + A_3I_t + A_4G_t + u_{1t} \quad (15.45)$$

Money supply function: $$M_t = B_1 + B_2Y_t + u_{2t} \quad (15.46)$$

where

Y = income
M = stock of money
I = investment expenditure
G = government expenditure on goods and services
u_1, u_2 = stochastic error terms.

In this model, the variables I and G are assumed exogenous.

The *income function,* a hybrid of the quantity-theory and the Keynesian approaches to income determination states that income is determined by the money supply, investment expenditure, and government expenditure. The *money supply* function states that the stock of the money supply is determined by the FED (Federal Reserve System) on the basis of the level of income. Obviously, we have a *simultaneity problem* here because of the feedback between income and the money supply.

Applying the order condition of identification, we can check that the income equation is unidentified (because it excludes no variable in the model), whereas the money supply function is overidentified, as it excludes two variables in the system (note that $m = 2$ in this model).

Since the income equation is underidentified, there is nothing we can do to estimate its parameters. What about the money supply function? Since it is overidentified, if we use ILS to estimate its parameters, we will not obtain unique estimates for the parameters; actually, B_2 will have two values. What about OLS? Because of the likely correlation between income Y and the stochastic error term u_2, OLS estimates will be inconsistent in view of our earlier discussion. What, then, is the alternative?

Suppose in the money supply function (15.46) we find a *surrogate* or *proxy* or an *instrumental variable* for Y such that, although resembling Y, it is uncorrelated with u_2. If one can find such a proxy, OLS can be used straightforwardly to estimate the parameters of the money supply function. (Why?) But how does one obtain such a proxy or instrumental variable? One answer is provided by the method of **two-stage least squares (2SLS).** As the name indicates, the method involves two successive applications of OLS. The process is as follows:

Stage 1. To get rid of the likely correlation between income Y and the error term u_2, first regress Y on *all* predetermined variables in the whole model, not just that equation. In the present case, this means regressing Y on the predetermined variables I (gross private domestic investment) and G (government expenditure) as follows:

$$Y_t = \Pi_1 + \Pi_2 I_t + \Pi_3 G_t + w_t \tag{15.47}$$

where w is a stochastic error term. From Equation (15.47), we obtain

$$\widehat{Y_t} = \widehat{\Pi_1} + \widehat{\Pi_2} I_t + \widehat{\Pi_3} G_t \tag{15.48}$$

where $\widehat{Y_t}$ is the estimated mean value of Y, given the values of I and G. Note, the ^ over the Π coefficients indicate that these are the estimated values of the true Πs.

We can write Eq. (15.47) therefore as

$$Y_t = \widehat{Y_t} + w_t \tag{15.49}$$

which shows that the (stochastic) Y consists of two parts: $\widehat{Y_t}$, which from Equation (15.48) is a linear combination of the predetermined variables I and G and a random component w_t. Following OLS theory, $\hat{Y}$ and w are therefore uncorrelated. (Why? See problem 5.13.)

Stage 2. The overidentified money supply function can now be written as

$$\begin{aligned} M_t &= B_1 + B_2(\widehat{Y_t} + w_t) + u_{2t} \\ &= B_1 + B_2\widehat{Y_t} + (u_{2t} + B_2 w_t) \\ &= B_1 + B_2\widehat{Y_t} + v_t \end{aligned} \tag{15.50}$$

where $v_t = u_{2t} + B_2 w_t$

Comparing Equations (15.50) and (15.46), we see that they are very similar in appearance, the only difference being that Y is replaced by $\hat{Y}$, the latter being obtained from Eq. (15.48). What is the advantage of this? It can be shown that although Y in the original money supply function (15.46) is likely to be correlated with the stochastic error term u_2 (hence rendering OLS inappropriate), $\hat{Y}$ in Eq. (15.50) is uncorrelated with v_t *asymptotically*, that is, in large sample (or, more accurately, as the sample size increases indefinitely). As a result, OLS can now be applied to Eq. (15.50), which will give *consistent estimates* of the parameters of the money supply function (15.46). This is an improvement over the direct application of OLS to Eq. (15.46), for in that situation the estimates are likely to be biased as well as inconsistent.[6]

15.8 2SLS: A NUMERICAL EXAMPLE

Let us continue with the money supply and income models of (15.45) and (15.46). Table 15-2 gives data on Y (income, as measured by GDP), M (money supply, as measured by the $M2$ measure of money supply), I (investment as measured by gross private domestic investment, GPDI), and G (federal government expenditure). The data are in billions of dollars, except the interest rate (as measured by the 6-month Treasury bill rate), which is a percentage.

[6]For further discussion of this somewhat technical point, see Damodar N. Gujarati, *Basic Econometrics*, 3d ed., McGraw-Hill, New York, 1995, Chap. 20.

TABLE 15-2
Macroeconomic data, United States, 1970–1994

obs	GDP	GPDI	Government	M2	TB
1970	1035.600	150.2000	115.9000	628.1000	6.562000
1971	1125.400	176.0000	117.1000	712.7000	4.511000
1972	1237.300	205.6000	125.1000	805.2000	4.466000
1973	1382.600	242.9000	128.2000	861.0000	7.178000
1974	1496.900	245.6000	139.9000	908.5000	7.926000
1975	1630.600	225.4000	154.5000	1023.200	6.122000
1976	1819.000	286.6000	162.7000	1163.700	5.266000
1977	2026.900	356.6000	178.4000	1286.500	5.510000
1978	2291.400	430.8000	194.4000	1388.600	7.572000
1979	2557.500	480.9000	215.0000	1496.900	10.01700
1980	2784.200	465.9000	248.4000	1629.300	11.37400
1981	3115.900	556.2000	284.1000	1793.300	13.77600
1982	3242.100	501.1000	313.2000	1953.200	11.08400
1983	3514.500	547.1000	344.5000	2187.700	8.750000
1984	3902.400	715.6000	372.6000	2378.400	9.800000
1985	4180.700	715.1000	410.1000	2576.000	7.660000
1986	4422.200	722.5000	435.2000	2820.300	6.030000
1987	4692.300	747.2000	455.7000	2922.300	6.050000
1988	5049.600	773.9000	457.3000	3083.500	6.920000
1989	5438.700	829.2000	477.2000	3243.000	8.040000
1990	5743.800	799.7000	503.6000	3356.000	7.470000
1991	5916.700	736.2000	522.6000	3457.900	5.490000
1992	6244.400	790.4000	528.0000	3515.300	3.570000
1993	6550.200	871.1000	522.1000	3583.600	3.140000
1994	6931.400	1014.400	516.3000	3617.000	4.660000

Source: Economic Report of the President, 1996,
GDP: Gross domestic product, billions of dollars, Table B-1, p. 280.
GPDI: Gross private domestic investment, billions of dollars, Table B-1, p. 280.
Government: Federal government expenditure, billions of dollars, Table B-1, p. 281.
M2: M2 money stock, billions of dollars, Table B-65, p. 355.
TB: 6-month Treasury bill rate, Table B-69, p. 360.

The data on interest rate are given for some problems at the end of the chapter. These data are annual and are for the period 1970 to 1994.

Stage 1 regression. To estimate the parameters of the money supply function (15.46), we first regress the stochastic variable Y (income) on the proxy variables I and G, which are treated as exogenous or predetermined. The results of this regression are

$$\begin{aligned} \widehat{Y_t} &= -333.1789 + 2.0307 I_t + 8.7188 G_t \\ \text{se} &= (151.1741) \quad (1.0054) \quad (1.6550) \qquad R^2 = 0.9742 \\ t &= (-2.2039) \quad (2.0198) \quad (5.2682) \end{aligned} \tag{15.51}$$

The reader can interpret these results in the usual manner. Notice that all the coefficients are statistically significant at the 5% level of significance.

Stage 2 regression. We estimate the money supply function (15.46) by regressing M not on the original income Y but on the Y as estimated in (15.48). The results are

$$\widehat{M_t} = 115.0327 + 0.5605\,\widehat{Y_t}$$
$$\text{se} = (54.4021) \quad (0.0136) \qquad r^2 = 0.9863 \qquad (15.52)^7$$
$$t = (2.1144) \quad (40.9735)$$

Note: Observe that there is a ^ on Y on the right-hand side.

OLS regression. For a comparison, we give the results of the regression (15.46) based on the *inappropriately* applied OLS:

$$\widehat{M_t} = 146.8179 + 0.5515\,Y_t$$
$$\text{se} = (53.3235) \quad (0.0133) \qquad r^2 = 0.9866 \qquad (15.53)$$
$$t = (2.7533) \quad (41.2458)$$

Comparing the 2SLS and the OLS results, you might say that the results are not vastly different. This may be so in the present case, but there is no guarantee that this always will be the case. Besides, we know that in theory 2SLS is better than OLS, especially in large samples.

We conclude our somewhat nontechnical discussion of the simultaneous equation models by noting that besides ILS and 2SLS there are other methods of estimating such models. But a discussion of these methods (e.g., the method of full information maximum likelihood) is beyond the scope of this introductory book.[8] Our primary purpose in this chapter was to introduce readers to the bare bones of the topic of simultaneous equation models to make them aware that on occasion we may have to go beyond the single equation regression modeling considered in the previous chapters.

15.9 SUMMARY

In contrast to the single equation models discussed in the preceding chapters, in simultaneous equation regression models what is a dependent (endogenous) variable in one equation appears as explanatory variable in another equation. Thus, there is a feedback relationship between the variables. This feedback creates the *simultaneity problem,* rendering OLS inapplicable to estimate the parameters of each equation individually. This is because the endogenous variable that appears as an explanatory variable in another equation may

[7]These standard errors are corrected to reflect the nature of the error term v_t. This is a technical point. The interested reader may consult Damodar N. Gujarati, *Basic Econometrics,* 3d ed., McGraw-Hill, New York, 1995, p. 705.

[8]The interested reader may refer to William H. Greene, *Econometric Analysis,* 3d ed., Prentice-Hall, New Jersey, 1997, Chap. 16.

be correlated with the stochastic error term of that equation. This violates one of the critical assumptions of OLS that the explanatory variable be either fixed, or nonrandom, or if random, it be uncorrelated with the error term. Because of this, if one uses OLS, the estimates thus obtained will be biased as well as inconsistent.

Besides the simultaneity problem, a simultaneous equation model may have an *identification problem.* An identification problem means we cannot uniquely estimate the values of the parameters of an equation. Therefore, before one estimates a simultaneous equation model, one must find out if an equation in such a model is identified.

One cumbersome method of finding out whether an equation is identified, is to obtain the *reduced form* equations of the model. A reduced form equation expresses a dependent (or endogenous) variable solely as a function of *exogenous,* or *predetermined, variables,* that is, variables whose values are determined outside the model. If there is a one-for-one correspondence between the reduced form coefficients and the coefficients of the original equation, then the original equation is identified.

A short-cut to determining identification is via the *order condition of identification.* The order condition counts the number of equations in the model and the number of variables in the model (both endogenous and exogenous). Then based on whether some variables are excluded from an equation but included in other equations of the model, the order condition decides whether an equation in the model is *underidentified, exactly identified,* or *overidentified.* An equation in a model is underidentified if we cannot estimate the values of the parameters of that equation. If we can obtain unique values of parameters of an equation, that equation is said to be exactly identified. If, on the other hand, the estimates of one or more parameters of an equation are not unique in the sense that there is more than one value of some parameters, that equation is said to be overidentified.

If an equation is underidentified, it is a dead-end case. There is not much one can do, short of changing the specification of the model (i.e., developing another model). If an equation is exactly identified, we can estimate it by the method of *indirect least squares (ILS).* ILS is a two-step procedure. In step 1, we apply OLS to the reduced form equations of the model, and then the original structural coefficients are retrieved from the reduced form coefficients. ILS estimators are consistent, that is, as the sample size increases indefinitely, the estimators converge to their true values. The parameters of the overidentified equation can be estimated by the method of *two-stage least squares (2SLS).* The basic idea behind 2SLS is to replace the explanatory variable that is correlated with the error term of the equation in which that variable appears by a variable that is not so correlated. Such a variable is called a *proxy,* or *instrumental, variable.* 2SLS estimators, like the ILS estimators, are consistent estimators.

Key Terms and Concepts

The key terms and concepts introduced in this chapter are

Simultaneous equation model
Endogenous variable
Exogenous (predetermined) variable
Structural or behavioral equation
Identities
Simultaneity problem
Reduced form equation
Indirect least squares (ILS)
Identification problem
  a) Underidentification
  b) Just or exact identification
  c) Overidentification
Identification rules
  a) Order condition of identification
Two-stage least squares (2SLS)

QUESTIONS

15.1. What is meant by the simultaneity problem?
15.2. What is the meaning of endogenous and exogenous variables?
15.3. Why is OLS generally inapplicable to estimate an equation embedded in a simultaneous equation model?
15.4. What happens if OLS is applied to estimate an equation in a simultaneous equation model?
15.5. What is meant by a reduced form (regression) equation? What is its use?
15.6. What is the meaning of structural, or behavioral, equation?
15.7. What is meant by indirect least squares? When is it used?
15.8. What is the nature of the identification problem? Why is it important?
15.9. What is the order condition of identification?
15.10. What may be meant by the statement that the order condition of identification "is a necessary but not sufficient condition" for identification?
15.11. Explain carefully the meaning of (1) underidentification, (2) exact identification, and (3) overidentification.
15.12. How does one estimate an underidentified equation?
15.13. What method(s) is used to estimate an exactly identified equation?
15.14. What is 2SLS used for?
15.15. Can 2SLS be also used to estimate an exactly identified equation?

PROBLEMS

15.16. Consider the following two-equation model:

$$Y_{1t} = A_1 + A_2Y_{2t} + A_3X_{1t} + u_{1t}$$

$$Y_{2t} = B_1 + B_2Y_{1t} + B_3X_{2t} + u_{2t}$$

where the Ys are the endogenous variables, the Xs the exogenous variables, and us the stochastic error terms.

(*a*) Obtain the reduced form regressions.
(*b*) Determine which of the equations is identified.

(*c*) For the identified equation, which method of estimation would you use and why?
(*d*) Suppose, a priori it is known that $A_3 = 0$. How would your answers to the preceding questions change? Why?

15.17. Consider the following model:

$$Y_{1t} = A_1 + A_2Y_{2t} + A_3X_{1t} + u_{1t}$$

$$Y_{2t} = B_1 + B_2Y_{1t} + u_{2t}$$

where the Ys are the endogenous variables, the Xs the exogenous, and the us the stochastic error terms. Based on this model, the following reduced form regressions were obtained

$$Y_{1t} = 6 + 8\ X_{1t}$$

$$Y_{2t} = 4 + 12\ X_{1t}$$

(*a*) Which structural coefficients, if any, can be estimated from these reduced form equations?
(*b*) How will our answer change if it is known a priori that (1) $A_2 = 0$ and (2) $A_1 = 0$.

15.18. Consider the following model:

$$R_t = A_1 + A_2M_t + A_3Y_t + u_{1t}$$

$$Y_t = B_1 + B_2R_t + u_{2t}$$

where Y = income (measured by GDP), R = interest rate (measured by 6-month Treasury bill rate, %), and M = money supply (measured by $M2$). Assume that M is determined exogenously.

(*a*) What economic rationale lies behind this model? (*Hint:* see any macroeconomics textbook.)
(*b*) Are the preceding equations identified?
(*c*) Using the data given in Table 15-2, estimate the parameters of the identified equation(s). Justify the method(s) you use.

15.19. Consider the following reformulation of the model given in problem 15.18.

$$R_t = A_1 + A_2M_t + A_3Y_t + u_{1t}$$

$$Y_t = B_1 + B_2R_t + B_3I_t + u_{2t}$$

where in addition to the variables defined in the preceding problem, I stands for investment (measured by gross private domestic investment, GPDI). Assume that M and I are exogenous.

(*a*) Which of the preceding equations is identified?
(*b*) Using the data of Table 15-2, estimate the parameters of the identified equation(s).
(*c*) Comment on the difference in the results of this and the preceding problem.

APPENDIX 15A
INCONSISTENCY OF OLS ESTIMATORS

To show that the OLS estimator of b_2 is an inconsistent estimator of B_2 because of correlation between Y_t and u_t, we start with the OLS estimator (15.6):

$$b_2 = \frac{\sum(C_t - \bar{C})(Y_t - \bar{Y})}{\sum(Y_t - \bar{Y})^2}$$

$$= \frac{\sum C_t y_t}{\sum y_t^2} \tag{15A.1}$$

where $y_t = (Y_t - \bar{Y})$.

Now substituting for C_t from Eq. (15.1), we obtain

$$b_2 = \frac{\sum(B_1 + B_2 Y_t + u_t)y_t}{\sum y_t^2}$$

$$= B_2 + \frac{\sum y_t u_t}{\sum y_t^2} \tag{15A.2}$$

where in the last step use is made of the fact that $\Sigma y_t = 0$ and $(\Sigma Y_t y_t / \Sigma y_t^2) = 1$ (why?).

Taking the expectation of Equation (15A.2), we get

$$E(b_2) = B_2 + E\left[\frac{\sum y_t u_t}{\sum y_t^2}\right] \tag{15A.3}$$

Unfortunately, we cannot readily evaluate the expectation of the second term in Equation (15A.3), since the expectations operator E is a linear operator. [*Note:* $E(A/B) \neq E(A)/E(B)$] But intuitively it should be clear that unless the second term in Eq. (15A.3) is zero, b_2 is a biased estimator of B_2.

Not only is b_2 biased, but it is inconsistent as well. An estimator is said to be consistent if its *probability limit (plim)* is equal to its true (population) value.[9] Using the properties of the plim, we can express[10]

$$\text{plim}(b_2) = \text{plim}(B_2) + \text{plim}\left[\frac{\sum y_t u_t}{\sum y_t^2}\right]$$

$$= B_2 + \text{plim}\left[\frac{\sum y_t u_t / n}{\sum y_t^2 / n}\right]$$

[9] If $\lim_{n \to \infty}$ Probability $\{|b_2 - B_2| < d\} = 1$, where $d > 0$ and n is sample size, we say that b_2 is a consistent estimator of B_2, which, for short, we write as $n \to \infty$ plim $(b_2) = B_2$. For further details, see Damodar N. Gujarati, *Basic Econometrics*, 3d ed., McGraw-Hill, New York 1995, pp. 782–783.

[10] Although $E(A/B) \neq E(A)/E(B)$, we can write plim $(A/B) =$ plim(A)/plim(B).

$$= B_2 + \frac{\text{plim}(\sum y_t u_t / n)}{\text{plim}(\sum y_t^2 / n)} \tag{15A.4}$$

where use is made of the properties of the plim operator that the plim of a constant (such as B_2) is that constant itself and that the plim of the ratio of two entities is the ratio of the plim of those entities.

Now as n increases indefinitely, it can be shown that

$$\text{plim}(b_2) = B_2 + \frac{1}{1 - B_2}\left(\frac{\sigma^2}{\sigma_y^2}\right) \tag{15A.5}$$

where σ^2 is the variance of u and σ_y^2 is the variance of Y.

Since B_2 (MPC) lies between 0 and 1, and since the two variance terms in Equation (15A.5) are positive, it is obvious from Eq. (15A.5) that plim (b_2) will be always greater than B_2, that is, b_2 will overestimate B_2 and the bias will not disappear no matter how large the sample size.