CHAPTER 11

Simultaneous Equations Models

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Explain why estimation of a supply and demand model requires an alternative to ordinary least squares (OLS).
2. Explain the difference between exogenous and endogenous variables.
3. Define the ''identification'' problem in simultaneous equations models.
4. Define the reduced form of a simultaneous equations model and explain its usefulness.
5. Explain why it is acceptable to estimate reduced-form equations by least squares.
6. Describe the two-stage least squares estimation procedure for estimating an equation in a simultaneous equations model, and explain how it resolves the estimation problem for least squares.

KEYWORDS

contemporaneous correlation
endogenous variables
exogenous variables
first-stage equation
identification
instrumental variables (IV) estimator
instruments
predetermined variables
reduced-form equation
reduced-form errors
reduced-form parameters
simultaneous equations
structural parameters
two-stage least squares

For most of us, our first encounter with economic models comes through studying supply and demand models, in which the market price and quantity of goods sold are *jointly determined* by the equilibrium of supply and demand. In this chapter, we consider econometric models for data that are jointly determined by two or more economic relations. These **simultaneous equations** models differ from those we have considered in previous chapters because in each model there are *two* or more dependent variables rather than just one.

Simultaneous equations models also differ from most of the econometric models we have considered so far, because they consist of a *set of equations*. For example, price and quantity are determined by the interaction of two equations, one for supply and the other for demand. Simultaneous equations models, which contain more than one dependent variable and more than

one equation, require special statistical treatment. The least squares estimation procedure *is not* appropriate in these models, and we must develop new ways to obtain reliable estimates of economic parameters.

Some of the concepts in this chapter were introduced in Chapter 10. However, reading Chapter 10 is *not* an absolute prerequisite for reading Chapter 11, which is largely self-contained. If you *have* read Chapter 10, you will observe that much of what you learned there will carry over to this chapter, including how simultaneous equations models fit into the big picture. If you *have not* read Chapter 10, referring back to portions of it will provide a deeper understanding of material presented in this chapter. This chapter on simultaneous equations is presented separately because its treatment was the first major contribution of econometrics to the wider field of statistics, and because of its importance in economic analysis.

11.1 A Supply and Demand Model

Supply and demand *jointly* determine the market price of a good and the quantity of it that is sold. Graphically, you recall that market equilibrium occurs at the intersection of the supply and demand curves, as shown in Figure 11.1. An econometric model that explains market price and quantity should consist of two equations, one for supply and the other for demand. It will be a simultaneous equations model, since both equations working together determine price and quantity. A very simple model might look like the following:

$$\text{Demand: } Q_i = \alpha_1 P_i + \alpha_2 X_i + e_{di} \tag{11.1}$$

$$\text{Supply: } Q_i = \beta_1 P_i + e_{si} \tag{11.2}$$

Based on economic theory, we expect the supply curve to be positively sloped, $\beta_1 > 0$, and the demand curve to be negatively sloped, $\alpha_1 < 0$. In this model, we assume that the quantity demanded (Q) is a function of price (P) and income (X). Quantity supplied is taken to be a function of only price. (We have omitted the intercepts to make the algebra easier. In practice, we would include intercept terms in these models.) The observation index $i = 1, \ldots, N$ may represent the market place at different points in time, or at different locations.

The point we wish to make very clear is that it takes *two* equations to describe the supply and demand equilibrium. The *two* equilibrium values, for price and quantity, P^* and Q^*, respectively, are determined at the same time. In this model, the variables P and Q are called **endogenous variables** because their values are determined within the system we have created. The endogenous variables P and Q are *dependent* variables and both are random variables. The income variable X has a value that is determined outside this system. Such variables are said to be **exogenous**, and these variables are treated like usual "x" explanatory variables.

Random errors are added to the supply and demand equations for the usual reasons.

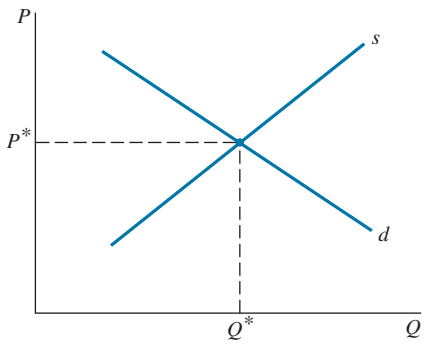

FIGURE 11.1 **Supply and demand equilibrium.**

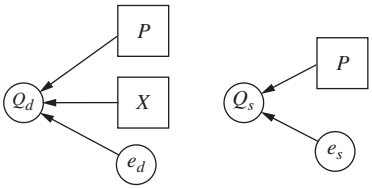

FIGURE 11.2 **Influence diagrams for two regression models.**

We adopt assumption SR2 from Chapter 2 for both the demand and supply equations, given any value of the exogenous variable $X_i, i = 1,\ldots, N$. To simplify notation, we refer to all the values of X_i as $\mathbf{X}$, where $\mathbf{X} = (X_1, X_2, \ldots, X_N)$. Then

$$E(e_{di}|\mathbf{X}) = 0, \qquad E(e_{si}|\mathbf{X}) = 0 \tag{11.3}$$

In Section 2.10, we coined the term "strictly exogenous" for an exogenous variable like this. It implies that $E(e_{di}) = E(e_{si}) = 0$; the unconditional expected value of each error equals zero. It also implies that any value of the exogenous variable X_j is uncorrelated with the error terms in the demand and supply equations, so $\text{cov}(e_{di}, X_j) = 0$ and $\text{cov}(e_{si}, X_j) = 0$. Further, the error terms in the demand and supply equations are assumed to be homoskedastic, $\text{var}(e_{di}|\mathbf{X}) = \sigma_d^2$, and $\text{var}(e_{si}|\mathbf{X}) = \sigma_s^2$. Finally, we also assume that there is no serial correlation and no correlation between the error terms of the two equations.

Let us emphasize the difference between simultaneous equations models and regression models using influence diagrams. An "influence diagram" is a graphical representation of relationships between model components. In the previous chapters, we would have modeled the supply and demand relationships as separate regressions, implying the influence diagrams in Figure 11.2. In this diagram the circles represent endogenous dependent variables and error terms. The squares represent exogenous explanatory variables. In regression analysis, the direction of the influence is one way: from the explanatory variable and the error term to the dependent variable. In this case there is no equilibrating mechanism that will lead quantity demanded to equal quantity supplied at a market-clearing price. For price to adjust to the market-clearing equilibrium, there must be an influence running from P to Q and from Q to P.

Recognizing that price P and quantity Q are *jointly determined*, and that there is feedback between them, suggests the influence diagram in Figure 11.3. In the simultaneous equations model we see the two-way influence, or feedback, between P and Q because they are jointly determined. The random error terms e_d and e_s affect both P and Q, suggesting a correlation between each of the endogenous variables and each of the random error terms. As we will see, this leads to failure of the ordinary least squares (OLS) estimator in simultaneous equations models. Income X is an exogenous variable that affects the endogenous variables, but there is no feedback from P and Q to X.

The fact that P is an endogenous variable on the right-hand side of the supply and demand equations means that we have an explanatory variable that is random. Not only is P random but it is

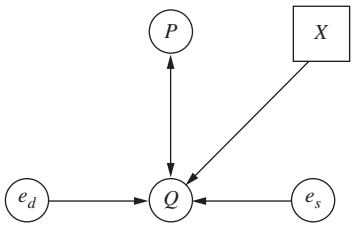

FIGURE 11.3 **Influence diagram for a simultaneous equations model.**

also **contemporaneously correlated** with the random errors in the demand and supply equations, that is, $\text{cov}(P_i, e_{di}) = E(P_i e_{di}) \neq 0$ and $\text{cov}(P_i, e_{si}) = E(P_i e_{si}) \neq 0$. When an explanatory variable is contemporaneously correlated with the regression error term then the OLS estimator is biased and inconsistent. We provide an intuitive argument for why this outcome is true in Section 11.3, and we prove it in Section 11.3.1.

11.2 The Reduced-Form Equations

The two structural equations (11.1) and (11.2) can be solved to express the endogenous variables P and Q as functions of the exogenous variable X. This reformulation of the model is called the **reduced form** of the structural equation system. The reduced form is very important in its own right, and also helps us understand the structural equation system. To find the reduced form, we solve equations (11.1) and (11.2) simultaneously for P and Q.

To solve for P, set Q in the demand and supply equations to be equal,

$$\beta_1 P_i + e_{si} = \alpha_1 P_i + \alpha_2 X_i + e_{di}$$

Then solve for P_i,

$$P_i = \frac{\alpha_2}{(\beta_1 - \alpha_1)} X_i + \frac{e_{di} - e_{si}}{(\beta_1 - \alpha_1)} = \pi_1 X_i + v_{1i} \tag{11.4}$$

To solve for Q_i, substitute the value of P_i in (11.4) into either the demand or supply equation. The supply equation is simpler, so substitute P_i into (11.2) and simplify:

$$\begin{aligned} Q_i = \beta_1 P_i + e_{si} &= \beta_1 \left[\frac{\alpha_2}{(\beta_1 - \alpha_1)} X_i + \frac{e_{di} - e_{si}}{(\beta_1 - \alpha_1)} \right] + e_{si} \\ &= \frac{\beta_1 \alpha_2}{(\beta_1 - \alpha_1)} X_i + \frac{\beta_1 e_{di} - \alpha_1 e_{si}}{(\beta_1 - \alpha_1)} = \pi_2 X_i + v_{2i} \end{aligned} \tag{11.5}$$

The parameters π_1 and π_2 in (11.4) and (11.5) are called **reduced-form parameters**. The errors v_{1i} and v_{2i} are **reduced-form errors**. The reduced forms are predictive equations. We assume that $E(P_i|X_i) = \pi_1 X_i$ and $E(Q_i|X_i) = \pi_2 X_i$. By definition $E(v_{1i}|X_i) = 0$ and $E(v_{2i}|X_i) = 0$, using assumptions (11.3), and also they are homoskedastic and serially uncorrelated if the same holds true for the structural equation errors e_{di} and e_{si}. Under these conditions, the ordinary least squares (OLS) estimators of the reduced-form parameters π_1 and π_2 are consistent, and have approximate normal distributions in large samples, whether the structural equation errors are normal or not. The most important aspect of the OLS estimators for the reduced-form parameters is that they are consistent estimators.

The reduced-form equations (11.4) and (11.5) have an endogenous variable on the left-hand side and exogenous variables, and a random error term, on the right-hand side. These are **first-stage equations** in the language of Chapter 10. We explain the term in Section 11.5 if you have not read Chapter 10. The terms **reduced-form equation** and **first-stage equation** are interchangeable.

The reduced-form equations are important for economic analysis. These equations relate the *equilibrium* values of the endogenous variables to the exogenous variables. Thus, if there is an increase in income X, π_1 is the expected increase in price, after market adjustments lead to a new equilibrium for P and Q. Similarly, π_2 is the expected increase in the expected equilibrium value of Q. (*Question*: how did we determine the directions of these changes?) Secondly, and using the same logic, the estimated reduced-form equations can be used to *predict* values of equilibrium price and quantity for different levels of income. Clearly CEOs and other market analysts are interested in the ability to forecast both prices and quantities sold of their products. Estimating the reduced-form equations makes such predictions possible.

11.3 The Failure of Least Squares Estimation

In this section, we explain why the OLS estimator should not be used to estimate an equation in a simultaneous equations model. For reasons that will become clear in the next section, we focus on the supply equation. In the supply equation (11.2), the endogenous variable P_i on the right-hand side of the equation is *contemporaneously correlated* with the error term e_{si}. Suppose there is a small change, or blip, in the error term e_{si}, say Δe_{si}. Trace the effect of this change through the system. The blip Δe_{si} in the error term of (11.2) is directly transmitted to the equilibrium value of P_i. This follows from the reduced form (11.4) that has P_i on the left and e_{si} on the right. Every change in the supply equation error term, e_{si}, has a direct effect on P_i. Because $\beta_1 > 0$ and $\alpha_1 < 0$, if $\Delta e_{si} > 0$, then $\Delta P_i < 0$. Thus, every time there is a change in e_{si} there is an associated change in P_i in the opposite direction. Consequently, P_i and e_{si} are negatively correlated.

The failure of OLS estimation for the supply equation can be explained as follows: OLS estimation of the relation between Q_i and P_i gives "credit" to price (P_i) for the effect of changes in the error term (e_{si}). This occurs because we do not observe the change in the error term, but only the change in P_i resulting from its correlation with the error e_{si}. The OLS estimator of β_1 will *understate* the true parameter value in this model because of the negative contemporaneous correlation between the endogenous variable P_i and the error term e_{si}. This occurs because we do not observe the change in the error term, but only the change in P_i resulting from its correlation with the error e_{si}. The least squares estimator of β_1 will *understate* the true parameter value in this model because of the negative contemporaneous correlation between the endogenous variable P_i and the error term e_{si}. In large samples, the least squares estimator will tend to be negatively biased in this model. This bias persists even if the sample size goes to infinity, and thus the least squares estimator is inconsistent. This means that the probability distribution of the least squares estimator will ultimately "collapse" about a point that is not the true parameter value as the sample size $N \to \infty$. See Section 5.7 for a general discussion of "large sample" properties of estimators. Here, we summarize by saying:

> The least squares estimator of parameters in a structural simultaneous equation is biased and inconsistent because of the contemporaneous correlation between the random error and the endogenous variables on the right-hand side of the equation.

11.3.1 Proving the Failure of OLS

Consider the supply and demand model in (11.1) and (11.2). To explain the failure of the OLS estimator of the supply equation, let us first obtain the conditional covariance between P_i and e_{si}.

$$
\begin{aligned}
\operatorname{cov}(P_i, e_{si}|\mathbf{X}) &= E\left\{\left[P_i - E(P_i|\mathbf{X})\right]\left[e_{si} - E(e_{si}|\mathbf{X})\right]\middle|\mathbf{X}\right\} \\
&= E(P_i e_{si}|\mathbf{X}) && \left[\text{since } E(e_{si}|\mathbf{X}) = 0\right] \\
&= E\left[(\pi_1 X_i + v_{1i})e_{si}|\mathbf{X}\right] && \left[\text{substitute for } P_i\right] \\
&= E\left[\left(\frac{e_{di} - e_{si}}{\beta_1 - \alpha_1}\right)e_{si}\middle|\mathbf{X}\right] && \left[\text{since } \pi_1 X_i \text{ is fixed}\right] \\
&= \frac{-E(e_{si}^2|\mathbf{X})}{\beta_1 - \alpha_1} && \left[\text{since } e_d, e_s \text{ assumed uncorrelated}\right] \\
&= \frac{-\sigma_s^2}{\beta_1 - \alpha_1} < 0
\end{aligned}
$$

What impact does the negative contemporaneous covariance have on the least squares estimator? The OLS estimator of the supply equation (11.2) (which does not have an intercept term) is

$$b_1 = \frac{\sum P_i Q_i}{\sum P_i^2}$$

Substitute for Q from the reduced-form equation (11.5) and simplify,

$$b_1 = \frac{\sum P_i(\beta_1 P_i + e_{si})}{\sum P_i^2} = \beta_1 + \sum\left(\frac{P_i}{\sum P_i^2}\right)e_{si}$$

The expected value of the least squares estimator is

$$E(b_1|\mathbf{X}) = \beta_1 + E\left[\sum\left(\frac{P_i}{\sum P_i^2}\right)e_{si}\middle|\mathbf{X}\right] = \beta_1 + E\left[\sum\left(\frac{P_i e_{si}}{\sum P_i^2}\right)\middle|\mathbf{X}\right] \quad \text{[move error to numerator]}$$

$$= \beta_1 + \sum\left[E\left(\frac{P_i e_{si}}{\sum P_i^2}\right)\middle|\mathbf{X}\right] \quad \text{[expected value of the sum is sum of expected values]}$$

$$\neq \beta_1 \quad \text{[expected value terms in the sum are not zero]}$$

In the final step, we have $E\left[\left(P_i e_{si}/\sum P_i^2\right)|\mathbf{X}\right] = E\left[g(P_i)e_{si}|\mathbf{X}\right] \neq 0$, where $g(P_i) = P_i/\sum P_i^2$. When finding the covariance between P_i and the random error e_{si}, we showed that $E(P_i e_{si}|\mathbf{X}) = E(P_i e_{si}) = -\sigma_s^2/(\beta_1 - \alpha_1) < 0$ and thus we suspect that $E\left[\left(P_i e_{si}/\sum P_i^2\right)|\mathbf{X}\right] < 0$, because $\sum P_i^2 > 0$, so that we suspect the least squares estimator exhibits a negative bias. However, the expected value of the ratio is not the ratio of expected values, so all we can really conclude is that the least squares estimator is biased, because e_{si} and P_i are contemporaneously correlated.

This bias does not disappear in larger samples, so the OLS estimator of the supply equation is inconsistent as well. The OLS estimator converges to a value less than β_1 and this is easier to show using asymptotic analysis similar to that in Chapter 5, equation (5.41). Rewrite the OLS estimators

$$b_1 = \beta_1 + \sum\left(\frac{P_i}{\sum P_i^2}\right)e_{si} = \beta_1 + \frac{\sum P_i e_{si}}{\sum P_i^2} = \beta_1 + \frac{\sum P_i e_{si}/N}{\sum P_i^2/N} = \beta_1 + \frac{\widehat{E(P_i e_{si})}}{\widehat{E(P_i^2)}}$$

Using the Law of Large Numbers, sample moments (averages) converge to population moments (expected values), so that

$$\widehat{E(P_i e_{si})} \xrightarrow{p} E(P_i e_{si}) = -\sigma_s^2/(\beta_1 - \alpha_1) < 0$$

and

$$\widehat{E(P_i^2)} \xrightarrow{p} E(P_i^2) > 0$$

Therefore

$$b_1 \xrightarrow{p} \beta_1 - \frac{\sigma_s^2/(\beta_1 - \alpha_1)}{E(P_i^2)} < \beta_1$$

11.4 The Identification Problem

In the supply and demand model given by (11.1) and (11.2),

- The parameters of the demand equation, α_1 and α_2, *cannot* be consistently estimated by *any* estimation method.
- The slope of the supply equation, β_1, *can* be consistently estimated.

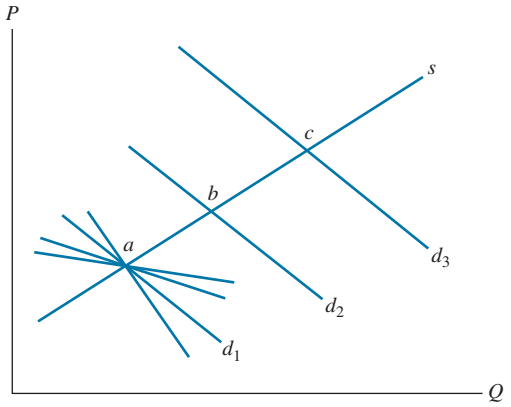

FIGURE 11.4 **The effect of changing income.**

How are we able to make such statements? The answer is quite intuitive, and it can be illustrated graphically. What happens when income X changes? The demand curve shifts and a new equilibrium price and quantity are created. In Figure 11.4 we show the demand curves d_1, d_2, and d_3 and equilibria, at points a, b, and c, for three levels of income. As income changes, data on price and quantity will be observed around the intersections of supply and demand. The random errors e_d and e_s cause small shifts in the supply and demand curves, creating equilibrium observations on price and quantity that are scattered about the intersections at points a, b, and c.

The data values will trace out the *supply curve*, suggesting that we can fit a line through them to estimate the slope β_1. The data values fall along the supply curve because income is *present* in the demand curve and *absent* from the supply curve. As income changes, the demand curve shifts but the supply curve remains fixed, resulting in observations along the supply curve.

There are *no* data values falling along any of the demand curves, and there is no way to estimate their slope. Any one of the infinite number of demand curves passing through the equilibrium points could be correct. Given the data, there is no way to distinguish the true demand curve from all the rest. Through the equilibrium point a we have drawn a few demand curves, each of which *could* have generated the data we observe.

The problem lies with the model that we are using. There is no variable in the supply equation that will shift it relative to the demand curve. If we were to add a variable to the supply curve, say W, then each time W is changed, the supply curve would shift, and the demand curve would stay fixed. The shifting of supply relative to a fixed demand curve (since W is *absent* from the demand equation) would create equilibrium observations along the demand curve, making it possible to estimate the slope of the demand curve and the effect of income on demand.

It is the *absence* of variables in one equation that are *present* in another equation that makes parameter estimation possible. A general rule, which is called a **necessary condition** for **identification** of an equation, is this:

A Necessary Condition for Identification

In a system of M simultaneous equations, which jointly determine the values of M endogenous variables, at least $M - 1$ variables must be absent from an equation for estimation of its parameters to be possible. When estimation of an equation's parameters is possible, then the equation is said to be *identified*, and its parameters can be estimated consistently. If fewer than $M - 1$ variables are omitted from an equation, then it is said to be *unidentified*, and its parameters cannot be consistently estimated.

In our supply and demand model there are $M = 2$ equations, so we require at least $M - 1 = 1$ variable to be omitted from an equation to identify it. There are a total of three variables: P, Q, and X. In the demand equation none of the variables are omitted; thus it is unidentified and its parameters cannot be estimated consistently. In the supply equation, one variable, income (X), is omitted; the supply curve is identified, and its parameter can be estimated.

The identification condition must be checked *before* trying to estimate an equation. If an equation is not identified, then changing the model must be considered before it is estimated. However, changing the model should not be done in a haphazard way; no important variable should be omitted from an equation just to identify it. The structure of a simultaneous equations model should reflect your understanding of how equilibrium is achieved and should be consistent with economic theory. Creating a false model is not a good solution to the identification problem.

This paragraph is for those who have read Chapter 10. The necessary condition for identification can be expressed in an alternative but equivalent fashion. The two-stage least squares estimation procedure was developed in Chapter 10 and shown to be an **instrumental variables estimator**. This procedure is developed further in the next section. The number of instrumental variables (IVs) required for estimation of an equation within a simultaneous equations model is equal to the number of right-hand side endogenous variables. In a typical equation within a simultaneous equations model, several **exogenous variables** appear on the right-hand side. Thus **instruments** must come from those exogenous variables omitted from the equation in question. Consequently, identification requires that the number of excluded exogenous variables in an equation be at least as large as the number of included right-hand side endogenous variables. This ensures an adequate number of IVs.

11.5 Two-Stage Least Squares Estimation

The most widely used method for estimating the parameters of an identified structural equation is called **two-stage least squares**, which is often abbreviated as 2SLS or TSLS. The name comes from the fact that it can be calculated using two OLS regressions. We will explain how it works by considering the supply equation in (11.2). Recall that we should not apply the usual OLS procedure to estimate β_1 in this equation because the endogenous variable P_i on the right-hand side of the equation is contemporaneously correlated with the error term e_{si}, causing the OLS estimator to be biased and inconsistent.

The variable P_i is composed of a systematic part, which is its expected value $E(P_i|X_i)$, and a random part, which is the reduced-form random error v_{1i}, that is,

$$P_i = E(P_i|X_i) + v_{1i} \tag{11.6}$$

In the supply equation (11.2), the portion of P_i that causes problems for the OLS estimator is v_{1i}, the random part. It is v_{1i} that causes P_i to be correlated with the error term e_{si}. If we knew $E(P_i|X_i)$, then we could replace P_i in (11.2) by (11.6) to obtain

$$Q_i = \beta_1 [E(P_i|X_i) + v_{1i}] + e_{si} = \beta_1 E(P_i|X_i) + (\beta_1 v_{1i} + e_{si}) \tag{11.7}$$

In (11.7) the explanatory variable on the right-hand side is $E(P_i|X_i)$. It depends only on the exogenous variable, and it is not correlated with the error term. We could apply OLS to (11.7) to consistently estimate β_1.

Of course, we cannot use the variable $E(P_i|X_i)$ in place of P_i since we do not know it. However, we can consistently estimate $E(P_i|X_i)$. Let $\hat{\pi}_1$ come from the fitted OLS estimation of the reduced-form equation for P_i. A consistent estimator for $E(P_i|X_i)$ is

$$\hat{P}_i = \hat{\pi}_1 X_i$$

Using $\hat{P}_i$ as a replacement for $E(P_i|X_i)$ in (11.7), we obtain

$$Q_i = \beta_1 \hat{P}_i + \hat{e}_{*i} \tag{11.8}$$

In large samples, $\hat{P}_i$ and the random error $\hat{e}_{*i}$ are uncorrelated, and consequently the parameter β_1 can be consistently estimated by applying OLS to (11.8).

The OLS estimator of (11.8) is the **two-stage least squares** estimator of β_1, which is consistent and asymptotically normal. Because the two-stage least squares estimator is consistent it converges to the true value in large samples. That the estimator is asymptotically normal means that if we have a large sample, the usual tests and confidence interval estimators can be used. To summarize, the two stages of the estimation procedure are:

1. OLS estimation of the reduced-form equation for P_i and the calculation of its predicted value, $\hat{P}_i$
2. OLS estimation of the structural equation in which the right-hand side endogenous variable P_i is replaced by its predicted value $\hat{P}_i$[1]

In practice always use software that is designed for 2SLS, so that standard errors and t-values will be calculated correctly.

11.5.1 The General Two-Stage Least Squares Estimation Procedure

The two-stage least squares estimation procedure can be used to estimate the parameters of any identified equation within a simultaneous equations system. In a system of M simultaneous equations, let the endogenous variables be y_{i1}, y_{i2},…, y_{iM}. There must always be as many equations in a simultaneous system as there are endogenous variables. Let there be K exogenous variables, $x_{i1}, x_{i2},\ldots, x_{iK}$. To illustrate, suppose $M = 3$ and the first structural equation within this system is

$$y_{i1} = \alpha_2 y_{i2} + \alpha_3 y_{i3} + \beta_1 x_{i1} + \beta_2 x_{i2} + e_{i1} \tag{11.9}$$

If this equation is identified, then its parameters can be estimated in two steps:

1. Use OLS to estimate the parameters of the reduced-form equations

$$y_{i2} = \pi_{12} x_{i1} + \pi_{22} x_{i2} + \cdots + \pi_{K2} x_{iK} + v_{i2}$$

$$y_{i3} = \pi_{13} x_{i1} + \pi_{23} x_{i2} + \cdots + \pi_{K3} x_{iK} + v_{i3}$$

   Obtain the predicted values

$$\hat{y}_{i2} = \hat{\pi}_{12} x_{i1} + \hat{\pi}_{22} x_{i2} + \cdots + \hat{\pi}_{K2} x_{iK}$$

$$\hat{y}_{i3} = \hat{\pi}_{13} x_{i1} + \hat{\pi}_{23} x_{i2} + \cdots + \hat{\pi}_{K3} x_{iK} \tag{11.10}$$

2. Replace the endogenous variables, y_{i2} and y_{i3}, on the right-hand side of the structural (11.9) by their predicted values from (11.10)

$$y_{i1} = \alpha_2 \hat{y}_{i2} + \alpha_3 \hat{y}_{i3} + \beta_1 x_{i1} + \beta_2 x_{i2} + e_{i1}^*$$

   Estimate the parameters of this equation by OLS.

In practice, we should always use software designed for 2SLS or IV estimation. It will correctly carry out the calculations of the 2SLS estimates and their standard errors.

Equation (11.9) has two right-hand side endogenous variables and two exogenous variables. K is the total number of exogenous variables. How large must K be so that equation (11.9) is identified? The identification "necessary" condition is that in a system of M equations at

[1]The discussion above is an intuitive explanation of the two-stage least squares estimator. For a general explanation of this estimation method, see Section 10.3. There we derive the two-stage least squares estimator and discuss its properties.

least $M - 1$ variables that appear elsewhere in the system must be omitted from each equation. There are $M = 3$ equations so $M - 1 = 2$ variables must be omitted from each equation. Let $K = K_1 + K_1^*$, where $K_1 = 2$ is the number of included exogenous variables in the first structural equation, and K_1^* is the number of exogenous variables excluded from the first structural equation. Identification of the first equation requires $K_1^* \geq 2$ and $K \geq 4$. In Chapter 10's terminology, K_1^* is the number of instrumental variables for the first equation.

The alternative description of the condition for identification is that the number of omitted exogenous variables, K_1^*, must be greater than, or equal to, the number of included, right-hand side endogenous variables. Let $M = 1 + M_1 + M_1^*$, where $M_1 = 2$ is the number of included right-hand side endogenous variables, and M_1^* is the number of endogenous variables excluded from the first equation. In this example, $M_1^* = 0$ because the first equation contains all three endogenous variables, including the left-hand side variable y_1. The identification rule is that $K_1^* \geq M_1$. In Chapter 10's language, there must be as many instrumental variables, K_1^*, as endogenous variables on the right-hand side of the equation, M_1.

Remark

Simultaneous equations models were developed in the early 1940s and for many years were the cornerstone of econometric analysis. The subject of Chapter 10 is regression equations with endogenous variables, which can be thought of as one equation from a system of equations. Because building and estimating complete systems are difficult, more researchers in recent years have relied on estimating individual equations by $2SLS/IV$, which is why the content of Chapter 10 precedes this treatment of simultaneous equations. However, the concepts and methods used in Chapters 10 and 11 are the same. Just keep in mind that:

1. **Two-stage least squares** and **instrumental variables estimation** are identical.
2. **IVs**, or just **instruments**, are exogenous variables that do not appear in the equation. Instruments are **excluded exogenous variables**.
3. **The reduced-form equations** in simultaneous equations modeling are the **first-stage equations** in instrumental variables, two-stage least squares, estimation.

11.5.2 The Properties of the Two-Stage Least Squares Estimator

We have described how to obtain estimates for structural equation parameters in identified equations. The properties of the two-stage least squares estimator are as follows:

- The 2SLS estimator is a biased estimator, but it is consistent.
- In large samples the 2SLS estimator is approximately normally distributed.
- The variances and covariances of the 2SLS estimator are unknown in small samples, but for large samples, we have expressions for them that we can use as approximations. These formulas are built into econometric software packages, which report standard errors and t-values, just like an OLS regression program.
- If you obtain 2SLS estimates by applying two least squares regressions using OLS regression software, the standard errors and t-values reported in the *second* regression are *not* correct for the 2SLS estimator. Always use specialized 2SLS or IV software when obtaining estimates of structural equations.

EXAMPLE 11.1 | Supply and Demand for Truffles

Truffles are a gourmet delight. They are edible fungi that grow below the ground. In France they are often located by collectors who use pigs to sniff out the truffles and "point" to them. Actually the pigs dig frantically for the truffles because pigs have an insatiable taste for them, as do the French, and they must be restrained from "pigging out" on them. Consider a supply and demand model for truffles:

$$\text{Demand: } Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 PS_i + \alpha_4 DI_i + e_{di} \quad (11.11)$$

$$\text{Supply: } Q_i = \beta_1 + \beta_2 P_i + \beta_3 PF_i + e_{si} \quad (11.12)$$

In the demand equation Q is the quantity of truffles traded in a particular French marketplace, indexed by i, P is the market price of truffles, PS is the market price of a substitute for real truffles (another fungus much less highly prized), and DI is per capita monthly disposable income of local residents. The supply equation contains the market price and quantity supplied. Also it includes PF, the price of a factor of production, which in this case is the hourly rental price of truffle-pigs used in the search process. In this model, we assume that P and Q are endogenous variables. The exogenous variables are PS, DI, PF, and the intercept.

Identification

Before thinking about estimation, check the identification of each equation. The rule for identifying an equation is that in a system of M equations at least $M-1$ variables must be omitted from each equation in order for it to be identified. In the demand equation the variable PF is not included; thus the necessary $M-1=1$ variable is omitted. In the supply equation both PS and DI are absent; more than enough to satisfy the identification condition. Note too that the variables that are omitted are different for each equation, ensuring that each contains at least one *shift* variable not present in the other. We conclude that each equation in this system is identified and can thus be estimated by two-stage least squares.

Why are the variables omitted from their respective equations? Because economic theory says that the price of a factor of production should affect supply but not demand, and that the price of substitute goods and income should affect demand and not supply. The specifications we used are based on the microeconomic theory of supply and demand.

The reduced-form equations

The reduced-form equations express each endogenous variable, P and Q, in terms of the exogenous variables PS, DI, PF, and the intercept, plus an error term. They are

$$Q_i = \pi_{11} + \pi_{21} PS_i + \pi_{31} DI_i + \pi_{41} PF_i + v_{i1}$$

$$P_i = \pi_{12} + \pi_{22} PS_i + \pi_{32} DI_i + \pi_{42} PF_i + v_{i2}$$

We can estimate these equations by OLS since the right-hand side variables are exogenous and contemporaneously uncorrelated with the random errors v_{i1} and v_{i2}. The data file *truffles* contains 30 observations on each of the endogenous and exogenous variables. The units of measurement are \$ per ounce for price P, ounces for Q, \$ per ounce for PS, and thousands of dollars for DI; PF is the hourly rental rate (\$) for a truffle-finding pig. A few of the observations are shown in Table 11.1. The results of the least squares estimations of the reduced-form equations for Q and P are reported in Tables 11.2a and 11.2b.

In Table 11.2a, we see that the estimated coefficients are statistically significant, and thus we conclude that the exogenous variables affect the quantity of truffles traded, Q, in this reduced-form equation. The $R^2 = 0.697$, and the overall F-statistic is 19.973, which has a p-value of less than 0.0001. In Table 11.2b the estimated coefficients

TABLE 11.1 Representative Truffle Data

OBS	P	Q	PS	DI	PF
1	29.64	19.89	19.97	2.103	10.52
2	40.23	13.04	18.04	2.043	19.67
3	34.71	19.61	22.36	1.870	13.74
4	41.43	17.13	20.87	1.525	17.95
5	53.37	22.55	19.79	2.709	13.71
Summary Statistics					
Mean	62.72	18.46	22.02	3.53	22.75
Std. Dev.	18.72	4.61	4.08	1.04	5.33

TABLE 11.2a Reduced Form for Quantity of Truffles (Q)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.8951	3.2434	2.4342	0.0221
PS	0.6564	0.1425	4.6051	0.0001
DI	2.1672	0.7005	3.0938	0.0047
PF	−0.5070	0.1213	−4.1809	0.0003

TABLE 11.2b Reduced Form for Price of Truffles (P)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−32.5124	7.9842	−4.0721	0.0004
PS	1.7081	0.3509	4.8682	0.0000
DI	7.6025	1.7243	4.4089	0.0002
PF	1.3539	0.2985	4.5356	0.0001

are statistically significant, indicating that the exogenous variables have an effect on market price P. The $R^2 = 0.889$ implies a good fit of the reduced-form equation to the data. The overall F-statistic value is 69.189 that has a p-value of less than 0.0001, indicating that the model has statistically significant explanatory power.

The structural equations

The reduced-form equations are used to obtain $\hat{P}$ that will be used in place of P on the right-hand side of the supply and demand equations in the second stage of two-stage least squares. From Table 11.2b, we have

$$\hat{P}_i = \hat{\pi}_{12} + \hat{\pi}_{22}PS_i + \hat{\pi}_{32}DI_i + \hat{\pi}_{42}PF_i$$
$$= -32.512 + 1.708PS_i + 7.603DI_i + 1.354PF_i$$

The 2SLS results are given in Tables 11.3a and 11.3b. The estimated demand curve results are in Table 11.3a. Note that the coefficient of price is negative, indicating that as the market price rises, the quantity demanded of truffles declines, as predicted by the law of demand. The standard errors that are reported are obtained from 2SLS software. They and the t-values are valid in large samples. The p-value indicates that the estimated slope of the demand curve is significantly different from zero. Increases in the price of the substitute for truffles increase the demand for truffles, which is a characteristic of substitute goods. Finally the effect of income is positive, indicating that truffles are a normal good. All of these variables have statistically significant coefficients and thus have an effect upon the quantity demanded.

TABLE 11.3a **2SLS Estimates for Truffle Demand**

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−4.2795	5.5439	−0.7719	0.4471
P	−0.3745	0.1648	−2.2729	0.0315
PS	1.2960	0.3552	3.6488	0.0012
DI	5.0140	2.2836	2.1957	0.0372

TABLE 11.3b **2SLS Estimates for Truffle Supply**

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	20.0328	1.2231	16.3785	0.0000
P	0.3380	0.0249	13.5629	0.0000
PF	−1.0009	0.0825	−12.1281	0.0000

The supply equation results appear in Table 11.3b. As anticipated, increases in the price of truffles increase the quantity supplied, and increases in the rental rate for truffle-seeking pigs, which is an increase in the cost of a factor of production, reduces supply. Both of these variables have statistically significant coefficient estimates.

EXAMPLE 11.2 | Supply and Demand at the Fulton Fish Market

The Fulton Fish Market has operated in New York City for over 150 years. The prices for fish are determined daily by the forces of supply and demand. Kathryn Graddy[2] collected daily data on the price of whiting (a common type of fish), quantities sold, and weather conditions during the period December 2, 1991, to May 8, 1992. These data are in the file *fultonfish*. Fresh fish arrive at the market about midnight. The wholesalers, or dealers, sell to buyers for retail shops and restaurants. The first interesting feature of this example is to consider whether prices and quantities are *simultaneously* determined by supply and demand at all.[3] We might consider this a market with a fixed, perfectly inelastic supply. At the start of the day, when the market is opened, the supply of fish available for the day is fixed. If supply is fixed, with a vertical supply curve, then price is demand-determined, with higher demand leading to higher prices but no increase in the quantity supplied. If this is true, then the feedback between prices and quantities is eliminated. Such models are said to be **recursive** and the demand equation can be estimated by OLS rather than the more complicated two-stage least squares procedure.

However whiting fish can be kept for several days before going bad, and dealers can decide to sell less, and add to their inventory, or buffer stock, if the price is judged too low, in hope for better prices the next day. Or, if the price is unusually high on a given day, then sellers can increase the day's catch with additional fish from their buffer stock. Thus despite the perishable nature of the product, and the daily resupply of fresh fish, daily price is simultaneously determined by supply and demand forces. The key point here is that "simultaneity" does not require that events occur at a simultaneous moment in time.

[2] See Kathryn Graddy (2006), "The Fulton Fish Market," *Journal of Economic Perspectives*, 20(2), 207–220.

[3] See Kathryn Graddy and Peter E. Kennedy (2010), "When Are Supply and Demand Determined Recursively Rather than Simultaneously?," *Eastern Economic Journal*, 36, 188–197.

Let us specify the demand equation for this market as

$$\ln(QUAN_t) = \alpha_1 + \alpha_2 \ln(PRICE_t) + \alpha_3 MON_t + \alpha_4 TUE_t + \alpha_5 WED_t + \alpha_6 THU_t + e_{dt} \quad (11.13)$$

where $QUAN_t$ is the quantity sold, in pounds, and $PRICE_t$ is the average daily price per pound. Note that we are using the subscript "t" to index observations for this relationship because of the time series nature of the data. The remaining variables are indicator variables for the days of the week, with Friday being omitted. The coefficient α_2 is the price elasticity of demand, which we expect to be negative. The daily indicator variables capture day-to-day shifts in demand. The supply equation is

$$\ln(QUAN_t) = \beta_1 + \beta_2 \ln(PRICE_t) + \beta_3 STORMY_t + e_{st} \quad (11.14)$$

The coefficient β_2 is the price elasticity of supply. The variable *STORMY* is an indicator variable indicating stormy weather during the previous three days. This variable is important in the supply equation because stormy weather makes fishing more difficult, reducing the supply of fish brought to market.

Identification

Prior to estimation, we must determine whether the supply and demand equation parameters are identified. The necessary condition for an equation to be identified is that in this system of $M = 2$ equations, it must be true that at least $M - 1 = 1$ variable must be omitted from each equation. In the demand equation the weather variable *STORMY* is omitted, and it does appear in the supply equation. In the supply equation, the four daily indicator variables that are included in the demand equation are omitted. Thus the demand equation shifts daily, while the supply remains fixed (since the supply equation does not contain the daily indicator variables), thus tracing out the supply curve, making it identified, as shown in Figure 11.4. Similarly, stormy conditions shift the supply curve relative to a fixed demand, tracing out the demand curve and making it identified.

The reduced-form equations

The reduced-form equations specify each endogenous variable as a function of all exogenous variables

$$\ln(QUAN_t) = \pi_{11} + \pi_{21} MON_t + \pi_{31} TUE_t + \pi_{41} WED_t + \pi_{51} THU_t + \pi_{61} STORMY_t + v_{t1} \quad (11.15)$$

$$\ln(PRICE_t) = \pi_{12} + \pi_{22} MON_t + \pi_{32} TUE_t + \pi_{42} WED_t + \pi_{52} THU_t + \pi_{62} STORMY_t + v_{t2} \quad (11.16)$$

These reduced-form equations can be estimated by OLS because the right-hand side variables are all exogenous and uncorrelated with the reduced-form errors v_{t1} and v_{t2}. Using the Graddys' data (*fultonfish*), we estimate these reduced-form equations and report them in Tables 11.4a and 11.4b. Estimation of the reduced-form equations is the first step of two-stage least squares estimation of the supply and demand equations. It is a requirement for successful two-stage least squares estimation that the estimated coefficients in the reduced form for the right-hand side endogenous variable be statistically significant. We have specified the structural equations (11.13) and (11.14) with $\ln(QUAN_t)$ as the left-hand side variable and $\ln(PRICE_t)$ as the right-hand side endogenous variable. Thus the key reduced-form equation is (11.16) for $\ln(PRICE_t)$. In this equation

- To identify the supply curve, the daily indicator variables must be jointly significant. This implies that at least one of their coefficients is statistically different from zero, meaning that there is at least one significant shift variable in the demand equation, which permits us to reliably estimate the supply equation.
- To identify the demand curve, the variable $STORMY_t$ must be statistically significant, meaning that supply has a significant shift variable, so that we can reliably estimate the demand equation.

Why is this so? The identification discussion in Section 11.4 requires only the presence of shift variables, not their significance. The answer comes from a great deal of econometric research in the past decade, which shows that the two-stage least squares estimator performs very poorly if the shift variables are not strongly significant.[4] Recall that to implement two-stage least squares we take the predicted value from the reduced-form regression and include it in the structural equations in place of the right-hand side endogenous variable, that is, we calculate

$$\widehat{\ln(PRICE_t)} = \hat{\pi}_{12} + \hat{\pi}_{22} MON_t + \hat{\pi}_{32} TUE_t + \hat{\pi}_{42} WED_t + \hat{\pi}_{52} THU_t + \hat{\pi}_{62} STORMY_t$$

where $\hat{\pi}_{k2}$ are the least squares estimates of the reduced-form coefficients, and then replace $\ln(PRICE_t)$ with $\widehat{\ln(PRICE_t)}$. To illustrate our point, let us focus on the problem of estimating the supply equation (11.14) and take the extreme case that $\hat{\pi}_{22} = \hat{\pi}_{32} = \hat{\pi}_{42} = \hat{\pi}_{52} = 0$, meaning that the coefficients on the daily indicator variables are all identically zero. Then

$$\widehat{\ln(PRICE_t)} = \hat{\pi}_{12} + \hat{\pi}_{62} STORMY_t$$

If we replace $\ln(PRICE_t)$ in the supply equation (11.14) with this predicted value, there will be *exact* collinearity between $\widehat{\ln(PRICE_t)}$ and the variable $STORMY_t$, which is already in the supply equation, and two-stage least squares will fail. If the coefficient estimates on the daily indicator

[4] See Section 10.3.9 for further discussion of this point.

variables are not exactly zero, but are jointly insignificant, it means there will be severe collinearity in the second stage, and although the two-stage least squares estimates of the supply equation can be computed, they will be unreliable. In Table 11.4b, showing the reduced-form estimates for (11.16), none of the daily indicator variables are statistically significant. Also, the joint F-test of significance of the daily indicator variables has p-value 0.65, so that we cannot reject the null hypothesis that all these coefficients are zero.[5] In this case the supply equation is not identified in practice, and we will not report estimates for it.

TABLE 11.4a Reduced Form for ln(Quantity) Fish

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.8101	0.1470	59.9225	0.0000
STORMY	−0.3878	0.1437	−2.6979	0.0081
MON	0.1010	0.2065	0.4891	0.6258
TUE	−0.4847	0.2011	−2.4097	0.0177
WED	−0.5531	0.2058	−2.6876	0.0084
THU	0.0537	0.2010	0.2671	0.7899

TABLE 11.4b Reduced Form for ln(Price) Fish

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−0.2717	0.0764	−3.5569	0.0006
STORMY	0.3464	0.0747	4.6387	0.0000
MON	−0.1129	0.1073	−1.0525	0.2950
TUE	−0.0411	0.1045	−0.3937	0.6946
WED	−0.0118	0.1069	−0.1106	0.9122
THU	0.0496	0.1045	0.4753	0.6356

However, $STORMY_t$ is statistically significant in Table 11.4b, meaning that the demand equation may be reliably estimated by two-stage least squares. An advantage of two-stage least squares estimation is that each equation can be treated and estimated separately, so the fact that the supply equation is not reliably estimable does not mean that we cannot proceed with estimation of the demand equation. The check of statistical significance of the sets of shift variables for the structural equations should be carried out each time a simultaneous equations model is formulated.

Two-stage least squares estimation of fish demand

Applying two-stage least squares estimation to the demand equation we obtain the results as given in Table 11.5. The price elasticity of demand is estimated to be −1.12, meaning that a 1% increase in fish price leads to about a 1.12% decrease in the quantity demanded; this estimate is statistically significant at the 5% level. The indicator variable coefficients are negative and statistically significant for Tuesday and Wednesday, meaning that demand is lower on these days relative to Friday.

TABLE 11.5 2SLS Estimates for Fish Demand

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.5059	0.1662	51.1890	0.0000
ln(*PRICE*)	−1.1194	0.4286	−2.6115	0.0103
MON	−0.0254	0.2148	−0.1183	0.9061
TUE	−0.5308	0.2080	−2.5518	0.0122
WED	−0.5664	0.2128	−2.6620	0.0090
THU	0.1093	0.2088	0.5233	0.6018

EXAMPLE 11.3 | Klein's Model I

One of the most widely used econometric examples in the past 50 years is the small, three equation, macroeconomic model of the U.S. economy proposed by Lawrence Klein, the 1980 Nobel Prize winner in Economics.[6] The model has three equations, which are estimated, and then a number of macroeconomic identities, or definitions, to complete the model. In all, there are eight endogenous variables and eight exogenous variables.

[5] Even if the variables are jointly significant, there may be a problem. The significance must be "strong." An F-value <10 is cause for concern. This problem is the same as that of weak instruments in instrumental variables estimation (see Section 10.3.9).

[6] Our presentation follows Ernst R. Berndt (1991), *The Practice of Econometrics: Classic and Contemporary*, Addison-Wesley Publishing, Section 10.5.

The first equation is a consumption function, in which aggregate consumption in year t, CN_t is related to total wages earned by all workers, W_t. Total wages are divided into wages of workers earned in the private sector, W_{1t}, and wages of workers earned in the public sector, W_{2t}, so that total wages $W_t = W_{1t} + W_{2t}$. Private sector wages W_{1t} are endogenous and determined within the structure of the model, as we will see below. Public sector wages W_{2t} are exogenous. In addition, consumption expenditures are related to nonwage income (profits) in the current year, P_t, which are endogenous, and profits from the previous year, P_{t-1}. Thus, the consumption function is

$$CN_t = \alpha_1 + \alpha_2\left(W_{1t} + W_{2t}\right) + \alpha_3 P_t + \alpha_4 P_{t-1} + e_{1t} \quad (11.17)$$

Now refer back to equation (5.44) in Section 5.7.3. There we introduced the term **contemporaneously uncorrelated** to describe the situation in which an explanatory variable observed at time t, x_{tk} is uncorrelated with the random error at time t, e_t. In the terminology of Chapter 10, the variable x_{tk} is **exogenous** if it is contemporaneously uncorrelated with the random error e_t. And the variable x_{tk} is **endogenous** if it is contemporaneously correlated with the random error e_t. In the consumption equation, W_{1t} and P_t are endogenous and contemporaneously correlated with the random error e_t. On the other hand, wages in the public sector, W_{2t}, are set by public authority and are assumed exogenous and uncorrelated with the current period random error e_{1t}. What about profits in the previous year, P_{t-1}? They are **not** correlated with the random error occurring one year later. Lagged endogenous variables are called **predetermined variables** and are treated just like exogenous variables.

The second equation in the model is the investment equation. Net investment, I_t, is specified to be a function of current and lagged profits, P_t and P_{t-1}, as well as the capital stock at the end of the previous year, K_{t-1}. This lagged variable is predetermined and treated as exogenous. The investment equation is

$$I_t = \beta_1 + \beta_2 P_t + \beta_3 P_{t-1} + \beta_4 K_{t-1} + e_{2t} \quad (11.18)$$

Finally, there is an equation for wages in the private sector, W_{1t}. Let $E_t = CN_t + I_t + \left(G_t - W_{2t}\right)$, where G_t is government spending. Consumption and investment are endogenous and government spending and public sector wages are exogenous. The sum, E_t, total national product minus public sector wages, is endogenous. Wages are taken to be related to E_t and the predetermined variable E_{t-1}, plus a time trend variable, $TIME_t = YEAR_t - 1931$, which is exogenous. The wage equation is

$$W_{1t} = \gamma_1 + \gamma_2 E_t + \gamma_3 E_{t-1} + \gamma_4 TIME_t + e_{3t} \quad (11.19)$$

Because there are eight endogenous variables in the entire system there must also be eight equations. Any system of M endogenous variables must have M equations to be complete. In addition to the three equations (11.17)–(11.19), which contain five endogenous variables, there are five other definitional equations to complete the system that introduce three further endogenous variables. In total, there are eight exogenous and predetermined variables, which can be used as IVs. The exogenous variables are government spending, G_t, public sector wages, W_{2t}, taxes, TX_t, and the time trend variable, $TIME_t$. Another exogenous variable is the constant term, the "intercept" variable in each equation, $X_{1t} \equiv 1$. The predetermined variables are lagged profits, P_{t-1}, the lagged capital stock, K_{t-1}, and the lagged total national product minus public sector wages, E_{t-1}.

11.6 Exercises

11.6.1 Problems

11.1 Our aim is to estimate the parameters of the simultaneous equations model

$$y_1 = \alpha_1 y_2 + e_1$$
$$y_2 = \alpha_2 y_1 + \beta_1 x_1 + \beta_2 x_2 + e_2$$

We assume that x_1 and x_2 are exogenous and uncorrelated with the error terms e_1 and e_2.

a. Solve the two structural equations for the reduced-form equation for y_2, that is, $y_2 = \pi_1 x_1 + \pi_2 x_2 + v_2$. Express the reduced-form parameters in terms of the **structural parameters** and the reduced-form error in terms of the structural parameters and e_1 and e_2. Show that y_2 is correlated with e_1.

b. Which equation parameters are consistently estimated using OLS? Explain.

c. Which parameters are "identified," in the simultaneous equations sense? Explain your reasoning.

d. To estimate the parameters of the reduced-form equation for y_2 using the method of moments (MOM), which was introduced in Section 10.3, the two moment equations are

$$N^{-1}\sum x_{i1}(y_2 - \pi_1 x_{i1} - \pi_2 x_{i2}) = 0$$
$$N^{-1}\sum x_{i2}(y_2 - \pi_1 x_{i1} - \pi_2 x_{i2}) = 0$$

Explain why these two moment conditions are a valid basis for obtaining consistent estimators of the reduced-form parameters.

e. Are the MOM estimators in part (d) the same as the OLS estimators? Form the sum of squared errors function for $y_2 = \pi_1 x_1 + \pi_2 x_2 + v_2$ and find the first derivatives. Set these to zero and show that they are equivalent to the two equations in part (d).

f. Using $\sum x_{i1}^2 = 1$, $\sum x_{i2}^2 = 1$, $\sum x_{i1}x_{i2} = 0$, $\sum x_{i1}y_{1i} = 2$, $\sum x_{i1}y_{2i} = 3$, $\sum x_{i2}y_{1i} = 3$, $\sum x_{i2}y_{2i} = 4$, and the two moment conditions in part (d) show that the MOM/OLS estimates of π_1 and π_2 are $\hat{\pi}_1 = 3$ and $\hat{\pi}_2 = 4$.

g. The fitted value $\hat{y}_2 = \hat{\pi}_1 x_1 + \hat{\pi}_2 x_2$. Explain why we can use the moment condition $\sum \hat{y}_{i2}(y_{i1} - \alpha_1 y_{i2}) = 0$ as a valid basis for consistently estimating α_1. Obtain the IV estimate of α_1.

h. Find the 2SLS estimate of α_1 by applying OLS to $y_1 = \alpha_1 \hat{y}_2 + e_1^*$. Compare your answer to that in part (g).

11.2 Consider a supply and demand model written in its most general implicit form, using capital Greek letters for the unknown parameters and E_i for the random errors,

$$\text{Demand: } \Gamma_{11}q + \Gamma_{21}p + \mathrm{B}_{11} + \mathrm{B}_{21}x + E_1 = 0$$
$$\text{Supply: } \Gamma_{12}q + \Gamma_{22}p + \mathrm{B}_{12} + \mathrm{B}_{22}x + E_2 = 0$$

a. Multiply each equation by 3. Do they remain true?

b. Multiply the demand equation by $-1/\Gamma_{11}$. Does it remain true?

c. Define $\alpha_{21} = -\Gamma_{21}/\Gamma_{11}$, $\beta_{11} = -\mathrm{B}_{11}/\Gamma_{11}$, $\beta_{21} = -\mathrm{B}_{21}/\Gamma_{11}$, $e_1 = -E_1/\Gamma_{11}$ and write the demand equation with q on the left-hand side and the remaining terms on the right-hand side. By choosing q to be on the left-hand side of the equation, we have chosen a **normalization rule**.

d. Repeat the process for the supply equation, beginning by multiplying through by $-1/\Gamma_{22}$, and obtain the normalized supply curve with

$$\alpha_{12} = -\Gamma_{12}/\Gamma_{22},\ \beta_{12} = -\mathrm{B}_{12}/\Gamma_{22},\ \beta_{22} = -\mathrm{B}_{22}/\Gamma_{22},\ \text{and}\ e_2 = -E_2/\Gamma_{22}$$

Write the normalized supply equation with p on the left-hand side and the remaining terms on the right side.

e. Mathematically, in a system of jointly determined variables, it does not matter which variable appears on the left side of each normalized equation. True or false?

11.3 Consider a supply and demand model written in its most general implicit form, using capital Greek letters for the unknown parameters and E_i for the random errors:

$$\text{Demand: } \Gamma_{11}q + \Gamma_{21}p + \mathrm{B}_{11} + \mathrm{B}_{21}x + E_1 = 0$$
$$\text{Supply: } \Gamma_{12}q + \Gamma_{22}p + \mathrm{B}_{12} + \mathrm{B}_{22}x + E_2 = 0$$

a. Find the reduced-form equation for p, $p = \pi_1 + \pi_2 x + v$. Express π_1 and π_2 in terms of parameters Γ_{ij} and B_{ij}.

b. Suppose we replace the "true" demand equation with an equation that is a mixture of the demand and supply equations, that is, multiply through the demand equation by 3 and the supply equation by 2 and then add the two equations together to obtain

$$(3\Gamma_{11} + 2\Gamma_{12})q + (3\Gamma_{21} + 2\Gamma_{22})p + (3\mathrm{B}_{11} + 2\mathrm{B}_{12}) + (3\mathrm{B}_{21} + 2\mathrm{B}_{22})x + (3E_1 + 2E_2) = 0$$

or $\Gamma'_{11}q + \Gamma'_{21}p + \mathrm{B}'_{11} + \mathrm{B}'_{21}x + E'_1 = 0$, with $'$ denoting the new parameters. Using the new demand equation, and the original supply equation, find the reduced-form equation for p, $p = \pi_1^* + \pi_2^* x + v^*$. Express π_1^* and π_2^* in terms of parameters Γ_{ij} and B_{ij}. Compare the solution to that in (a).

11.4 Consider the supply and demand model below:

$$\text{Demand: } q = -p + 3 + 2x + e_1$$
$$\text{Supply: } p = q + 1 + x + e_2$$

a. Find the reduced-form equations for p and q as a function of the exogenous variable x.
b. Now suppose that the demand equation is $q = -5p + 11 + 8x + e_1^*$. Find the reduced-form equations for p and q using this demand equation and the original supply equation.
c. Show that the new demand equation is a mixture of the original supply and demand equations. Specifically, it is three times the original demand equation plus two times the supply equation. [*Hint*: It is simpler to put the demand and supply equations into implicit form, with everything on the left side and zero on the right side, before doing the multiplying and adding.]
d. If we have N observations on p, q, and x, can we consistently estimate the demand equation by OLS? Why?
e. If we have N observations on p, q, and x, can we consistently estimate the reduced-form equations by OLS? Why?
f. Given the true reduced-form equations, can we deduce whether $q = -p + 3 + 2x + e_1$ or $q = -5p + 11 + 8x + e_1^*$ is the true demand equation?
g. Is the demand equation "identified" using the necessary condition?

11.5 Consider the supply and demand model below:

$$\text{Demand: } q = -p + 3 + 2x + e_1$$
$$\text{Supply: } p = q + 1 + e_2$$

a. Find the reduced-form equations for p and q as a function of the exogenous variable x.
b. Now suppose that the demand equation is $q = -5p + 11 + 6x + e_1^*$. Find the reduced-form equations for p and q using this demand equation and the original supply equation.
c. Show that the new demand equation is a mixture of the original supply and demand equations. Specifically, it is three times the original demand equation plus two times the supply equation. [*Hint*: It is simpler to put the demand and supply equations into implicit form, with everything on the left side and zero on the right side, before doing the multiplying and adding.]
d. If we have N observations on p, q, and x, can we consistently estimate the supply equation by OLS? Why?
e. If we have N observations on p, q, and x, can we consistently estimate the reduced-form equations by OLS? Why?
f. Given the economic supply and demand model proposed in the question, is it possible for the mixture equation $q = -5p + 11 + 6x + e_1^*$ to be a supply curve? Explain.
g. Is the demand equation "identified" using the necessary condition? Is the supply equation "identified" using the necessary condition?

11.6 Consider the supply and demand model below, where x is exogenous.

$$\text{Demand: } q = \alpha_1 p + \alpha_2 + \alpha_3 x + e_1$$
$$\text{Supply: } p = \beta_1 q + \beta_2 + e_2$$

a. Find the reduced-form equations for p and q, $q = \pi_{11} + \pi_{21}x + v_1$ and $p = \pi_{12} + \pi_{22}x + v_2$, expressing the reduced-form parameters in terms of α's and β's.
b. Suppose $\pi_{11} = 1/5$, $\pi_{21} = 3/5$, $\pi_{12} = 2/5$, and $\pi_{22} = 6/5$. Solve for as many of the α's and β's as you can.

11.7 Consider the supply and demand model below, where x and w are exogenous.

$$\text{Demand: } q = \alpha_1 p + \alpha_1 x + \alpha_2 w + e_1$$
$$\text{Supply: } p = \beta_1 q + e_2$$

a. Find the reduced-form equations for p and q, $q = \pi_{11}x + \pi_{21}w + v_1$ and $p = \pi_{12}x + \pi_{22}w + v_2$, expressing the reduced-form parameters in terms of α's and β's.
b. Suppose $\pi_{11} = 1/5$, $\pi_{21} = 1/5$, $\pi_{12} = 2/5$, and $\pi_{22} = 2/5$. Solve for as many of the α's and β's as you can.

11.8 In macroeconomics, the simple "consumption function" relates national expenditure on consumption goods, $CONSUMP_t$ = aggregate consumption, in period t to national income, $INCOME_t = GNP_t$. Specify the consumption function $CONSUMP_t = \beta_1 + \beta_2 INCOME_t + e_t$. Suppose that INV_t is aggregate investment. In the simplest model, the income identity is $INCOME_t = CONSUMP_t + INV_t$.

a. Substitute the income identity into the consumption function and solve for consumption in terms of investment.

b. Find the covariance between $INCOME_t$ and the random error e_t.

c. Find the covariance between INV_t and $INCOME_t$.

d. Suppose INV_t is uncorrelated with the random error e_t. Does it satisfy the conditions for an IV?

11.9 Consider the simultaneous equations model, where x is exogenous.

$$y_{i1} = \alpha_1 y_{i2} + \alpha_2 x_{i1} + e_{i1}$$
$$y_{i2} = \alpha_2 y_{i1} + \beta_1 x_{i1} + e_{i2}$$

Assume that $E(e_{i1}|\mathbf{x}_1) = E(e_{i2}|\mathbf{x}_1) = 0$, $\text{var}(e_{i1}|\mathbf{x}_1) = \sigma_1^2$, $\text{var}(e_{i2}|\mathbf{x}_1) = \sigma_2^2$, and $\text{cov}(e_{i1}, e_{i2}|\mathbf{x}_1) = \sigma_{12}$.

a. Substitute the second equation into the first and find the reduced-form equation for y_{i1}.

b. Multiply the reduced-form equation for y_{i1} from part (a) by e_{i2} and find $\text{cov}(y_{i1}, e_{i2}|\mathbf{x}_1) = E(y_{i1}e_{i2}|\mathbf{x}_1)$.

c. Show that $\text{cov}(y_{i1}, e_{i2}|\mathbf{x}_1) = 0$ if $\alpha_1 = 0$ and $\sigma_{12} = 0$. Such a system is said to be **recursive**.

d. Is the OLS estimator of the first equation consistent under the conditions in (c)? Explain.

e. Is the OLS estimator of the second equation consistent under the conditions in (c)? Explain.

11.10 Reconsider the Truffle supply and demand model in Example 11.1. Modify the demand equation as $Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 PS_i + e_i^d$, keeping the supply equation unchanged. The estimates are given in Table 11.6.

TABLE 11.6 Estimates for Exercise 11.10

	(1)	(2)	(3)	(4)
C	5.6169	−40.5043	0.4460	19.9625
	(3.6256)	(10.0873)	(4.1596)	(1.2371)
PF	−0.2762	2.1635		−1.0425
	(0.1097)	(0.3053)		(0.0907)
PS	0.8685	2.4522	1.1815	
	(0.1434)	(0.3991)	(0.2765)	
P			−0.1277	0.3542
			(0.0671)	(0.0288)

Standard errors in parentheses.

a. Are the demand and supply equations identified using the necessary condition in Section 11.4? Explain.

b. Column (1) contains the OLS estimates of the reduced-form equation for Q, and column (2) contains the OLS estimates of the reduced-form equation for P. Compute the first-stage F-test used to decide upon instrument strength for each equation. Is the F-value greater than the rule of thumb threshold, $F > 10$? [*Hint*: Recall the relationship between t- and F-tests.]

c. Using the estimates accurately sketch the supply and demand equations, with Q on the vertical axis and P on the horizontal axis. For these sketches set the values of the exogenous variables, PS and PF, to be $PF^* = 23$ and $PS^* = 22$.

d. What are the equilibrium values of P and Q from (c)?

e. On the graph from part (c) show the consequences of increasing the price of the factor of production (the truffle-seeking pig's rental rate) from $PF^* = 23$ to $PF^* = 30$, holding the value of PS constant.

f. Calculate the change in equilibrium price P and quantity Q in (d). What is the percentage change in equilibrium quantity divided by the percentage change in PF?

g. Calculate a 95% interval estimate for the elasticity of Q with respect to PF using the reduced-form equation estimates, at $PF^* = 23$ and $PS^* = 22$. Is the elasticity in (f) within the 95% interval estimate?

11.11 Reconsider the Truffle supply and demand model in Example 11.1. Suppose we modify the supply equation to be $Q_i = \beta_1 + \beta_2 P_i + e_i^s$, keeping the demand equation unchanged.

a. Are the supply and demand equations identified using the necessary condition in Section 11.4? Explain.

b. The estimated first-stage, reduced form, equation becomes

$$\underset{(t)}{\hat{P}_i = -13.50 + 1.47PS_i + 12.41DI_i} \qquad F = 54.21$$
$$(3.23) \quad (6.95)$$

Do you judge the omitted exogenous variables (instruments) strong enough to estimate the identified equation(s)? Explain.

c. The estimated supply equation using 2SLS is

$$\underset{(\text{se})}{\hat{Q}_i = 8.6455 + 0.1564P_i}$$
$$(2.89) \quad (0.045)$$

Verify that the point of the means (see Table 11.1) falls on the estimated supply curve.

d. Calculate the price elasticity of supply at the means and compare it to the elasticity computed from the 2SLS estimates in Table 11.3b.

e. Comparing the results in parts (b) and (c) to those in Example 11.1, do you think we should include PF in the supply equation? Explain.

11.12 Suppose you want to estimate a wage equation for married women of the form

$$\ln(WAGE) = \beta_1 + \beta_2 HOURS + \beta_3 EDUC + \beta_4 EXPER + \beta_5 EXPER^2 + e_1$$

where $WAGE$ is the hourly wage, $HOURS$ is number of hours worked per week, $EDUC$ is years of education, and $EXPER$ is years of experience. Your classmate observes that higher wages can bring forth increased work effort, and that married women with small children may reduce their hours of work to take care of them. It may also be true that a husband's wage rate has an effect on a wife's hours of work supplied, so that there may be an auxiliary relationship such as

$$HOURS = \alpha_1 + \alpha_2 \ln(WAGE) + \alpha_3 KIDS + \alpha_4 \ln(HWAGE) + e_2$$

where $KIDS$ is the number of children under the age of six in the woman's household and $HWAGE$ is her husband's wage rate.

a. Can the wage equation be estimated satisfactorily using the OLS estimator? If not, why not?

b. Is the wage equation "identified"? What does the term *identification* mean in this context?

c. If you seek an alternative to least squares estimation for the wage equation, suggest an estimation procedure and how (step by step, and NOT a computer command) it is carried out.

d. Other than the identification condition in part (b), are there any other conditions that must be met so that we can confidently use the estimation procedure in part (c)? What are those conditions?

11.13 In the post-World War II period, monetary policy effects and the supply and demand for money were important topics. Consider the following model, where M is the money stock, R is short-term rate of interest, GNP is national income, and R_d is Federal Reserve's discount rate, which it charges commercial banks. The endogenous variables are the money supply, M, and the short-term rate of interest, R. The exogenous variables are GNP and the Federal Reserve's discount rate, R_d. The lagged money stock M_{t-1} is a **predetermined** variable. It is treated as an exogenous variable and uncorrelated with the current period error. Using quarterly data from the post-war period, the 2SLS estimated money demand, omitting seasonal and other dummy variables, is

$$\underset{(\text{se})}{\hat{M}_t = 23.06 + 0.0618GNP_t - 0.0025\left(R \times GNP_t\right) + 0.686M_{t-1} + \cdots}$$
$$(0.0126) \quad (0.0007) \quad (0.0728)$$

The supply equation is taken to be proportional to the difference between the short-term interest rate R and the discount rate, R_d, with the factor of proportionality being the maximum potential money

stock, M^*, which is a known constant. The estimated supply equation, omitting seasonal and other dummy variables, is

$$\hat{M}_t = 0.8522 + 0.0751\left[M_t^*\left(R_t - R_{d,t}\right)\right] + \cdots$$
$$\text{(se)} \qquad\qquad (0.0159)$$

a. If the preceding period's money supply increases by one unit, what happens to the money demand function? In a graph like Figure 11.4, with M on the vertical axis and R on the horizontal axis, does the money demand curve shift right or left or not at all? Does the money supply curve shift if $\Delta M_{t-1} > 0$? If so, which direction?

b. If GNP increases by one unit, what happens to the money demand function? In a graph like Figure 11.4, does the money demand curve shift right or left or not at all? Does the money supply curve shift if $\Delta GNP_t > 0$? If so, which direction?

c. If the discount rate, R_d, is increased does the money demand curve shift right or left or not at all? Does the money supply curve shift if $\Delta R_{d,t} > 0$? If so, which direction?

d. Explain how your answers to (a), (b), and (c) imply that both the supply and demand for money functions are identified.

11.14 Australian wine is popular in Australia and worldwide. Using annual data on wine grape transactions Q(10,000 tonne units) and price P(\$AU100 per tonne) of wine produced in warm inland Australia, an estimated demand equation is

$$\hat{Q}_t = -0.278P_t + 2.884INCOME_t - 3.131XRATE_t - 2.766STOCKS_{t-1} + \cdots$$
$$(t) \quad (-2.85) \qquad (6.34) \qquad\quad (-3.04) \qquad\quad (-2.24)$$

INCOME (US\$1,000,000) is weighted household consumption expenditure, *XRATE* is the exchange rate per \$AU, and *STOCKS* (1000 million litres) are from the previous year. An estimated supply equation is

$$\hat{Q}_t = 0.824P_t + 0.682Q_{t-4} + 0.598TIME_t - 1.688TEMP_t + 1.793NON_{t-4} - 1.570PREM_{t-4} + \cdots$$
$$(t) \quad (4.82) \qquad (3.68) \qquad (5.87) \qquad (-1.19) \qquad\quad (4.21) \qquad\quad (-2.42)$$

TEMP is mean January temperature (mid-summer "Down Under"), NON_{t-4} is the price of the regional wine grape relative to other non-premium grapes, lagged four years, and $PREM_{t-4}$ is the price of the regional wine grapes relative to other premium wine grapes, lagged four years. Production at time $t-4$ is on the right-hand side reflecting the four years required between planting grape vines and producing wine. This is a **partial adjustment** model as discussed in Exercise 9.30. In both equations, we have omitted the intercept and indicator variables for specific regions.

a. Which variables in the model cause the demand equation to shift relative to the supply equation?

b. Which variables in the model cause the supply equation to shift relative to the demand equation?

c. Discuss the signs of the estimated coefficients in the demand equation.

d. Sample means of Q, P, and *INCOME* are $\overline{Q} = 4.98$, $\overline{P} = 6.06$, and $\overline{INCOME} = 1.66$. Calculate the price and income elasticity of demand at the means.

e. Discuss the signs of the estimated coefficients in the supply equation.

f. Calculate the elasticity of equilibrium supply with respect to price at the means.

11.15 Consider the supply and demand for labor, and in particular that for married women. Wages and hours worked are jointly determined by supply and demand. Let the supply equation be

$$HOURS = \beta_1 + \beta_2 \ln(WAGE) + \beta_3 EDUC + \beta_4 AGE + \beta_5 KIDSL6 + \beta_6 KIDS618 + \beta_7 NWIFEINC + e_e$$

KIDSL6 are the number of children less than 6 years old, *KIDS618* are the number of children who are 6 to 18 years old, *NWIFEINC* is household income other than the wife's earnings. Let the demand equation be

$$\ln(WAGE) = \alpha_1 + \alpha_2 HOURS + \alpha_3 EDUC + \alpha_4 EXPER + \alpha_5 EXPER^2 + e_d$$

a. Imagine a supply and demand graph, like Figure 11.4, with *HOURS* on the vertical axis and ln(*WAGE*) on the horizontal axis. Describe the anticipated effects on the graph of increases in the number of small children on the woman's supply and demand curves. What is the anticipated effect on equilibrium wage and hours worked?

b. Describe the anticipated effects on the graph of increases in experience on the woman's supply and demand curves. What is the anticipated effect on equilibrium wage and hours worked?

c. Does the necessary condition for identification appear to hold for the supply equation? What are the IVs used in 2SLS? Write out the econometric form of the reduced-form equation for ln(*WAGE*), letting the coefficients be denoted as π_1, π_2, etc. What hypotheses would you test to evaluate the strength of IVs used in 2SLS estimation of the supply equation?

d. Does the necessary condition for identification appear to hold for the demand equation? What are the IVs used in 2SLS? Write out the econometric form of the reduced-form equation for *HOURS*, letting the coefficients be denoted as γ_1, γ_2, etc. What hypotheses would you test to evaluate the strength of IVs used in 2SLS estimation of the demand equation?

11.16 Consider the following supply and demand model

$$\text{Demand: } Q_i = \alpha_1 + \alpha_2 P_i + e_{di}, \quad \text{Supply: } Q_i = \beta_1 + \beta_2 P_i + \beta_3 W_i + e_{si}$$

where Q is the quantity, P is the price, and W is the wage rate, which is assumed exogenous. Data on these variables are in Table 11.7.

TABLE 11.7 Data for Exercise 11.16

Q	P	W
4	2	2
6	4	3
9	3	1
3	5	1
8	8	3

a. Derive the algebraic form of the reduced-form equations, $Q = \theta_1 + \theta_2 W + v_2$ and $P = \pi_1 + \pi_2 W + v_1$, expressing the reduced-form parameters in terms of the structural parameters.

b. Which structural parameters can you solve for from the results in part (a)? Which equation is "identified"?

c. The estimated reduced-form equations are $\hat{Q} = 5 + 0.5W$ and $\hat{P} = 2.4 + 1W$. Solve for the identified structural parameters. This is the method of **indirect least squares**.

d. Obtain the fitted values from the reduced-form equation for P, and apply 2SLS to obtain estimates of the demand equation.

11.17 Example 11.3 introduces Klein's Model I.

a. Do we have an adequate number of IVs to estimate each equation? Check the necessary condition for the identification of each equation. The necessary condition for identification is that in a system of M equations at least $M - 1$ variables must be omitted from each equation.

b. An equivalent identification condition is that the number of excluded exogenous variables from the equation must be at least as large as the number of included right-hand side endogenous variables. Check that this condition is satisfied for each equation.

c. Write down in econometric notation the first-stage equation, the reduced form, for W_{1t}, wages of workers earned in the private sector. Call the parameters π_1, π_2,…

d. Describe the two regression steps of 2SLS estimation of the consumption function. This is not a question about a computer software command.

e. Does following the steps in part (d) produce regression results that are identical to the 2SLS estimates provided by software specifically designed for 2SLS estimation? In particular, will the t-values be the same?

11.6.2 Computer Exercises

11.18 Example 11.3 introduces Klein's Model I. Here we examine a simplified model that excludes the government sector and allows further practice with simultaneous equations models. Suppose the model is

reduced to the following two equations, for two endogenous variables, consumption, *CN*, and investment, *I*. The two estimable equations are the consumption and investment functions:

$$CN_t = \alpha_1 + \alpha_2 I_t + \alpha_3 TIME_t + e_{1t}$$

$$I_t = \beta_1 + \beta_2 CN_t + \beta_3 K_{t-1} + e_{2t}$$

a. Check the identification of the consumption and investment functions.

b. Solve for the reduced-form equation for *CN*. Call the parameters π_1, π_2, π_3 and express them in terms of the structural parameters, similar to equations (11.4) and (11.5).

c. Using the data file *klein*, estimate each of the structural equations by OLS. Comment on the signs and significance of the coefficients.

d. Estimate each of the structural equations by 2SLS. Comment on the signs and significance of the coefficients.

e. Estimate the first-stage, reduced form, equation. In the reduced-form equation for consumption is K_{t-1} statistically significant? In the reduced-form equation for investment is $TIME_t$ statistically significant? Do these results help explain the differences in the OLS and 2SLS estimates?

11.19 The labor supply of married women has been a subject of a great deal of economic research. The data file is *mroz*, and the variable definitions are in the file *mroz.def*. The data file contains information on women who have worked in the previous year and those who have not. The variable indicating whether a woman worked *LFP*, labor force participation, takes the value 1 if a woman worked and 0 if she did not.

a. Calculate the summary statistics for the variables: wife's age, the number of less than 6-year-old children, and the income from other sources than from the wife's employment, *NWIFEINC*, for the women who worked (*LFP* = 1) and those who did not (*LFP* = 0). Define *NWIFEINC* = *FAMINC* – *WAGE* × *HOURS*. Comment on any differences you observe.

b. Consider the following supply equation specification:

$$HOURS = \beta_1 + \beta_2 \ln(WAGE) + \beta_3 EDUC + \beta_4 AGE + \beta_5 KIDSL6 + \beta_6 KIDS618 + \beta_7 NWIFEINC + e$$

What signs do you expect each of the coefficients to have, and why? What does *NWIFEINC* measure?

c. Estimate the supply equation in (b) using OLS regression on *only the women who worked* (*LFP* = 1). Did things come out as expected? If not, why not?

d. Estimate the reduced-form equation by OLS for the women who worked, using work experience, *EXPER*, as an additional exogenous variable.

$$\ln(WAGE) = \pi_1 + \pi_2 EDUC + \pi_3 AGE + \pi_4 KIDSL6 + \pi_5 KIDS618 + \pi_6 NWIFEINC + \pi_7 EXPER + v$$

Based on the estimated reduced form, what is the effect upon wage of an additional year of education?

e. Check the identification of the supply equation, considering the availability of instrument *EXPER*.

f. Estimate the supply equation by two-stage least squares, using software designed for this purpose. Discuss the signs and significance of the estimated coefficients.

11.20 This exercise examines a supply and demand model for edible chicken, which the U.S. Department of Agriculture calls "broilers." The data for this exercise are in the file *newbroiler*, which is adapted from the data provided by Epple and McCallum (2006). We consider the demand equation in this exercise and the supply equation in Exercise 11.21.

a. Consider the demand equation:

$$\ln(Q_t) = \alpha_1 + \alpha_2 \ln(P_t) + \alpha_3 \ln(Y_t) + \alpha_4 \ln(PB_t) + \alpha_5 POPGRO_t + e_t^d$$

where Q = per capita consumption of chicken, in pounds; Y = real per capita income; P = real price of chicken; PB = real price of beef; and $POPGRO$ = rate of population growth. What are the endogenous variables? What are the exogenous variables?

b. Using data from 1960 to 1999, estimate the demand equation by OLS. Comment on the signs and significance of the estimates.

c. Test the OLS residuals from part (b) for serial correlation by constructing a correlogram and carrying out the $T \times R^2$ test. What do you conclude about the presence of serial correlation?

d. Estimate the demand equation by 2SLS using as instruments $\ln(PF_t)$, $TIME_t = YEAR_t - 1949$, $\ln(QPROD_{t-1})$, and $\ln(EXPTS_{t-1})$. Compare and contrast these estimates to the OLS estimates in part (a).

e. Estimate the reduced-form, first-stage, equation and test the joint significance of $\ln(PF_t)$, $TIME_t$, $\ln(QPROD_{t-1})$, and $\ln(EXPTS_{t-1})$. Can we conclude that at least one instrument is strong?

f. Test the reduced-form equation for serial correlation using the $T \times R^2$ test.

g. Estimate the reduced-form, first-stage, equation using HAC standard errors and test the joint significance of $\ln(PF_t)$, $TIME_t$, $\ln(QPROD_{t-1})$, and $\ln(EXPTS_{t-1})$.

h. Obtain the 2SLS residuals from part (d). Construct a correlogram. Is there evidence of serial correlation? Obtain 2SLS estimates with HAC standard errors and compare the results to those in (d).

i. Test the validity of the surplus instruments using the Sargan test, discussed in Section 10.4.3, and the 2SLS estimates in part (d).

11.21 This exercise examines a supply and demand model for edible chicken, which the U.S. Department of Agriculture calls "broilers." The data for this exercise are in the file *newbroiler*, which is adapted from the data provided by Epple and McCallum (2006). We considered the demand equation in Exercise 11.20. The supply equation is

$$\ln(QPROD_t) = \beta_1 + \beta_2 \ln(P_t) + \beta_3 \ln(PF_t) + \beta_4 TIME_t + \ln(QPROD_{t-1}) + e_t^s$$

where *QPROD* is the aggregate production of young chickens, *PF* is nominal price index of broiler feed, and *TIME* = time index with $1950 = 1, \ldots, 2001 = 52$. This supply equation is dynamic, with lagged production on the right-hand side. This predetermined variable is exogenous. *TIME* is included to capture technical progress in production.

a. What are the endogenous variables? What are the exogenous variables? What is the interpretation of the parameter β_2? What signs do you expect for each of the parameters?

b. Using data from 1960 to 1999, estimate the supply equation by OLS. Comment on the signs and significance of the estimates. Test the residuals for serial correlation. Is serial correlation present?

c. Estimate the reduced-form, first-stage, regression by OLS using the IVs $\ln(Y_t)$, $\ln(PB_t)$, *POPGRO*, and $\ln(EXPTS_{t-1})$. Test the joint significance of these variables. Can we conclude that we have at least one strong instrument?

d. Estimate the supply equation by 2SLS using the instruments listed in part (c). Compare and contrast these results to those in part (b).

e. Test the validity of the surplus instruments using the Sargan test, discussed in Section 10.4.3.

11.22 This exercise examines a supply and demand model for edible chicken, which the U.S. Department of Agriculture calls "broilers." The data for this exercise are in the file *newbroiler*, which is adapted from the data provided by Epple and McCallum (2006). We considered the demand equation in Exercise 11.20. It is

$$\ln(Q_t) = \alpha_1 + \alpha_2 \ln(P_t) + \alpha_3 \ln(Y_t) + \alpha_4 \ln(PB_t) + \alpha_5 POPGRO_t + e_t^d$$

where Q is the per capita consumption of chicken, in pounds; Y is real per capita income; P is real price of chicken; *PB* is real price of beef, and *POPGRO* is rate of population growth. What are the endogenous variables? What are the exogenous variables? The demand equation suffers from serial correlation. In the AR(1) model $e_t^d = \rho e_{t-1}^d + v_t^d$ the value of ρ is large. Epple and McCallum estimate the model in "first difference" form:

$$\begin{aligned} \ln(Q_t) &= \alpha_1 + \alpha_2 \ln(Y_t) + \alpha_3 \ln(P_t) + \alpha_4 \ln(PB_t) + e_t^d \\ -\left[\ln(Q_t) = \alpha_1 + \alpha_2 \ln(Y_t) + \alpha_3 \ln(P_t) + \alpha_4 \ln(PB_t) + e_t^d\right] & \\ \hline \Delta \ln(Q_t) = \alpha_2 \Delta \ln(Y_t) + \alpha_3 \Delta \ln(P_t) + \alpha_4 \Delta \ln(PB_t) + v_t^d & \end{aligned}$$

a. Regarding this specification (i) what changes do you notice after this transformation? (ii) Are the parameters of interest affected? (iii) If $\rho = 1$, have we solved the serial correlation problem? (iv) What is the interpretation of the "Δ" variables like $\Delta \ln(Q_t)$? [*Hint*: See Appendix A.4.6.] (v) What is the interpretation of the parameter α_2? (vi) What signs do you expect for each of the coefficients? Explain.

b. Using data from 1960 to 1999, estimate the reduced-form, first-stage, equation for $\Delta\ln(P_t)$ using instruments $\ln(PF_t)$, $TIME_t$, $\ln(QPROD_{t-1})$, and $\ln(EXPTS_{t-1})$. Can we conclude that at least one instrument is strong?

c. Estimate the first-stage equation for $\Delta\ln(P_t)$ using instruments $\Delta\ln(PF_t)$, $\Delta\ln(QPROD_{t-1})$, and $\Delta\ln(EXPTS_{t-1})$. Can we conclude that at least one instrument is strong? On logical grounds, why might we prefer these instruments to those in (b)?

d. Estimate the first-stage equation for $\Delta\ln(P_t)$ using instrument $\Delta\ln(PF_t)$. Can we conclude that the one instrument is strong?

e. Obtain the 2SLS estimates of the first-differenced demand equation using $\Delta\ln(PF_t)$ as the instrument. In this estimation omit the constant term.

f. Obtain the 2SLS estimates of the first-differenced demand equation using $\Delta\ln(PF_t)$ as the instrument including a constant term.

g. Compare the estimates of the key demand parameters in parts (e) and (f). Are the signs consistent with expectations? What are the interpretations of the estimated coefficients? Should an intercept be included in the differenced demand equation? Explain.

h. Construct a correlogram for the 2SLS residuals in part (e). Is there any evidence of serial correlation?

11.23 Reconsider Example 11.2 on the supply and demand for fish at the Fulton Fish Market. The data are in the file *fultonfish*.

a. Obtain OLS estimates of the supply equation. Comment on the coefficient signs and significance. Do you anticipate the OLS estimator to have a positive bias or a negative bias or no bias? Explain.

b. It is possible that bad weather on shore reduces attendance at restaurants, which in turn may reduce the demand for fish at the market. Add the variables *RAINY* and *COLD* to the demand equation in (11.13). Derive the algebraic reduced form for ln(*PRICE*) for this new specification.

c. Estimate the reduced-form equation in part (b). Test the joint significance of *RAINY* and *COLD*. Are these variables jointly significant at the 5% level?

d. Using the estimates from part (c), test the joint significance of *MON*, *TUE*, *WED*, *THU*, *RAINY*, and *COLD*. Are these variables jointly significant at the $\alpha = 0.05$ level?

e. Estimate the supply equation by 2SLS using instruments *MON*, *TUE*, *WED*, *THU*, *RAINY*, and *COLD*. Compare these estimates to the OLS estimates in part (a). Given the results in part (d), can we conclude that the supply equation is identified?

11.24 Reconsider Example 11.2 on the supply and demand for fish at the Fulton Fish Market. The data are in the file *fultonfish*.

a. Add the variable *MIXED*, which indicates poor but not *STORMY* weather conditions, to the supply equation in equation (11.14). Estimate the new reduced-form equation for ln(*PRICE*), adding the variable *MIXED* to equation (11.16). Is it statistically significant at the 5% level? Test the joint significance of *STORMY* and *MIXED*. Is the resulting *F*-value greater than 10?

b. Estimate the demand equation using *STORMY* and *MIXED* as IVs. Compare the coefficient estimates to those in Table 11.5.

c. Test the validity of the surplus instrument using the Sargan test, discussed in Section 10.3.4.

d. In the reduced-form equation in part (a), test the joint significance of the indicator variables *MON*, *TUE*, *WED*, and *THU* at the 5% level. What do you conclude? Are we now able to estimate the supply equation by 2SLS with confidence in our procedure?

11.25 Reconsider Example 11.2 on the supply and demand for fish at the Fulton Fish Market. The data are in the file *fultonfish*. In this exercise, we explore the behavior of the market on days in which changes in fish inventories are large relative to those days on which inventory changes are small. Graddy and Kennedy (2006) anticipate that prices and quantities will demonstrate simultaneity on days with large changes in inventories, as these are days when sellers are demonstrating their responsiveness to prices. On days when inventory changes are small, the anticipation is that feedback between prices and quantities is broken, and simultaneity is no longer an issue.

a. Use the subset of data for days in which inventory change is large, as indicated by the variable $CHANGE = 1$. Estimate the reduced-form equation (11.16) and test the significance of *STORMY*. Discuss the importance of this test for the purpose of estimating the demand equation by two-stage least squares.

b. Obtain the OLS residuals $\hat{v}_{t2}$ from the reduced-form equation estimated in (a). Carry out a Hausman test, as discussed in Section 10.4.1, for the endogeneity of ln(*PRICE*) by adding $\hat{v}_{t2}$

as an extra variable to the demand equation in (11.13), estimating the resulting model by OLS, and testing the significance of $\hat{v}_{t2}$ using a standard t-test. If $\hat{v}_{t2}$ is a significant variable in this augmented regression then we may conclude that ln(*PRICE*) is endogenous. Based on this test, what do you conclude?

c. Estimate the demand equation using two-stage least squares and OLS using the data when $CHANGE = 1$, and discuss these estimates. Compare them to the estimates in Table 11.5.

d. Estimate the reduced-form equation (11.16) for the data when $CHANGE = 0$. Compare these reduced-form estimates to those in (a) and those in Table 11.4b.

e. Obtain the OLS residuals $\hat{v}_{t2}$ from the reduced-form equation estimated in (d). Carry out a Hausman test for the endogeneity of ln(*PRICE*), as described in part (b). Based on this test, what do you conclude?

f. Obtain the two-stage least squares and the OLS estimates for the demand equation for the data when $CHANGE = 0$. Compare these estimates to each other and to the estimates in (c). Discuss the relationships between them.

11.26 Use your computer software for two-stage least squares or IVs estimation, and the 30 observations in the data file *truffles* to obtain 2SLS estimates of the system in equations (11.4) and (11.5). Compare your results to those in Tables 11.3a and 11.3b.

a. Using the 2SLS estimated equations, compute the price elasticity of supply and demand "at the means." Comment on the signs and magnitudes of these elasticities.

b. Using the 2SLS estimates for the demand equation, obtain the squared 2SLS residuals, $\hat{e}_d^2$. Carry out the Breusch–Pagan NR^2 test for heteroskedasticity using just the exogenous variables in the variance function. Is there any evidence of heteroskedasticity?

c. Using the 2SLS estimates for the supply equation, obtain the squared 2SLS residuals, $\hat{e}_s^2$. Carry out the Breusch–Pagan NR^2 test for heteroskedasticity using just the exogenous variables in the variance function. Is there any evidence of heteroskedasticity?

d. Plot the squared supply equation residuals $\hat{e}_s^2$ versus each of the three exogenous variables. Discuss the visual evidence of heteroskedasticity.

e. Obtain 2SLS estimates of the supply equation using robust standard errors. How do the t-statistic values compare to those in Table 11.3b? Do you think it is a good idea to use robust standard errors for this equation? Explain.

11.27 Estimate equations (11.4) and (11.5) by OLS, ignoring the fact that they form a simultaneous system. Use the data file *truffles*. Compare your results to those in Table 11.3. Do the signs of the least squares estimates agree with economic reasoning?

11.28 Supply and demand curves as traditionally drawn in economics principles classes have price (P) on the vertical axis and quantity (Q) on the horizontal axis.

a. Rewrite the truffle demand and supply equations in (11.11) and (11.12) with price P on the left-hand side. What are the anticipated signs of the parameters in this rewritten system of equations?

b. Using the data in the file *truffles*, estimate the supply and demand equations that you have formulated in (a) using two-stage least squares. Are the signs correct? Are the estimated coefficients significantly different from zero?

c. Estimate the price elasticity of demand "at the means" using the results from (b).

d. Accurately sketch the supply and demand equations, with P on the vertical axis and Q on the horizontal axis, using the estimates from part (b). For these sketches set the values of the exogenous variables DI, PS, and PF to be $DI^* = 3.5$, $PF^* = 23$, and $PS^* = 22$.

e. What are the equilibrium values of P and Q obtained in part (d)? Calculate the predicted equilibrium values of P and Q using the estimated reduced-form equations from Table 11.2, using the same values of the exogenous variables. How well do they agree?

f. Estimate the supply and demand equations that you have formulated in (a) using OLS. Are the signs correct? Are the estimated coefficients significantly different from zero? Compare the results to those in part (b).

11.29 Example 11.3 introduces Klein's Model I. Use the data file *klein* to answer the following questions.

a. Estimate the consumption function in equation (11.17) by OLS. Comment on the signs and significance of the coefficients.

b. Estimate the reduced-form equation for wages of workers in the private sector, W_{1t}, using all eight exogenous and predetermined variables as explanatory variables. Test the joint significance of all

the variables except wages of workers in the public sector, W_{2t}, and lagged profits, P_{t-1}. Save the residuals, $\hat{v}_{1t}$.

c. Estimate the reduced-form equation for profits, P_t, using all eight exogenous and predetermined variables as explanatory variables. Test the joint significance of all the variables except wages of workers in the public sector, W_{2t}, and lagged profits, P_{t-1}. Save the residuals, $\hat{v}_{2t}$.

d. The Hausman test for the presence of endogenous explanatory variables is discussed in Section 10.4.1. It is implemented by adding the reduced-form residuals to the structural equation and testing their significance, that is, using OLS, estimate the model

$$CN_t = \alpha_1 + \alpha_2(W_{1t} + W_{2t}) + \alpha_3 P_t + \alpha_4 P_{t-1} + \delta_1 \hat{v}_{1t} + \delta_2 \hat{v}_{2t} + e_{1t}$$

Use an F-test for the null hypothesis $H_0: \delta_1 = 0, \delta_2 = 0$ at the 5% level of significance. By rejecting the null hypothesis, we conclude that either W_{1t} or P_t is endogenous, or both are endogenous. What do we conclude from the test? In the context of this simultaneous equations model what result should we find?

e. Obtain the 2SLS estimates of the consumption equation using all eight exogenous and predetermined variables as IVs. Compare the estimates to the OLS estimates in part (a). Do you find any important differences?

f. Let the 2SLS residuals from part (e) be $\hat{e}_{1t}$. Regress these residuals on all the exogenous and predetermined variables. If these instruments are valid, then the R^2 from this regression should be low, and none of the variables are statistically significant. The Sargan test for instrument validity is discussed in Section 10.4.3. The test statistic TR^2 has a chi-square distribution with degrees of freedom equal to the number of "surplus" IVs if the surplus instruments are valid. The consumption equation includes three exogenous and/or predetermined variables of the total of eight possible. There are $L = 5$ external instruments and $B = 2$ right-hand side endogenous variables. Compare the value of the test statistic to the 95th percentile value from the $\chi^2_{(3)}$ distribution. What do we conclude about the validity of the surplus instruments in this case?

11.30 Example 11.3 introduces Klein's Model I. Use the data file *klein* to answer the following questions.

a. Estimate the investment function in equation (11.18) by OLS. Comment on the signs and significance of the coefficients.

b. Estimate the reduced-form equation for profits, P_t, using all eight exogenous and predetermined variables as explanatory variables. Test the joint significance of all the variables except lagged profits, P_{t-1}, and lagged capital stock, K_{t-1}. Save the residuals, $\hat{v}_t$ and compute the fitted values, $\hat{P}_t$.

c. The Hausman test for the presence of endogenous explanatory variables is discussed in Section 10.4.1. It is implemented by adding the reduced-form residuals to the structural equation and testing their significance, that is, using OLS estimate the model

$$I_t = \beta_1 + \beta_2 P_t + \beta_3 P_{t-1} + \beta_4 K_{t-1} + \delta \hat{v}_t + e_{2t}$$

Use a t-test for the null hypothesis $H_0: \delta = 0$ versus $H_1: \delta \neq 0$ at the 5% level of significance. By rejecting the null hypothesis, we conclude that P_t is endogenous. What do we conclude from the test? In the context of this simultaneous equations model what result should we find?

d. Obtain the 2SLS estimates of the investment equation using all eight exogenous and predetermined variables as IVs and software designed for 2SLS. Compare the estimates to the OLS estimates in part (a). Do you find any important differences?

e. Estimate the second-stage model $I_t = \beta_1 + \beta_2 \hat{P}_t + \beta_3 P_{t-1} + \beta_4 K_{t-1} + e_{2t}$ by OLS. Compare the estimates and standard errors from this estimation to those in part (d). What differences are there?

f. Let the 2SLS residuals from part (e) be $\hat{e}_{2t}$. Regress these residuals on all the exogenous and predetermined variables. If these instruments are valid, then the R^2 from this regression should be low, and none of the variables are statistically significant. The Sargan test for instrument validity is discussed in Section 10.4.3. The test statistic TR^2 has a chi-square distribution with degrees of freedom equal to the number of "surplus" IVs if the surplus instruments are valid. The investment equation includes three exogenous and/or predetermined variables out of the total of eight possible. There are $L = 5$ external instruments and $B = 1$ right-hand side endogenous variables. Compare the value of the test statistic to the 95th percentile value from the $\chi^2_{(4)}$ distribution. What do we conclude about the validity of the surplus instruments in this case?

Appendix 11A 2SLS Alternatives

There has always been great interest in alternatives to the standard IV/2SLS estimator. The search for better alternatives was energized by the discovery of the problems weak instruments pose for the usual IV/2SLS estimator. In this appendix, we examine a few alternative estimators for a single equation with endogenous regressors. The equation might be part of a simultaneous equations system, or a standalone equation with an endogenous regressor, as we studied in Chapter 10. The limited information maximum likelihood (LIML) estimator was first derived by Anderson and Rubin in 1949.[1] It has played a "back seat" role relative to 2SLS over the years, but this is no longer true. There is renewed interest in LIML in the presence of weak instruments. Several modifications of LIML have been suggested by Fuller (1977) and others. These estimators are unified in a common framework, along with 2SLS, using the idea of a ***k*-class** of estimators. Later in this appendix, we provide Stock–Yogo tables of critical values for weak instruments that apply to the LIML estimator and Fuller modifications. What is illustrated by these tables is that LIML suffers less from test size aberrations than the 2SLS estimator, and that the Fuller modification suffers less from bias.

11A.1 The *k*-Class of Estimators

In a system of M simultaneous equations let the endogenous variables be $y_1, y_2, \ldots, y_M$. Let there be K exogenous variables, $x_1, x_2, \ldots, x_K$. Suppose the first structural equation within this system is

$$y_1 = \alpha_2 y_2 + \beta_1 x_1 + \beta_2 x_2 + e_1 \tag{11A.1}$$

If this equation is identified, then its parameters can be estimated. The variable y_2 is endogenous because it is correlated with the regression error term e_1. The endogenous variable y_2 has reduced form $y_2 = \pi_{12}x_1 + \pi_{22}x_2 + \cdots + \pi_{K2}x_K + v_2 = E(y_2|\mathbf{X}) + v_2$. The source of the endogeneity of y_2 is not the systematic portion $E(y_2|\mathbf{X})$, which is exogenous. The random component v_2 is the source of the endogeneity problem. One way to think about developing an IV for y_2 is to remove, or "purge," v_2 from it, that is, use the IV $y_2 - v_2 = E(y_2|\mathbf{X})$. This instrument has the essential properties of an instrument: It is correlated with the endogenous variable y_2 and it is uncorrelated with the structural equation error e_1. The difficulty is that $E(y_2|\mathbf{X})$ is unknown. However, the parameters of the reduced-form equation are consistently estimated by OLS, so that

$$\widehat{E(y_2|\mathbf{X})} = \hat{\pi}_{12}x_1 + \hat{\pi}_{22}x_2 + \cdots + \hat{\pi}_{K2}x_K \tag{11A.2}$$

The reduced-form residuals are

$$\hat{v}_2 = y_2 - \widehat{E(y_2|\mathbf{X})}$$

In large samples the reduced-form estimators $\hat{\pi}_{k2}$ converge in probability to their true values. This means that in large samples we can substitute for $E(y_2|\mathbf{X})$ its estimated value

$$\widehat{E(y_2|\mathbf{X})} = y_2 - \hat{v}_2 \tag{11A.3}$$

The two-stage least squares estimator is an IV estimator using $\widehat{E(y_2|\mathbf{X})}$ as an instrument. Equation (11A.3) shows that the instrument used in 2SLS can be thought of as the endogenous variable y_2 "purged" of the troublesome error term v_2.

The ***k*-class** of estimators is a unifying framework. A *k-class* estimator is an IV estimator using IV $y_2 - k\hat{v}_2$. It is called a *class* of estimators because it represents the OLS estimator if

[1] Anderson, T.W. and Rubin, H. (1949), "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 21, 46–63.

$k = 0$ and the 2SLS estimator if $k = 1$. Why would we be interested in using values of k other than 1? Hopefully by adjusting this value we can improve upon the performance of the k-class estimator relative to the 2SLS estimator.

11A.2 The LIML Estimator

As noted earlier, the LIML estimator is one of the oldest estimators for an equation within a system of simultaneous equations, or any equation with an endogenous variable on the right-hand side. Rather than obtaining the LIML estimates by maximizing a likelihood function (see Appendix C.8 for an introduction to maximum likelihood estimation) we will exploit the fact that the LIML estimator is a member of the *k-class*.

The equation $y_1 = \alpha_2 y_2 + \beta_1 x_1 + \beta_2 x_2 + e_1$ is in **normalized form**, meaning that we have chosen one variable to appear as the dependent variable. In general the first equation can be written in **implicit form** as $\alpha_1 y_1 + \alpha_2 y_2 + \beta_1 x_1 + \beta_2 x_2 + e_1 = 0$. There is no rule that says y_1 has to be the dependent variable in the first equation. **Normalization** amounts to setting α_1 or α_2 to the value -1. One parameter α_i must be set to -1 so that we can identify the equation, but it does not matter which one. Let $y^* = \alpha_1 y_1 + \alpha_2 y_2$, then the unnormalized equation can be written as $y^* + \beta_1 x_1 + \beta_2 x_2 + e_1 = 0$, or

$$y^* = -\beta_1 x_1 - \beta_2 x_2 - e_1 = \theta_1 x_1 + \theta_2 x_2 + \eta \tag{11A.4}$$

In (11A.1) the exogenous variables $x_3, \ldots, x_K$ were omitted. If we had included them, (11A.4) would be

$$y^* = \theta_1 x_1 + \cdots + \theta_K x_K + \eta \tag{11A.5}$$

The **least variance ratio** estimator chooses α_1 and α_2 so that the ratio of the sum of squared residuals from (11A.4) relative to the sum of squared residuals from (11A.5) is as small as possible. Define the ratio of sum of squared residuals from the two models as

$$\ell = \frac{SSE \text{ from regression of } y^* \text{ on } x_1, x_2}{SSE \text{ from regression of } y^* \text{ on } x_1, \ldots, x_K} \geq 1 \tag{11A.6}$$

We assume that the variables $x_3, \ldots, x_K$ were omitted from (11A.1) for a reason based in economic theory. The estimates of α_1 and α_2, one of which will be set to -1, should be chosen so to make the reduced regression (11A.4) fit the data as well as possible while still imposing the condition that $x_3, \ldots, x_K$ are omitted.

The algebra required for the solution is beyond the scope of this book.[2] The interesting result is that the minimum value of ℓ in (11A.6), call it $\hat{\ell}$, results in the LIML estimator when used as k in the *k-class* estimator, that is, use $k = \hat{\ell}$ when forming the instrument $y_2 - k\hat{v}_2$, and the resulting IV estimator is the LIML estimator.

11A.2.1 Fuller-Modified LIML

A modification suggested by Wayne Fuller (1977)[3] uses the *k-class* value

$$k = \hat{\ell} - \frac{a}{N - K} \tag{11A.7}$$

where K is the total number of IVs (included and excluded exogenous variables) and N is the sample size. The value of a is a constant. Fuller says (1977, p. 951), "If one desires estimates that are nearly unbiased 'a' is set equal to 1. Presumably 'a' = 1 would be used when one is interested in testing hypotheses or setting approximate confidence intervals for the parameters." Fuller also showed that the value $a = 4$ leads to an estimator that minimizes the "mean square

[2] Advanced students should consider reading Peter Schmidt's *Econometrics*, 1976, Chapter 4, New York, NY: Marcel Dekker. Inc.

[3] Wayne Fuller, "Some Properties of a Modification of the Limited Information Estimator," *Econometrica*, 45, 939–953.

error" of estimation. If we are estimating some parameter δ using an estimator $\hat{\delta}$, then the mean square error of estimation is

$$MSE\left(\hat{\delta}\right) = E\left(\hat{\delta} - \delta\right)^2 = \text{var}\left(\hat{\delta}\right) + \left[E\left(\hat{\delta}\right) - \delta\right]^2 = \text{var}\left(\hat{\delta}\right) + \left[bias\left(\hat{\delta}\right)\right]^2$$

Estimator *MSE* combines both variance and bias into a single measure.

11A.2.2 Advantages of LIML

A great deal of research has been devoted to the performance of the LIML estimator relative to the 2SLS estimator when instruments are weak and/or there are a large number of instruments. Stock and Yogo (2005, p. 106) say, "Our findings support the view that LIML is far superior to *(2)SLS* when the researcher has weak instruments …" when using interval estimates' coverage rate as the criterion. Also "… the Fuller-*k* estimator is more robust to weak instruments than *(2)SLS* when viewed from the perspective of bias." Some other findings are discussed by Mariano (2001)[4]:

- For the 2SLS estimator the amount of bias is an increasing function of the degree of over identification. The distributions of the 2SLS and least squares estimators tend to become similar when overidentification is large. LIML has the advantage over 2SLS when there are a large number of instruments.
- The LIML estimator converges to normality faster than the 2SLS estimator and is generally more symmetric.

11A.2.3 Stock–Yogo Weak IV Tests for LIML

Tables 11A.1 and 11A.2 contain Stock–Yogo critical values for testing weak instruments. These tests are discussed in Chapter 10, Appendix A. Table 11A.1 contains the critical values using the criterion of maximum LIML test size for a 5% test. Note that for $L > 1$, LIML critical values are lower than the 2SLS critical values in Table 10A.1. This means that the Cragg–Donald F-test statistic does not have to be as large for us to reject the null hypothesis that the instruments are weak when using LIML instead of 2SLS. Table 11A.2 contains the critical values for the test of weak instruments using the relative bias criterion for the Fuller modification of LIML, using $a = 1$. There is no similar table for LIML, because the LIML estimator does not have a finite expected value, and thus the concept of bias breaks down.

TABLE 11A.1 **Critical Values for the Weak Instrument Test Based on LIML Test Size (5% Level of Significance)[5]**

	B = 1				*B = 2*			
	Maximum Test Size				Maximum Test Size			
L	0.10	0.15	0.20	0.25	0.10	0.15	0.20	0.25
1	16.38	8.96	6.66	5.53				
2	8.68	5.33	4.42	3.92	7.03	4.58	3.95	3.63
3	6.46	4.36	3.69	3.32	5.44	3.81	3.32	3.09
4	5.44	3.87	3.30	2.98	4.72	3.39	2.99	2.79

[4]Mariano, R. S. (2001), "Simultaneous equation model estimators," in *The Companion to Theoretical Econometrics*, Badi Baltagi ed., Oxford: Blackwell Publishing, pp. 139–142.

[5]These values are from Table 5.4, page 103, in Stock and Yogo (2005), *op. cit*. The authors thank James Stock and Motohiro Yogo for permission to use these results. Their tables are more extensive than the ones we provide. The significance level of the test for weak instruments is 5%.

TABLE 11A.2 **Critical Values for the Weak Instrument Test Based on Fuller-*k* Relative Bias (5% Level of Significance)[6]**

	$B = 1$				$B = 2$			
	Maximum Relative Bias				Maximum Relative Bias			
L	0.05	0.10	0.20	0.30	0.05	0.10	0.20	0.30
1	24.09	19.36	15.64	12.71				
2	13.46	10.89	9.00	7.49	15.50	12.55	9.72	8.03
3	9.61	7.90	6.61	5.60	10.83	8.96	7.18	6.15
4	7.63	6.37	5.38	4.63	8.53	7.15	5.85	5.10

EXAMPLE 11.4 | Testing for Weak Instruments Using LIML

This illustration was introduced in Example 10.8. With the Mroz data we estimate the *HOURS* supply equation

$$HOURS = \beta_1 + \beta_2 MTR + \beta_3 EDUC + \beta_4 KIDSL6 + \beta_5 NWIFEINC + e \quad (11A.8)$$

The reduced-form estimates are in Table 10A.3. The LIML estimates are given in Table 11A.3. The models we consider are as follows:

Model 1: endogenous: *MTR*; IV: *EXPER*

Model 2: endogenous: *MTR*; IV: *EXPER*, $EXPER^2$, *LARGECITY*

Model 3: endogenous: *MTR*, *EDUC*; IV: *MOTHEREDUC*, *FATHEREDUC*

Model 4: endogenous: *MTR*, *EDUC*; IV: *MOTHEREDUC*, *FATHEREDUC*, *EXPER*

TABLE 11A.3 **LIML Estimations**

MODEL	(1)	(2)	(3)	(4)
C	17423.7211	16191.3338	−24491.5972	18587.9064
	(5.56)	(5.40)	(−0.31)	(5.05)
MTR	−18456.5896	−17023.8164	29709.4652	−19196.5172
	(−5.08)	(−4.90)	(0.33)	(−4.79)
EDUC	−145.2928	−134.5504	258.5590	−197.2591
	(−4.40)	(−4.26)	(0.31)	(−3.05)
KIDSL6	151.0229	113.5034	−1144.4778	207.5531
	(1.07)	(0.84)	(−0.46)	(1.27)
NWIFEINC	−103.8983	−96.2895	149.2325	−104.9415
	(−5.27)	(−5.11)	(0.32)	(−5.07)
N	428	428	428	428
$\hat{\ell}$	1.0000	1.0195	1.0000	1.0029
CRAGG–DONALD *F*	30.61	13.22	0.10	8.60
NUMBER IV *L*	1	3	2	3
NUMBER ENDOG *B*	1	1	2	2

t-statistics in parentheses.

[6]These values are from Table 5.3, page 102, in James H. Stock and Motohiro Yogo (2005), *op. cit.*

First, for the just identified equations for which the number of instruments equals the number of endogenous variables in Models (1) and (3), the LIML estimates are identical to the 2SLS estimators. This identity is always true for just-identified equations. For the overidentified Models (2) and (4), the estimated values $\hat{\ell}$ are close to 1, so that the estimates are not too far from the 2SLS estimates.

The estimates are not the important aspect of this illustration. The Cragg–Donald F-test statistic is the same for all the estimators. For convenience its values for each equation are given at the bottom of Table 11A.3. In Model (2), we have $B = 1$ endogenous variable and $L = 3$ instruments. Using the LIML maximum size of 10% as our criterion, the Stock–Yogo critical value is 6.46. The Cragg–Donald F-test statistic 13.22 exceeds this value, so we reject the null hypothesis that the instruments are weak and conclude that they are not weak. This is not the conclusion we would have drawn based on IV/2SLS estimation. The critical value from Table 10A.1 is 22.30, and we would have not rejected the null hypothesis that the instruments are weak.

In Model (4) there are $B = 2$ endogenous variables and $L = 3$ instruments. Using the maximum size of 10% critical value from Table 11A.1 of 5.44, we reject the null hypothesis that the instruments are weak using the Cragg–Donald F-test statistic of 8.60. If we were using the 2SLS/IV estimator, we would have not rejected the hypothesis that the instruments are weak because the critical value from Table 10A.1 is 13.43.

What is indicated by these examples is that the LIML estimator performs better, at least potentially, in the face of weak instruments. We cannot prove anything based on one result from one sample, which is why we present a Monte Carlo simulation experiment in Appendix 11A.3.

EXAMPLE 11.5 | Testing for Weak Instruments with Fuller-Modified LIML

Using the Fuller modification of LIML, and setting the constant $a = 1$, we obtain the estimates in Table 11A.4. All the results are at least somewhat different from the 2SLS/IV estimations, because even for just-identified equations, the Fuller estimator is different from the 2SLS estimator. The only extremely dramatic change now comes in Model (3), where coefficient signs become more in line with the other models, although still nothing is significant. In Model (4), if we adopt the criterion of 10% maximum relative bias, then the Stock–Yogo critical value is 8.96. The Cragg–Donald F-test statistic is 8.6, so we fail to reject the null hypothesis that the instruments are weak.

TABLE 11A.4 **Fuller ($a = 1$) Estimations**

MODEL	(1)	(2)	(3)	(4)
C	17108.0110	15924.1895	2817.5400	18156.7850
	(5.60)	(5.44)	(0.20)	(5.10)
MTR	−18089.5451	−16713.2345	−1304.8205	−18730.1617
	(−5.11)	(−4.93)	(−0.08)	(−4.84)
EDUC	−142.5409	−132.2218	−29.6043	−191.1248
	(−4.41)	(−4.27)	(−0.20)	(−3.05)
KIDSL6	141.4113	105.3703	−287.7915	193.2295
	(1.02)	(0.79)	(−0.65)	(1.21)
NWIFEINC	−101.9491	−94.6401	−12.0108	−102.6290
	(−5.31)	(−5.14)	(−0.15)	(−5.12)
N	428	428	428	428
k	0.9976	1.0172	0.9976	1.0005
FULLER *a*	1.0000	1.0000	1.0000	1.0000
NUMBER IV *L*	1	3	2	3
CRAGG–DONALD *F*	30.61	13.22	0.10	8.60
NUMBER ENDOG *B*	1	1	2	2

t-statistics in parentheses

11A.3 Monte Carlo Simulation Results

In Appendix 10B.2, we carried out a Monte Carlo simulation to explore the properties of the IV/2SLS estimators. Here we employ the same experiment, adding aspects of the new estimators we have introduced in this appendix.

First, examine the percentage rejections of the true null hypothesis $\beta_2 = 1$ using a two-tail test at the 5% level of significance. The Monte Carlo rejection rate for the IV/2SLS estimator is in the column labeled $t\left(\hat{\beta}_2\right)$, and for the LIML estimator in the column $t\left(\hat{\beta}_{2,\text{LIML}}\right)$. The largest difference is in the case of strong endogeneity with weak instruments, in which the test based upon the two-stage least squares estimator rejects 28.86% of the time, while the test based on the LIML estimator rejects 13.47% of the time. Recall that a two-tail test at the 5% level of significance corresponds to determining whether the 95% interval estimate contains the hypothesized parameter value. In these Monte Carlo experiments, the 95% interval estimate based on the LIML estimator contains the true parameter 86.53% of the time, whereas the 95% interval estimate using IV/2SLS contains the true parameter only 71.14% of the time. This finding is consistent with Stock and Yogo's conclusion about coverage rates of the two interval estimation approaches.

In these experiments, there is little difference between the averages of the two-stage least squares estimates, $\bar{\hat{\beta}}_2$ and the Fuller modified ($a = 1$) LIML estimates $\bar{\hat{\beta}}_{2,\text{F}}$. A greater contrast shows up when comparing how close the estimates are to the true parameter value using the mean square error criterion. In Table 11A.5, we report the empirical mean square error for the IV/2SLS estimator, $\text{mse}\left(\hat{\beta}_2\right)$ and that for the Fuller modification of LIML with $a = 4$, $\text{mse}\left(\hat{\beta}_{2,\text{F}}\right)$. Recall that the mean square error measures how close the estimates are to the true parameter value. For the IV/2SLS estimator, the empirical mean square error is

$$\text{mse}\left(\hat{\beta}_2\right) = \sum_{m=1}^{10000}\left(\hat{\beta}_{2m} - \beta_2\right)^2 \Big/ 10{,}000$$

The Fuller-modified LIML has lower mean square error than the IV/2SLS estimator in each experiment, and when the instruments are weak, the improvement is large.

TABLE 11A.5 Monte Carlo Simulation Results

ρ	π	$\bar{F}$	$\bar{\hat{\beta}}_2$	$t\left(\hat{\beta}_2\right)$	$t\left(\hat{\beta}_{2,\text{LIML}}\right)$	$\bar{\hat{\beta}}_{2,\text{F}}$	$\text{mse}\left(\hat{\beta}_2\right)$	$\text{mse}\left(\hat{\beta}_{2,\text{F}}\right)$
0.0	0.1	1.98	0.9941	0.0049	0.0049	0.9941	0.4068	0.0748
0.0	0.5	21.17	0.9998	0.0441	0.0473	0.9997	0.0140	0.0132
0.8	0.1	2.00	1.3311	0.2886	0.1347	1.3375	1.0088	0.3289
0.8	0.5	21.18	1.0111	0.0636	0.0509	1.0000	0.0139	0.0127