8

ENDOGENEITY AND INSTRUMENTAL VARIABLE ESTIMATION

8.1 INTRODUCTION

The assumption that $\mathbf{x}_i$ and ε_i are uncorrelated in the linear regression model,

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon, \tag{8-1}$$

has been crucial in the development thus far. But there are many applications in which this assumption is untenable. Examples include models of treatment effects such as those in Examples 6.8–6.13, models that contain variables that are measured with error, dynamic models involving expectations, and a large variety of common situations that involve variables that are unobserved, or for other reasons are omitted from the equation. Without the assumption that the disturbances and the regressors are uncorrelated, none of the proofs of consistency or unbiasedness of the least squares estimator that were obtained in Chapter 4 will remain valid, so the least squares estimator loses its appeal. This chapter will develop an estimation method that arises in situations such as these.

It is convenient to partition $\mathbf{x}$ in (8-1) into two sets of variables, $\mathbf{x}_1$ and $\mathbf{x}_2$, with the assumption that $\mathbf{x}_1$ is not correlated with ε and $\mathbf{x}_2$ is, or may be (part of the empirical investigation). We are assuming that $\mathbf{x}_1$ is **exogenous** in the model—see Assumption A.3 in the statement of the linear regression model in Section 2.3. It will follow that $\mathbf{x}_2$ is, by this definition, **endogenous** in the model. How does endogeneity arise? Example 8.1 suggests some common settings.

Example 8.1 Models with Endogenous Right-Hand-Side Variables

The following models and settings will appear at various points in this book.

Omitted Variables: In Example 4.2, we examined an equation for gasoline consumption of the form

$$\ln G = \beta_1 + \beta_2 \ln \mathit{Price} + \beta_3 \ln \mathit{Income} + \varepsilon.$$

When income is improperly omitted from this (any) demand equation, the resulting "model" is

$$\ln G = \beta_1 + \beta_2 \ln \mathit{Price} + w,$$

where $w = \beta_3 \ln \mathit{Income} + \varepsilon$. Linear regression of ln$G$ on a constant and ln*Price* does not consistently estimate (β_1, β_2) if ln*Price* is correlated with w. It surely will be in aggregate time-series data. The omitted variable reappears in the equation, in the disturbance, causing **omitted variable bias** in the least squares estimator of the misspecified equation.

Berry, Levinsohn, and Pakes (1995) examined the equilibrium in the U.S. automobile market. The centerpiece of the model is a random utility, multinomial choice model. For consumer i in market t, the utility of brand choice j is $U_{ijt} = U(w_i, p_{jt}, \mathbf{x}_{jt}, \mathbf{f}_{jt} | \boldsymbol{\beta})$, where w_i is individual heterogeneity, p_{jt} is the price, $\mathbf{x}_{jt}$ is a vector of observed attributes, and $\mathbf{f}_{jt}$ is a vector of unobserved features of the brand. Under the assumptions of random utility maximizing, and

aggregating over individuals, the model produces a market share equation, $s_{jt} = s_j(\mathbf{p}_t, \mathbf{X}_t, \mathbf{f}_t | \boldsymbol{\beta})$. Because $\mathbf{f}_t$ is unobserved features that consumers care about (i.e., $\mathbf{f}_{jt}$ influences the market share of brand j), and $\mathbf{f}_{jt}$ is reflected in the price of the brand, p_{jt}, $\mathbf{p}_t$ is endogenous in this choice model that is based on observed market shares.

Endogenous Treatment Effects: Krueger and Dale (1999) and Dale and Krueger (2002, 2011) examined the effect of attendance at an elite college on lifetime earnings. The regression model with a "treatment effect" dummy variable, T, which equals one for those who attended an elite college and zero otherwise, appears as

$$\ln y = \mathbf{x}'\boldsymbol{\beta} + \delta T + \varepsilon.$$

Least squares regression of a measure of earnings, ln y, on $\mathbf{x}$ and T attempts to produce an estimate of δ, the impact of the treatment. It seems inevitable, however, that some unobserved determinants of lifetime earnings, such as ambition, inherent abilities, persistence, and so on would also determine whether the individual had an opportunity to attend an elite college. If so, then the least squares estimator of δ will inappropriately attribute the effect to the treatment, rather than to these underlying factors. Least squares will not consistently estimate δ, ultimately because of the correlation between T and ε.

In order to quantify definitively the impact of attendance at an elite college on the individuals who did so, the researcher would have to conduct an impossible experiment. Individuals in the sample would have to be observed twice, once having attended the elite college and a second time (in a second lifetime) without having done so. Whether comparing individuals who attended elite colleges to other individuals who did not adequately measures the **effect of the treatment on the treated** individuals is the subject of a vast current literature. See, for example, Imbens and Wooldridge (2009) for a survey.

Simultaneous Equations: In an equilibrium model of price and output determination in a market, there would be equations for both supply and demand. For example, a model of output and price determination in a product market might appear,

$$\begin{aligned}
\text{(Demand)} \quad & \textit{Quantity}_D = \alpha_0 + \alpha_1\, \textit{Price} + \alpha_2\, \textit{Income} + \varepsilon_D, \\
\text{(Supply)} \quad & \textit{Quantity}_S = \beta_0 + \beta_1\, \textit{Price} + \beta_2\, \textit{Input Price} + \varepsilon_S, \\
\text{(Equilibrium)} \quad & \textit{Quantity}_D = \textit{Quantity}_S.
\end{aligned}$$

Consider attempting to estimate the parameters of the demand equation by regression of a time series of equilibrium quantities on equilibrium prices and incomes. The equilibrium price is determined by the equation of the two quantities. By imposing the equilibrium condition, we can solve for $\textit{Price} = (\alpha_0 - \beta_0 + \alpha_2\, \textit{Income} - \beta_2\, \textit{Input Price} + \varepsilon_D - \varepsilon_S)/(\beta_1 - \alpha_1)$. The implication is that *Price* is correlated with ε_D—if an external shock causes ε_D to change, that induces a shift in the demand curve and ultimately causes a new equilibrium *Price*. Least squares regression of quantity on *Price* and *Income* does not estimate the parameters of the demand equation consistently. This "feedback" between ε_D and *Price* in this model produces **simultaneous equations bias** in the least squares estimator.

Dynamic Panel Data Models: In Chapter 11, we will examine a dynamic **random effects** model of the form $y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \varepsilon_{it} + u_i$ where u_i contains the time-invariant unobserved features of individual i. Clearly, in this case, the regressor $y_{i,t-1}$ is correlated with the disturbance, $(\varepsilon_{it} + u_i)$—the unobserved heterogeneity is present in y_{it} in every period. In Chapter 13, we will examine a model for municipal expenditure of the form $S_{it} = f(S_{i,t-1}, \ldots) + \varepsilon_{it}$. The disturbances are assumed to be freely correlated across periods, so both $S_{i,t-1}$ and ε_{it} are correlated with $\varepsilon_{i,t-1}$. It follows that they are correlated with each other, which means that this model, even without time-persistent effects, does not satisfy the assumptions of the linear regression model. The regressors and disturbances are correlated.

Omitted Parameter Heterogeneity: Many cross-country studies of economic growth have the following structure (greatly simplified for purposes of this example),

$$\Delta \ln Y_{it} = \alpha_i + \theta_i t + \beta_i \Delta \ln Y_{i,t-1} + \varepsilon_{it},$$

where $\Delta \ln Y_{it}$ is the growth rate of country i in year t.[1] Note that the coefficients in the model are country specific. What does least squares regression of growth rates of income on a time trend and lagged growth rates estimate? Rewrite the growth equation as

$$\begin{aligned}\Delta \ln Y_{it} &= \alpha + \theta t + \beta(\Delta \ln Y_{i,t-1}) + (\alpha_i - \alpha) + (\theta_i - \theta)t + (\beta_i - \beta)(\Delta \ln Y_{i,t-1}) + \varepsilon_{it} \\ &= \alpha + \theta t + \beta(\Delta \ln Y_{i,t-1}) + w_{it}.\end{aligned}$$

We assume that the "average" parameters, α, θ, and β, are meaningful fixed parameters to be estimated. Does the least squares regression of $\Delta \ln Y_{it}$ on a constant, t, and $\Delta \ln Y_{i,t-1}$ estimate these parameters consistently? We might assume that the cross-country variation in the constant terms is purely random, and the time trends, θ_i, are driven by purely exogenous factors. But the differences across countries of the convergence parameters, β_i, are likely at least to be correlated with the growth in incomes in those countries, which will induce a correlation between the lagged income growth and the term $(\beta_i - \beta)$ embedded in w_{it}. If $(\beta_i - \beta)$ is random noise that is uncorrelated with $\Delta \ln Y_{i,t-1}$, then $(\beta_i - \beta)\, \Delta \ln Y_{i,t-1}$ will be also.

Measurement Error: Ashenfelter and Krueger (1994), Ashenfelter and Zimmerman (1997), and Bonjour et al. (2003) examined applications in which an earnings equation,

$$y_{i,t} = f(Education_{i,t},\ \ldots) + \varepsilon_{i,t},$$

is specified for sibling pairs (twins) $t = 1, 2$ for n families. Education is a variable that is inherently unmeasurable; years of schooling is typically the best **proxy variable** available. Consider, in a very simple model, attempting to estimate the parameters of

$$y_{it} = \beta_1 + \beta_2\, Education_{it} + \varepsilon_{it},$$

by a regression of $Earnings_{it}$ on a constant and $Schooling_{it,}$ with

$$Schooling_{it} = Education_{it} + u_{it},$$

where u_{it} is the measurement error. By a simple substitution, we find

$$y_{it} = \beta_1 + \beta_2\, Schooling_{it} + w_{it},$$

where $w_{it} = \varepsilon_{it} - \beta_2 u_{it}$. *Schooling* is clearly correlated with $w_{it} = (\varepsilon_{it} - \beta_2 u_{it})$. The interpretation is that at least some of the variation in *Schooling* is due to variation in the measurement error, u_{it}. Because schooling is correlated with w_{it}, it is endogenous in the earnings equation, and least squares is not a suitable estimator. As we will show later, in cases such as this one, the mismeasurement of a relevant variable causes a particular form of inconsistency, **attenuation bias**, in the estimator of β_2.

Nonrandom Sampling: In a model of the effect of a training program, an employment program, or the labor supply behavior of a particular segment of the labor force, the sample of observations may have voluntarily selected themselves into the observed sample. The Job Training Partnership Act (JTPA) was a job training program intended to provide employment assistance to disadvantaged youth. Anderson et al. (1991) found that for a sample that they examined, the program appeared to be administered most often to the best qualified applicants. In an earnings equation estimated for such a nonrandom sample, the implication is that the disturbances are not truly random. For the application just described, for example, on average, the disturbances are unusually high compared to the

[1] See, for example, Lee, Pesaran, and Smith (1997).

full population. Merely unusually high would not be a problem save for the general finding that the explanation for the nonrandomness is found at least in part in the variables that appear elsewhere in the model. This nonrandomness of the sample translates to a form of omitted variable bias known as **sample selection bias**.

Attrition: We can observe two closely related important cases of nonrandom sampling. In panel data studies of firm performance, the firms still in the sample at the end of the observation period are likely to be a subset of those present at the beginning—those firms that perform badly, "fail," or drop out of the sample. Those that remain are unusual in the same fashion as the previous sample of JTPA participants. In these cases, least squares regression of the performance variable on the covariates (whatever they are) suffers from a form of selection bias known as **survivorship bias**. In this case, the distribution of outcomes, firm performances for the survivors is systematically higher than that for the population of firms as a whole. This produces a phenomenon known as **truncation bias**. In clinical trials and other statistical analyses of health interventions, subjects often drop out of the study for reasons related to the intervention itself—for a quality of life intervention such as a drug treatment for cancer, subjects may leave because they recover and feel uninterested in returning for the exit interview, or they may pass away or become incapacitated and be unable to return. In either case, the statistical analysis is subject to **attrition bias**. The same phenomenon may impact the analysis of panel data in health econometrics studies. For example, Contoyannis, Jones, and Rice (2004) examined self-assessed health outcomes in a long panel data set extracted from the British Household Panel Survey. In each year of the study, a significant number of the observations were absent from the next year's data set, with the result that the sample was winnowed significantly from the beginning to the end of the study.

In all the cases listed in Example 8.1, the term *bias* refers to the result that least squares (or other conventional modifications of least squares) is an inconsistent (persistently biased) estimator of the coefficients of the model of interest. Though the source of the result differs considerably from setting to setting, all ultimately trace back to endogeneity of some or all of the right-hand-side variables and this, in turn, translates to correlation between the regressors and the disturbances. These can be broadly viewed in terms of some specific effects:

- Omitted variables, either observed or unobserved,
- Feedback effects,
- Dynamic effects,
- Endogenous sample design, and so on.

There are three general solutions to the problem of constructing a consistent estimator. In some cases, a more detailed, **structural specification** of the model can be developed. These usually involve specifying additional equations that explain the correlation between $\mathbf{x}_i$ and ε_i in a way that enables estimation of the full set of parameters of interest. We will develop a few of these models in later chapters, including, for example, Chapter 19, where we consider Heckman's (1979) model of sample selection. The second approach, which is becoming increasingly common in contemporary research, is the method of **instrumental variables**. The method of instrumental variables is developed around the following estimation strategy: Suppose that in the model of (8-1), the K variables $\mathbf{x}_i$ may be correlated with ε_i. Suppose as well that there exists a set of L variables $\mathbf{z}_i$, such that $\mathbf{z}_i$ is correlated with $\mathbf{x}_i$, but not with ε_i. We cannot estimate $\boldsymbol{\beta}$ consistently by using the familiar least squares estimator. But the assumed lack of correlation between $\mathbf{z}_i$ and ε_i implies a set of relationships that may allow us construct a consistent estimator

of $\boldsymbol{\beta}$ by using the assumed relationships among $\mathbf{z}_i$, $\mathbf{x}_i$, and ε_i. A third method that builds off the second augments the equation with a constructed exogenous variable (or set of variables), C_i, such that in the presence of the **control function**, $C, \mathbf{x}_{i2}$ is not correlated with ε_i. The best known approach to the sample selection problem turns out to be a control function estimator. The method of two-stage least squares can be construed as another.

This chapter will develop the method of instrumental variables as an extension of the models and estimators that have been considered in Chapters 2–7. Section 8.2 will formalize the model in a way that provides an estimation framework. The method of **instrumental variables (IV)** estimation and **two-stage least squares (2SLS)** is developed in detail in Section 8.3. Two tests of the model specification are considered in Section 8.4. A particular application of the estimation with measurement error is developed in detail in Section 8.5. Section 8.6 will consider nonlinear models and begin the development of the generalized method of moments (GMM) estimator. The IV estimator is a powerful tool that underlies a great deal of contemporary empirical research. A shortcoming, the problem of weak instruments, is considered in Section 8.7. Finally, some observations about instrumental variables and the search for causal effects are presented in Section 8.8.

This chapter will develop the fundamental results for IV estimation. The use of instrumental variables will appear in many applications in the chapters to follow, including multiple equations models in Chapter 10, the panel data methods in Chapter 11, and in the development of the generalized method of moments in Chapter 13.

8.2 ASSUMPTIONS OF THE EXTENDED MODEL

The assumptions of the linear regression model, laid out in Chapters 2 and 4, are:

A.1. Linearity: $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$.
A.2. Full rank: The $n \times K$ sample data matrix, $\mathbf{X}$, has full column rank.
A.3. Exogeneity of the independent variables: $E[\varepsilon_i | x_{j1}, x_{j2}, \ldots, x_{jk}] = 0, i, j = 1, \ldots, n$. There is no correlation between the disturbances and the independent variables.
A.4. Homoscedasticity and nonautocorrelation: Each disturbance, ε_i, has the same finite variance, σ^2, and is uncorrelated with every other disturbance, ε_j, conditioned on $\mathbf{X}$.
A.5. Stochastic or nonstochastic data: $(x_{i1}, x_{i2}, \ldots, x_{iK}), i = 1, \ldots, n$.
A.6. Normal distribution: The disturbances are normally distributed.

We will maintain the important result that plim $(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}_{\mathbf{xx}}$. The basic assumptions of the regression model have changed, however. First, A.3 (no correlation between $\mathbf{x}$ and ε) is, under our new assumptions,

$$\textbf{A.I3.} \quad E[\varepsilon_i | \mathbf{x}_i] = \eta.$$

We interpret Assumption A.I3 to mean that the regressors now provide information about the expectations of the disturbances. The important implication of A.I3 is that the disturbances and the regressors are now correlated. Assumption A.I3 implies that

$$E[\mathbf{x}_i \varepsilon_i] = \boldsymbol{\gamma} \tag{8-2}$$

for some nonzero $\boldsymbol{\gamma}$. If the data are well behaved, then we can apply Theorem D.5 (Khinchine's theorem) to assert that,

$$\text{plim } (1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \boldsymbol{\gamma}. \tag{8-3}$$

Notice that the original model results if $\eta = 0$. The implication of (8-3) is that the regressors, $\mathbf{X}$, are no longer exogenous. Assumptions A.4–A.6 will be secondary considerations in the discussion of this chapter. We will develop some essential results with A.4 in place, then turn to robust inference procedures that do not rely on it. As before, we will characterize the essential results based on random sampling from the joint distribution of y and $\mathbf{x}$ (and $\mathbf{z}$). Assumption A.6 is no longer relevant—all results from here forward will be based on asymptotic distributions.

We now assume that there is an additional set of variables, $\mathbf{z} = (z_1, \ldots, z_L)$, that have two essential properties:

1. **Relevance**: They are correlated with the independent variables, $\mathbf{X}$.
2. **Exogeneity**: They are uncorrelated with the disturbance.

We will formalize these notions as we proceed. In the context of our model, variables that have these two properties are instrumental variables. We assume the following:

A.I7. $[\mathbf{x}_i, \mathbf{z}_i, \varepsilon_i]$, $i = 1, \ldots, n$, are an i.i.d. sequence of random variables.
A.I8a. $E[x_{ik}^2] = \mathbf{Q}_{\mathbf{xx},kk} < \infty$, a finite constant, $k = 1, \ldots, K$.
A.I8b. $E[z_{il}^2] = \mathbf{Q}_{\mathbf{zz},ll} < \infty$, a finite constant, $l = 1, \ldots, L$.
A.I8c. $E[z_{il}x_{ik}] = \mathbf{Q}_{\mathbf{zx},lk} < \infty$, a finite constant, $l = 1, \ldots, L, k = 1, \ldots, K$.
A.I9. $E[\varepsilon_i | \mathbf{z}_i] = 0$.

In later work in time-series models, it will be important to relax assumption A.I7. Finite means of $\mathbf{z}_l$ follows from A.I8b. Using the same analysis as in Section 4.4, we have

$$\text{plim } (1/n)\mathbf{Z}'\mathbf{Z} = \mathbf{Q}_{\mathbf{zz}}, \text{ a finite, positive definite matrix (\textit{well-behaved data})},$$
$$\text{plim } (1/n)\mathbf{Z}'\mathbf{X} = \mathbf{Q}_{\mathbf{zx}}, \text{ a finite, } L \times K \text{ matrix with rank } K \text{ (\textit{relevance})},$$
$$\text{plim } (1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0} \text{ (\textit{exogeneity})}.$$

In our statement of the regression model, we have assumed thus far the special case of $\eta = 0$; $\boldsymbol{\gamma} = \mathbf{0}$ follows.

For the present, we will assume that $L = K$—there are the same number of instrumental variables as there are right-hand-side variables in the equation. Recall in the introduction and in Example 8.1, we partitioned $\mathbf{x}$ into $\mathbf{x}_1$, a set of K_1 exogenous variables, and $\mathbf{x}_2$, a set of K_2 endogenous variables, on the right-hand side of (8-1). In nearly all cases in practice, the problem of endogeneity is attributable to one or a small number of variables in $\mathbf{x}$. In the Krueger and Dale (1999) study of endogenous treatment effects in Example 8.1, we have a single endogenous variable in the equation, the treatment dummy variable, T. The implication for our formulation here is that in such a case, the K_1 variables $\mathbf{x}_1$ will be K_1 of the variables in $\mathbf{Z}$ and the K_2 remaining variables will be other exogenous variables that are not the same as $\mathbf{x}_2$. The usual interpretation will be that these K_2 variables, $\mathbf{z}_2$, are the instruments for $\mathbf{x}_2$ while the $\mathbf{x}_1$ variables are instruments for themselves. To continue the example, the matrix $\mathbf{Z}$ for the endogenous treatment effects model would contain the K_1 columns of $\mathbf{X}$ and an additional instrumental variable, $\mathbf{z}$, for the treatment dummy variable. In the simultaneous equations model of supply and demand, the endogenous right-hand-side variable is $x_2 = \textit{price}$ while the exogenous variables are (1, *Income*). One might suspect (correctly), that in this model, a set of instrumental variables would be $\mathbf{z} = (1, \textit{Income}, \textit{InputPrice})$. In terms of the underlying relationships among the variables, this intuitive understanding will provide a reliable

guide. For reasons that will be clear shortly, however, it is necessary statistically to treat $\mathbf{Z}$ as the instruments for $\mathbf{X}$ in its entirety.

There is a second subtle point about the use of instrumental variables that will likewise be more evident below. The relevance condition must actually be a statement of conditional correlation. Consider, once again, the treatment effects example, and suppose that z is the instrumental variable in question for the treatment dummy variable T. The relevance condition as stated implies that the correlation between z and $(\mathbf{x},T)$ is nonzero. Formally, what will be required is that the *conditional* correlation of z with $T|\mathbf{x}$ be nonzero. One way to view this is in terms of a projection; the instrumental variable z is relevant if the coefficient on z in the projection of T on $(\mathbf{x}, z)$ is nonzero. Intuitively, z must provide information about the movement of T that is not provided by the $\mathbf{x}$ variables that are already in the model.

8.3 INSTRUMENTAL VARIABLES ESTIMATION

For the general model of Section 8.2, we lose most of the useful results we had for least squares. We will consider the implications for least squares and then construct an alternative estimator for $\boldsymbol{\beta}$ in this extended model.

8.3.1 LEAST SQUARES

The least squares estimator, $\mathbf{b}$, is no longer unbiased,

$$E[\mathbf{b}|\mathbf{X}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\eta} \neq \boldsymbol{\beta},$$

so the Gauss–Markov theorem no longer holds. The estimator is also inconsistent,

$$\operatorname{plim}\mathbf{b} = \boldsymbol{\beta} + \operatorname{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\operatorname{plim}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right) = \boldsymbol{\beta} + \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\gamma} \neq \boldsymbol{\beta}. \tag{8-4}$$

(The asymptotic distribution is considered in the exercises.) The inconsistency of least squares is not confined to the coefficients on the endogenous variables. To see this, apply (8-4) to the treatment effects example discussed earlier. In that case, all but the last variable in $\mathbf{X}$ are uncorrelated with $\boldsymbol{\varepsilon}$. This means that

$$\operatorname{plim}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \gamma_K \end{pmatrix} = \gamma_K \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

It follows that for this special case, the result in (8-4) is

$$\operatorname{plim}\mathbf{b} = \boldsymbol{\beta} + \gamma_K \times \text{the last column of } \mathbf{Q}_{\mathbf{XX}}^{-1}.$$

There is no reason to expect that any of the elements of the last column of $\mathbf{Q}_{\mathbf{XX}}^{-1}$ will equal zero. The implication is that even though only one of the variables in $\mathbf{X}$ is correlated with $\boldsymbol{\varepsilon}$, all of the elements of $\mathbf{b}$ are inconsistent, not just the estimator of the coefficient

on the endogenous variable. This effect is called **smearing**; the inconsistency due to the endogeneity of the one variable is smeared across all of the least squares estimators.

8.3.2 THE INSTRUMENTAL VARIABLES ESTIMATOR

Because $E[\mathbf{z}_i\varepsilon_i] = 0$ and all terms have finite variances, it follows that $\operatorname{plim}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right) = 0$. Therefore,

$$\operatorname{plim}\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = \left[\operatorname{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]\boldsymbol{\beta} + \operatorname{plim}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right) = \left[\operatorname{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]\boldsymbol{\beta}. \quad \textbf{(8-5)}$$

We have assumed that $\mathbf{Z}$ has the same number of variables as $\mathbf{X}$. For example, suppose in our consumption function that $\mathbf{x}_t = [1, Y_t]$ when $\mathbf{z}_t = [1, Y_{t-1}]$. We have also assumed that the rank of $\mathbf{Z}'\mathbf{X}$ is K, so now $\mathbf{Z}'\mathbf{X}$ is a square matrix. It follows that

$$\left[\operatorname{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]^{-1}\operatorname{plim}\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = \boldsymbol{\beta},$$

which leads us to the **instrumental variable estimator**,

$$\mathbf{b}_{\mathrm{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}. \quad \textbf{(8-6)}$$

For a model with a constant term and a single x and instrumental variable z, we have

$$b_{\mathrm{IV}} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} = \frac{Cov(z, y)}{Cov(z, x)}.$$

We have already proved that $\mathbf{b}_{\mathrm{IV}}$ is consistent. We now turn to the **asymptotic distribution**. We will use the same method as in Section 4.4.3. First,

$$\sqrt{n}(\mathbf{b}_{\mathrm{IV}} - \boldsymbol{\beta}) = \left(\frac{\mathbf{Z}'\boldsymbol{X}}{n}\right)^{-1}\frac{1}{\sqrt{n}}\mathbf{Z}'\boldsymbol{\varepsilon},$$

which has the same **limiting distribution** as $\mathbf{Q}_{\mathbf{zx}}^{-1}[(1/\sqrt{n})\mathbf{Z}'\boldsymbol{\varepsilon}]$. Our analysis of $(1/\sqrt{n})\mathbf{Z}'\boldsymbol{\varepsilon}$ can be the same as that of $(1/\sqrt{n})\boldsymbol{X}'\boldsymbol{\varepsilon}$ in Section 4.4.3, so it follows that

$$\left(\frac{1}{\sqrt{n}}\mathbf{Z}'\boldsymbol{\varepsilon}\right) \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\sigma}^2\boldsymbol{Q}_{zz}],$$

and

$$\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{Z}'\boldsymbol{\varepsilon}\right) \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}_{\mathbf{zx}}^{-1}\mathbf{Q}_{\mathbf{zz}}\mathbf{Q}_{\mathbf{xz}}^{-1}].$$

This step completes the derivation for the next theorem.

THEOREM 8.1 Asymptotic Distribution of the Instrumental Variables Estimator

If Assumptions A.1–A5, A.I7, A.I8a–c, and A.I9 all hold for $[y_i, \mathbf{x}_i, \mathbf{z}_i, \varepsilon_i]$, *where* $\mathbf{z}$ *is a valid set of* $L = K$ *instrumental variables, then the asymptotic distribution of the instrumental variables estimator* $\mathbf{b}_{\mathrm{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$ *is*

$$\mathbf{b}_{\mathrm{IV}} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}_{\mathbf{zx}}^{-1}\mathbf{Q}_{\mathbf{zz}}\mathbf{Q}_{\mathbf{xz}}^{-1}\right]. \tag{8-7}$$

where $\mathbf{Q}_{\mathbf{zx}} = \mathrm{plim}(\mathbf{Z}'\mathbf{X}/n)$ *and* $\mathbf{Q}_{\mathbf{zz}} = \mathrm{plim}(\mathbf{Z}'\mathbf{Z}/n)$. *If Assumption A4 is dropped, then the asymptotic covariance matrix will be the population counterpart to the robust estimators in (8-8h) or (8-8c), below.*

8.3.3 ESTIMATING THE ASYMPTOTIC COVARIANCE MATRIX

To estimate the **asymptotic covariance matrix**, we will require an estimator of σ^2. The natural estimator is

$$\hat{\sigma}^2 = \frac{1}{n - K}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\mathbf{b}_{\mathrm{IV}})^2.$$

The correction for degrees of freedom is unnecessary, as all results here are asymptotic, and $\hat{\sigma}^2$ would not be unbiased in any event. Nonetheless, it is standard practice to make the degrees of freedom correction. Using the same approach as in Section 4.4.2 for the regression model, we find that $\hat{\sigma}^2$ is a **consistent estimator** of σ^2. We will estimate $\mathrm{Asy.Var}[\mathbf{b}_{\mathrm{IV}}]$ with

$$\begin{aligned}\mathrm{Est.Asy.Var}[\mathbf{b}_{\mathrm{IV}}] &= \frac{1}{n}\left(\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}\right)\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{Z}'\mathbf{Z}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{Z}}{n}\right)^{-1} \\ &= \hat{\sigma}^2(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Z})(\mathbf{X}'\mathbf{Z})^{-1}.\end{aligned} \tag{8-8}$$

The estimator in (8-8) is based on Assumption A.4, homoscedasticity and nonautocorrelation. By writing the IV estimator as

$$\mathbf{b}_{IV} = \boldsymbol{\beta} + \left[\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\right]^{-1}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i$$

we can use the same logic as in (4-35)–(4-37) and (4-40)–(4-42) to construct estimators of the asymptotic covariance matrix that are robust to heteroscedasticity,

$$\begin{aligned}\mathrm{Est.Asy.Var}[\mathbf{b}_{IV}] &= \left[\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\right]^{-1}\left[\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i'\hat{\varepsilon}_i^2\right]\left[\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{z}_i\right]^{-1} \\ &= n(\mathbf{Z}'\mathbf{X})^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i'\hat{\varepsilon}_i^2\right](\mathbf{X}'\mathbf{Z})^{-1},\end{aligned} \tag{8-8h}$$

and to clustering,

$$\text{Est.Asy.Var}[\mathbf{b}_{IV}] = C(\mathbf{Z}'\mathbf{X})^{-1}\left[\left(\frac{C}{C-1}\right)\frac{1}{C}\sum_{c=1}^{C}\left(\sum_{i=1}^{N_c}\mathbf{z}_{ic}\hat{\varepsilon}_{ic}\right)\left(\sum_{i=1}^{N_c}\mathbf{z}_{ic}\hat{\varepsilon}_{ic}\right)'\right](\mathbf{X}'\mathbf{Z})^{-1}, \quad \textbf{(8-8c)}$$

respectively.

8.3.4 MOTIVATING THE INSTRUMENTAL VARIABLES ESTIMATOR

In obtaining the IV estimator, we relied on the solutions to the equations in (8-5), $\text{plim}(\mathbf{Z}'\mathbf{y}/n) = \text{plim}(\mathbf{Z}'\mathbf{X}/n)\boldsymbol{\beta}$ or $\mathbf{Q}_{\mathbf{Zy}} = \mathbf{Q}_{\mathbf{ZX}}\boldsymbol{\beta}$. The IV estimator is obtained by solving this set of K **moment equations**. Because this is a set of K equations in K unknowns, if $\mathbf{Q}_{\mathbf{ZX}}^{-1}$ exists, then there is an exact solution for $\boldsymbol{\beta}$, given in (8-6). The corresponding moment equations if only $\mathbf{X}$ is used would be

$$\text{plim}(\mathbf{X}'\mathbf{y}/n) = \text{plim}(\mathbf{X}'\mathbf{X}/n)\boldsymbol{\beta} + \text{plim}(\mathbf{X}'\boldsymbol{\varepsilon}/n) = \text{plim}(\mathbf{X}'\mathbf{X}/n)\boldsymbol{\beta} + \boldsymbol{\gamma}$$

or

$$\mathbf{Q}_{\mathbf{Xy}} = \mathbf{Q}_{\mathbf{XX}}\boldsymbol{\beta} + \boldsymbol{\gamma},$$

which is, without further restrictions, K equations in $2K$ unknowns. There are insufficient equations to solve this system for either $\boldsymbol{\beta}$ or $\boldsymbol{\gamma}$. The further restrictions that would allow estimation of $\boldsymbol{\beta}$ would be $\boldsymbol{\gamma} = \mathbf{0}$; this is precisely the exogeneity assumption A.3. The implication is that the parameter vector $\boldsymbol{\beta}$ is not **identified** in terms of the moments of $\mathbf{X}$ and $\mathbf{y}$ alone—there does not exist a solution. But it is identified in terms of the moments of $\mathbf{Z}$, $\mathbf{X}$, and $\mathbf{y}$, plus the K restrictions imposed by the exogeneity assumption, and the relevance assumption that allows computation of $\mathbf{b}_{\text{IV}}$.

By far the most common application of IV estimation involves a single endogenous variable in a multiple regression model,

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots x_{i\,\text{K}}\beta_K + \varepsilon_i,$$

with $\text{Cov}(x_K, \varepsilon) \neq 0$. The instrumental variable estimator, based on instrument z, proceeds from two conditions:

- Relevance: $\text{Cov}(z, x_K | x_1, \ldots, x_{K-1}) \neq 0$,
- Exogeneity: $E(\varepsilon | z) = 0$.

In words, the relevance condition requires that the instrument provide explanatory power of the variation of the endogenous variable beyond that provided by the other exogenous variables already in the model. A theoretical basis for the relevance condition would be a projection of x_K on all of the exogenous variables in the model,

$$x_K = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_{K-1}x_{K-1} + \lambda z + u.$$

In this form, the relevance condition will require $\lambda \neq 0$. This can be verified empirically; in a linear regression of x_K on $(x_1, \ldots, x_{K-1}, z)$, one would expect the least squares estimate of λ to be statistically different from zero. The exogeneity condition is not directly testable. It is entirely theoretical. (The Hausman and Wu tests suggested below are only indirect.)

Consider these results in the context of a simplified model,

$$y = \beta x + \delta T + \varepsilon.$$

In order for least squares consistently to estimate δ (and β), it is assumed that movements in T are exogenous to the model, so that covariation of y and T is explainable by the movement of T and not by the movement of ε. When T and ε are correlated and ε varies through some factor not in the equation, the movement of y will appear to be induced by variation in T when it is actually induced by variation in ε which is transmitted through T. If T is exogenous, that is, not correlated with ε, then movements in ε will not "cause" movements in T (we use the term *cause* very loosely here) and will thus not be mistaken for exogenous variation in T. The exogeneity assumption plays precisely this role. What is needed, then, to identify δ is movement in T that is definitely not induced by movement in ε? Enter the instrumental variable, z. If z is an instrumental variable with $\text{Cov}(z, T) \neq 0$ and $\text{Cov}(z, \varepsilon) = 0$, then movement in z provides the variation that we need. If we can consider doing this exercise experimentally, in order to measure the "causal effect" of movement in T, we would change z and then measure the per unit change in y associated with the change in T, knowing that the change in T was induced only by the change in z, not ε. That is, the estimator of δ is $(\Delta y/\Delta z)/(\Delta T/\Delta z)$.

Example 8.2 Instrumental Variable Analysis

Grootendorst (2007) and Deaton (1997) recount what appears to be the earliest application of the method of instrumental variables:

Although IV theory has been developed primarily by economists, the method originated in epidemiology. IV was used to investigate the route of cholera transmission during the London cholera epidemic of 1853–54. A scientist from that era, John Snow, hypothesized that cholera was waterborne. To test this, he could have tested whether those who drank purer water had lower risk of contracting cholera. In other words, he could have assessed the correlation between water purity (x) and cholera incidence (y). Yet, as Deaton (1997) notes, this would not have been convincing: "The people who drank impure water were also more likely to be poor, and to live in an environment contaminated in many ways, not least by the 'poison miasmas' that were then thought to be the cause of cholera." Snow instead identified an instrument that was strongly correlated with water purity yet uncorrelated with other determinants of cholera incidence, both observed and unobserved. This instrument was the identity of the company supplying households with drinking water. At the time, Londoners received drinking water directly from the Thames River. One company, the Lambeth Water Company, drew water at a point in the Thames above the main sewage discharge; another, the Southwark and Vauxhall Company, took water below the discharge. Hence the instrument z was strongly correlated with water purity x. The instrument was also uncorrelated with the unobserved determinants of cholera incidence (y). According to Snow (1855, pp. 74–75), the households served by the two companies were quite similar; indeed: "the mixing of the supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. . . . The experiment, too, is on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and in most cases, without their knowledge; one group supplied with water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity.

A stylized sketch of Snow's experiment is useful for suggesting how the instrumental variable estimator works. The theory states that

$$\text{Cholera Occurrence} = f(\text{Impure Water, Other Factors}).$$

For simplicity, denote the occurrence of cholera in household i with

$$c_i = \alpha + \delta w_i + \varepsilon_i,$$

where c_i represents the presence of cholera, $w_i = 1$ if the household has (measurably) impure water, 0 if not, and δ is the sought after causal effect of the water impurity on the prevalence of cholera. It would seem that one could simply compute $d = (\bar{c}|w = 1) - (\bar{c}|w = 0)$, which would be the result of a regression of c on w, to assess the effect of impure water on the prevalence of cholera. The flaw in this strategy is that a cholera prone environment, u, affects both the water quality, w, and the other factors, ε. Interpret this to say that both Cov(w, u) and Cov(ε, u) are nonzero and therefore, Cov(w, ε) is nonzero. The endogeneity of w in the equation invalidates the regression estimator of δ. The pernicious effect of the common influence, u, works through the unobserved factors, ε. The implication is that $E[c|w] \neq \alpha + \delta w$ because $E[\varepsilon|w] \neq 0$. Rather,

$$E[c|w = 1] = \alpha + \delta + E[\varepsilon|w = 1]$$
$$E[c|w = 0] = \alpha + \cdots + E[\varepsilon|w = 0]$$

so,

$$E[c|w = 1] - E[c|w = 0] = \delta + \{E[\varepsilon|w = 1] - E[\varepsilon|w = 0]\}.$$

It follows that comparing the cholera rates of households with bad water to those with good water, $P[c|w = 1] - P[c|w = 0]$, does not reveal only the impact of the bad water on the prevalence of cholera. It partly reveals the impact of bad water on some other factor in ε that, in turn, impacts the cholera prevalence. Snow's IV approach based on the water supplying company works as follows: Define

$$\mathrm{I} = 1 \text{ if water is supplied by Lambeth,}$$
$$0 \text{ if Southwark and Vauxhall.}$$

To establish the *relevance* of this instrument, Snow argued that

$$E[w|I = 1] \neq E[w|I = 0].$$

Snow's theory was that water supply was the culprit, and Lambeth supplied purer water than Southwark. This can be verified observationally. The instrument is *exogenous* if

$$E[\varepsilon|I = 1] = E[\varepsilon|I = 0].$$

This is the theory of the instrument. Water is supplied randomly to houses. Homeowners do not even know who supplies their water. The assumption is not that the unobserved factor, ε, is unaffected by the water quality. It is that the other factors, not the water quality, are present in equal measure in households supplied by the two different water suppliers. This is Snow's argument that the households supplied by the two water companies are otherwise similar. The assignment is random. To use the instrument, we note $E[c|I] = \delta E[w|I] + E[\varepsilon|I]$, so

$$E[c|I = 1] = \alpha + \delta E[w|I = 1] + E[\varepsilon|I = 1],$$
$$E[c|I = 0] = \alpha + \delta E[w|I = 0] + E[\varepsilon|I = 0].$$

This produces an estimating equation,

$$E[c|I = 1] - E[c|I = 0] = \delta\{E[w|I = 1] - E[w|I = 0]\}$$
$$+ \{E[\varepsilon|I = 1] - E[\varepsilon|I = 0]\}.$$

The second term in braces is zero if I is exogenous, which was assumed. The IV estimator is then

$$\hat{\delta} = \frac{E[c|I = 1] - E[c|I = 0]}{E[w|I = 1] - E[w|I = 0]}.$$

Note that the nonzero denominator results from the relevance condition. We can see that δ is analogous to Cov(c,l)/Cov(w,l), which is (8-6).

To operationalize the estimator, we will use

$$P(c|l=1) = \hat{E}(c|l=1) = \bar{c}_1 = \text{proportion of households supplied by Lambeth that have cholera,}$$
$$P(w|l=1) = \hat{E}(w|l=1) = \bar{w}_1 = \text{proportion of households supplied by Lambeth that have bad water,}$$
$$P(c|l=0) = \hat{E}(c|l=0) = \bar{c}_0 = \text{proportion of households supplied by Vauxhall that have cholera,}$$
$$P(w|l=0) = \hat{E}(w|l=0) = \bar{w}_0 = \text{proportion of households supplied by Vauxhall that have bad water.}$$

To complete this development of Snow's experiment, we can show that the estimator $\hat{\delta}$ is an application of (8-6). Define three dummy variables, $c_i = 1$ if household i suffers from cholera and 0 if not, $w_i = 1$ if household i receives impure water and 0 if not, and $l_i = 1$ if household i receives its water from Lambeth and 0 if from Vauxhall; let **c**, **w**, and **l** denote the column vectors of n observations on the three variables; and let **i** denote a column of ones. For the model $c_i = \alpha + \delta w_i + \varepsilon_i$, we have $\mathbf{Z} = [\mathbf{i}, \mathbf{l}]$, $\mathbf{X} = [\mathbf{i}, \mathbf{w}]$, and $\mathbf{y} = \mathbf{c}$. The estimator is

$$\begin{pmatrix} a \\ d \end{pmatrix} = [\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{Z}'\mathbf{y} =$$
$$\begin{bmatrix} \mathbf{i}'\mathbf{i} & \mathbf{i}'\mathbf{w} \\ \mathbf{l}'\mathbf{i} & \mathbf{l}'\mathbf{w} \end{bmatrix}^{-1}\begin{pmatrix} \mathbf{i}'\mathbf{c} \\ \mathbf{l}'\mathbf{c} \end{pmatrix} = \begin{bmatrix} n & n\bar{w} \\ n_1 & n_1\bar{w}_1 \end{bmatrix}^{-1}\begin{pmatrix} n\bar{c} \\ n_1\bar{c}_1 \end{pmatrix} = \frac{1}{nn_1(\bar{w}_1 - \bar{w})}\begin{bmatrix} n_1\bar{w}_1 & -n\bar{w} \\ -n_1 & n \end{bmatrix}\begin{pmatrix} n\bar{c} \\ n_1\bar{c}_1 \end{pmatrix}.$$

Collecting terms, $d = (\bar{c}_1 - \bar{c})/(\bar{w}_1 - \bar{w})$. Because $n = n_0 + n_1$, $\bar{c}_1 = (n_0\bar{c}_1 + n_1\bar{c}_1)/n$ and $\bar{c} = (n_0\bar{c}_0 + n_1\bar{c}_1)/n$, so $\bar{c}_1 - \bar{c} = (n_0/n)(\bar{c}_1 - \bar{c}_0)$. Likewise, $\bar{w}_1 - \bar{w} = (n_0/n)(\bar{w}_1 - \bar{w}_0)$ so $d = (\bar{c}_1 - \bar{c}_0)/(\bar{w}_1 - \bar{w}_0) = \hat{\delta}$. This estimator based on the difference in means is the Wald (1940) estimator.

Example 8.3 Streams as Instruments

In Hoxby (2000), the author was interested in the effect of the amount of school "choice" in a school "market" on educational achievement in the market. The equations of interest were of the form

$$\frac{A_{ikm}}{\ln E_{km}} = \beta_1 C_m + \mathbf{x}'_{ikm}\boldsymbol{\beta}_2 + \bar{\mathbf{x}}'_{\cdot km}\boldsymbol{\beta}_3 + \bar{\mathbf{x}}'_{\cdot\cdot m}\boldsymbol{\beta}_4 + \varepsilon_{ikm} + \varepsilon_{km} + \varepsilon_m,$$

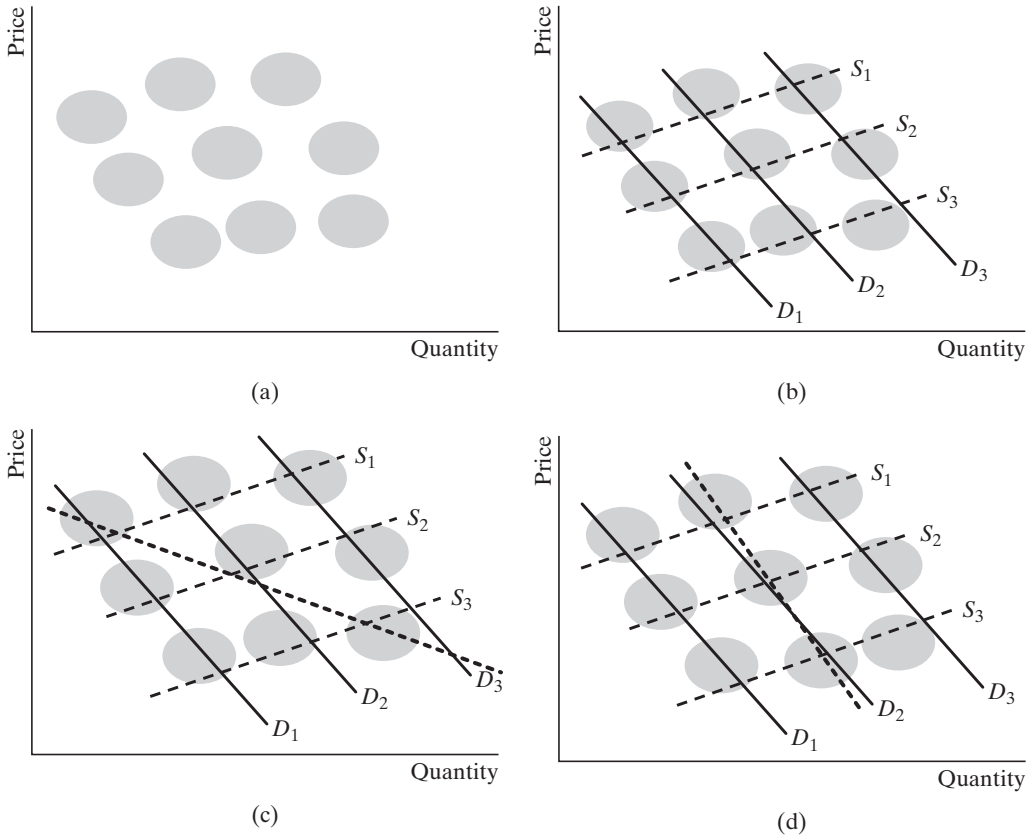
where "ikm" denotes household i in district k in market m, A_{ikm} is a measure of achievement, and E_{ikm} is per capita expenditures. The equation contains individual-level data, district means, and market means. The exogenous variables are intended to capture the different sources of heterogeneity at all three levels of aggregation. (The compound disturbance, which we will revisit when we examine panel data specifications in Chapter 10, is intended to allow for random effects at all three levels as well.) Reasoning that the amount of choice available to students, C_m, would be endogenous in this equation, the author sought a valid instrumental variable that would "explain" (be correlated with) C_m but uncorrelated with the disturbances in the equation. In the U.S. market, to a large degree, school district boundaries were set in the late 18th through the 19th centuries and handed down to present-day administrators by historical precedent. In the formative years, the author noted, district boundaries were set in response to natural travel barriers, such as rivers and streams. It follows, as she notes, that "the number of districts in a

given land area is an increasing function of the number of natural barriers"; hence, the number of streams in the physical market area provides the needed instrumental variable.[2] This study is an example of a "natural experiment," as described in Angrist and Pischke (2009).

Example 8.4 Instrumental Variable in Regression

The role of an instrumental variable in identifying parameters in regression models was developed in Working's (1926) classic application, adapted here for our market equilibrium example in Example 8.1. Figure 8.1a displays the observed data for the market equilibria in a market in which there are random disturbances (ε_S, ε_D) and variation in demanders' incomes and input prices faced by suppliers. The market equilibria in Figure 8.1a are scattered about as the aggregates of all these effects. Figure 8.1b suggests the underlying conditions of supply and demand that give rise to these equilibria. Different outcomes in the supply equation corresponding to different values of the input price and different outcomes on the demand side corresponding to different income values produce nine regimes, punctuated

FIGURE 8.1 Identifying a Demand Curve with an Instrumental Variable.



[2]The controversial topic of the study and the unconventional choice of instruments caught the attention of the popular press, for example, http://www.wsj.com/articles/SB113011672134577225 and http://www.thecrimson.com/article/2005/7/8/star-ec-prof-caught-in-academic/, and academic observers including Rothstein (2004).

by the random variation induced by the disturbances. Given the ambiguous mass of points, linear regression of quantity on price (and income) is likely to produce a result such as that shown by the heavy dotted line in Figure 8.1c. The slope of this regression barely resembles the slope of the demand equations. Faced with this prospect, how is it possible to learn about the slope of the demand curve? The experiment needed, shown in Figure 8.1d, would involve two elements: (1) Hold *Income* constant, so we can focus on the demand curve in a particular demand setting. That is the function of multiple regression—Income is included as a conditioning variable in the equation. (2) Now that we have focused on a particular set of demand outcomes (e.g., D2), move the supply curve so that the equilibria now trace out the demand function. That is the function of the changing *InputPrice*, which is the instrumental variable that we need for identification of the demand function(s) for this experiment.

8.4 TWO-STAGE LEAST SQUARES, CONTROL FUNCTIONS, AND LIMITED INFORMATION MAXIMUM LIKELIHOOD

Thus far, we have assumed that the number of instrumental variables in $\mathbf{Z}$ is the same as the number of variables (exogenous plus endogenous) in $\mathbf{X}$. In the typical application, there is one instrument for the single endogenous variable in the equation. The model specification may imply additional instruments. Recall the market equilibrium application considered in Examples 8.1 and 8.4. Suppose this were an agricultural market in which there are two exogenous conditions of supply, *InputPrice* and *Rainfall*. Then, the equations of the model are

$$\text{(Demand)} \quad Quantity_D = \alpha_0 + \alpha_1\, Price + \alpha_2\, Income + \varepsilon_D,$$

$$\text{(Supply)} \quad Quantity_S = \beta_0 + \beta_1\, Price + \beta_2\, Input\ Price + \beta_3\, Rain\ fall + \varepsilon_S,$$

$$\text{(Equilibrium)} \quad Quantity_D = Quantity_S.$$

Given the approach taken in Example 8.4, it would appear that the researcher could simply choose either of the two exogenous variables (instruments) in the supply equation for purpose of identifying the demand equation. Intuition should suggest that simply choosing a subset of the available instrumental variables would waste sample information—it seems inevitable that it will be preferable to use the full matrix $\mathbf{Z}$, even when $L > K$. (In the example above, $\mathbf{z} = (1, Income, InputPrice, Rainfall.)$ The method of two-stage least squares solves the problem of how to use all the information in the sample when $\mathbf{Z}$ contains more variables than are necessary to construct an instrumental variable estimator. We will also examine two other approaches to estimation. *The results developed here also apply to the case in which there is one endogenous variable and one instrument.*

In the model

$$y = \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda + \varepsilon,$$

where x_2 is a single variable, and there is a single instrument, z_1, that is relevant and exogenous, then the parameters of the model, $(\boldsymbol{\beta}, \lambda)$, can be estimated using the moments of $(y, \mathbf{x}_1, x_2, z_1)$. The IV estimator in (8-6) shows the one function of the moments that can be used for the estimation. In this case, $(\boldsymbol{\beta}, \lambda)$ are said to be *exactly identified*. There are exactly enough moments for estimation of the parameters. If there were a second exogenous and relevant instrument, say z_2, then we could use z_2 instead of z_1 in (8-6) and obtain a second, different estimator. In this case, the parameters are **overidentified**

in terms of the moments of $(y, \mathbf{x}_1, x_2, z_1, z_2)$. This does not mean that there is now simply a second estimator. If z_1 and z_2 are both exogenous and relevant, then any linear combination of them, $z_* = a_1 z_1 + a_2 z_2$, would also be a valid instrument. More than one IV estimator means an infinite number of possible estimators. Overidentification is qualitatively different from exact identification. The methods examined in this section are usable for overidentified models.

8.4.1 TWO-STAGE LEAST SQUARES

If $\mathbf{Z}$ contains more variables than $\mathbf{X}$, then $\mathbf{Z}'\mathbf{X}$ will be $L \times K$ with rank $K < L$ and will thus not have an inverse—(8-6) is not useable. The crucial result for estimation is plim $(\mathbf{Z}'\boldsymbol{\varepsilon}/n) = \mathbf{0}$. That is, every column of $\mathbf{Z}$ is asymptotically uncorrelated with $\boldsymbol{\varepsilon}$. That also means that every linear combination of the columns of $\mathbf{Z}$ is also uncorrelated with $\boldsymbol{\varepsilon}$, which suggests that one approach would be to choose K linear combinations of the columns of $\mathbf{Z}$. Which to choose? One obvious possibility is simply to choose K variables among the L in $\mathbf{Z}$. Discarding the information contained in the *extra* $L - K$ columns will turn out to be inefficient. A better choice that uses all of the instruments is the projection of the columns of $\mathbf{X}$ in the column space of $\mathbf{Z}$,

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{Z}\mathbf{F}. \tag{8-9}$$

The instruments in this case are linear combinations of the variables (columns) in $\mathbf{Z}$. With this choice of instrumental variables, we have

$$\begin{aligned}\mathbf{b}_{\mathrm{IV}} &= (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\ &= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}.\end{aligned} \tag{8-10}$$

The estimator of the asymptotic covariance matrix will be $\hat{\sigma}^2$ times the bracketed matrix in (8-10). The proofs of consistency and asymptotic normality for this estimator are exactly the same as before, because our proof was generic for any valid set of instruments, and $\hat{\mathbf{X}}$ qualifies.

There are two reasons for using this estimator—one practical, one theoretical. If any column of $\mathbf{X}$ also appears in $\mathbf{Z}$, then that column of $\mathbf{X}$ is reproduced exactly in $\hat{\mathbf{X}}$. This result is important and useful. Consider what is probably the typical application in which the regression contains K variables, only one of which, say, the kth, is correlated with the disturbances. We have one or more instrumental variables in hand, as well as the other $K - 1$ variables that certainly qualify as instrumental variables in their own right. Then what we would use is $\mathbf{Z} = [\mathbf{X}_{(k)}, \mathbf{z}_1, \mathbf{z}_2, \ \ldots\ ]$, where we indicate omission of the kth variable by (k) in the subscript. Another useful interpretation of $\hat{\mathbf{X}}$ is that each column is the set of fitted values when the corresponding column of $\mathbf{X}$ is regressed on all the columns of $\mathbf{Z}$. The coefficients for $\mathbf{x}_k$ are in the kth column of $\mathbf{F}$ in (8-9). It also makes clear why each $\mathbf{x}_k$ that appears in $\mathbf{Z}$ is perfectly replicated. Every $\mathbf{x}_k$ provides a perfect predictor for itself, without any help from the remaining variables in $\mathbf{Z}$. In the example, then, every column of $\mathbf{X}$ except the one that is omitted from $\mathbf{X}_{(k)}$ is replicated exactly, whereas the one that *is* omitted is replaced in $\hat{\mathbf{X}}$ by the predicted values in the regression of this variable on all the $\mathbf{z}$'s including the other $\mathbf{x}$ variables.

Of all the different linear combinations of $\mathbf{Z}$ that we might choose, $\hat{\mathbf{X}}$ is the most efficient in the sense that the asymptotic covariance matrix of an IV estimator based on a linear combination $\mathbf{Z}\mathbf{F}$ is smaller when $\mathbf{F} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ than with any other $\mathbf{F}$ that

uses all L columns of $\mathbf{Z}$; *a fortiori*, this result eliminates linear combinations obtained by dropping any columns of $\mathbf{Z}$.[3]

We close this section with some practical considerations in the use of the instrumental variables estimator. By just multiplying out the matrices in the expression, you can show that

$$\begin{aligned}\mathbf{b}_{\mathbf{IV}} &= (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\ &= (\mathbf{X}'(\mathbf{I} - \mathbf{M}_{\mathbf{Z}})\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{M}_{\mathbf{Z}})\mathbf{y} \\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}\end{aligned} \tag{8-11}$$

because $\mathbf{I} - \mathbf{M}_{\mathbf{Z}}$ is idempotent. Thus, when (*and only when*) $\hat{\mathbf{X}}$ is the set of instruments, the IV estimator is computed by least squares regression of $\mathbf{y}$ on $\hat{\mathbf{X}}$. This conclusion suggests that $\mathbf{b}_{\mathrm{IV}}$ can be computed in two steps, first by computing $\hat{\mathbf{X}}$, then by the least squares regression. For this reason, this is called the two-stage least squares (2SLS) estimator. One should be careful of this approach, however, in the computation of the asymptotic covariance matrix; $\hat{\sigma}^2$ should not be based on $\hat{\mathbf{X}}$. The estimator

$$s^2_{\mathrm{IV}} = \frac{\sum_{i=1}^{n}(y_i - \hat{\mathbf{x}}_i'\mathbf{b}_{\mathrm{IV}})^2}{n}$$

is inconsistent for σ^2, with or without a correction for degrees of freedom. (The appropriate calculation is built into modern software.)

An obvious question is where one is likely to find a suitable set of instrumental variables. The recent literature on natural experiments focuses on local policy changes such as the Mariel Boatlift (Example 6.9) or global policy changes that apply to the entire economy such as mandatory schooling (Example 6.13), or natural outcomes such as occurrences of streams (Example 8.3) or birthdays [Angrist and Krueger (1992)]. In many time-series settings, lagged values of the variables in the model provide natural candidates. In other cases, the answer is less than obvious and sometimes involves some creativity as in Examples 8.9 and 8.11. Unfortunately, there usually is not much choice in the selection of instrumental variables. The choice of $\mathbf{Z}$ is often ad hoc.

Example 8.5 Instrumental Variable Estimation of a Labor Supply Equation

Cornwell and Rupert (1988) analyzed the returns to schooling in a panel data set of 595 observations on heads of households. The sample data are drawn from years 1976 to 1982 from the "Non-Survey of Economic Opportunity" from the Panel Study of Income Dynamics. The estimating equation is

$$\begin{aligned}\ln Wage_{it} = {} & \alpha_1 + \alpha_2 Exp_{it} + \alpha_3 Exp_{it}^2 + \alpha_4 Wks_{it} + \alpha_5 Occ_{it} + \alpha_6 Ind_{it} + \alpha_7 South_{it} + \\ & \alpha_8 SMSA_{it} + \alpha_9 MS_{it} + \alpha_{10} Union_{it} + \alpha_{11} Ed_i + \alpha_{12} Fem_i + \alpha_{13} Blk_i + \varepsilon_{it}.\end{aligned}$$

(The variables are described in Example 4.6.) The main interest of the study, beyond comparing various estimation methods, is α_{11}, the return to education. The equation suggested is a **reduced form equation**; it contains all the variables in the model but does not specify the underlying structural relationships. In contrast, the three-equation model specified at the beginning of this section is a **structural equation system**. The reduced form for this model would consist of separate regressions of *Price* and *Quantity* on (1, *Income, InputPrice, Rainfall*). We will return to the idea of reduced forms in the setting of simultaneous equations models in Chapter 10. For the present, the implication for the suggested model is that this

[3]See Brundy and Jorgenson (1971) and Wooldridge (2010, pp. 103–104).

market equilibrium equation represents the outcome of the interplay of supply and demand in a labor market. Arguably, the supply side of this market might consist of a household labor supply equation such as

$$Wks_{it} = \beta_1 + \beta_2 \ln Wage_{it} + \beta_3 Ed_i + \beta_4 Union_{it} + \beta_5 Fem_i + \varepsilon_{it}.$$

(One might prefer a different set of right-hand-side variables in this structural equation.) Structural equations are more difficult to specify than reduced forms. If the number of weeks worked and the accepted wage offer are determined jointly, then $lnWage_{it}$ and u_{it} in this equation are correlated. We consider two instrumental variable estimators based on

$$\mathbf{z}_1 = [1, Ind_{it}, Ed_i, Union_{it}, Fem_i]$$

and

$$\mathbf{z}_2 = [1, Ind_{it}, Ed_i, Union_{it}, Fem_i, SMSA_{it}].$$

We begin by examining the relevance condition. In the regression of ln *Wage* on $\mathbf{z}_1$, the *t* ratio on *Ind* is +6.02. In the regression of ln *Wage* on $\mathbf{z}_2$, the Wald statistic for the joint test that the coefficients on *Ind* and *SMSA* are both zero is +240.932. In both cases, the hypothesis is rejected, and we conclude that the instruments are, indeed, relevant. Table 8.1 presents the three sets of estimates. The least squares estimates are computed using the standard results in Chapters 3 and 4. One noteworthy result is the very small coefficient on the log wage variable. The second set of results is the instrumental variable estimates. Note that, here, the single instrument is IND_{it}. As might be expected, the log wage coefficient becomes considerably larger. The other coefficients are, perhaps, contradictory. One might have different expectations about all three coefficients. The third set of coefficients are the two-stage least squares estimates based on the larger set of instrumental variables. In this case, *SMSA* and *Ind* are both used as instrumental variables.

8.4.2 A CONTROL FUNCTION APPROACH

A control function is a constructed variable that is added to a model to "control for" the correlation between an endogenous variable and the unobservable elements. In the presence of the control function, the endogenous variable becomes exogenous. Control functions appear in the estimators for several of the nonlinear models we will consider later in the book. For the linear model we are studying here, the approach provides a

TABLE 8.1 Estimated Labor Supply Equation

	OLS		*IV with* $\mathbf{Z}_1$		*IV with* $\mathbf{Z}_2$		*Control Function*	
Variable	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***
Constant	44.7665	1.2153	18.8987	13.0590	30.7044	4.9997	30.7044	4.9100
ln *Wage*	0.7326	0.1972	5.1828	2.2454	3.1518	0.8572	3.1518	0.8418
Education	−0.1532	0.03206	−0.4600	0.1578	−0.3200	0.0661	−0.3200	0.0649
Union	−1.9960	0.1701	−2.3602	0.2567	−2.1940	0.1860	−2.1940	0.1826
Female	−1.3498	0.2642	0.6957	1.0650	−0.2378	0.4679	−0.2378	0.4594
$\hat{u}$							−2.5594	0.8659
$\hat{\sigma}$[a]	1.0301		5.3195		5.1110		5.0187	

[a]Square root of sum of squared residuals/*n*.

useful view of the IV estimator. For the model underlying the preceding example, we have a structural equation,

$$Wks_{it} = \beta_1 + \beta_2 \ln Wage_{it} + \beta_3 Ed_i + \beta_4 Union_{it} + \beta_5 Fem_i + \varepsilon_{it},$$

and the projection (based on $\mathbf{z}_2$),

$$\ln Wage = \gamma_1 + \gamma_2 Ind_{it} + \gamma_3 Ed_i + \gamma_4 Union_{it} + \gamma_5 Fem_i + \gamma_6 SMSA_{it} + u_{it}.$$

The ultimate source of the endogeneity of ln *Wage* in the structural equation for *Wks* is the correlation of the unobservable variables, u and ε. If u were observable—we'll call this observed counterpart $\hat{u}$—then the parameters in the augmented equation,

$$Wks_{it} = \beta_1 + \beta_2 \ln Wage_{it} + \beta_3 Ed_i + \beta_4 Union_{it} + \beta_5 Fem_i + \rho\hat{u} + \widetilde{\varepsilon}_{it},$$

could be estimated consistently by least squares. In the presence of $\hat{u}$, ln *Wage* is uncorrelated with the unobservable in this equation—$\hat{u}$ would be the control function that we seek.

To formalize the approach, write the main equation as

$$y = \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda + \varepsilon, \tag{8-12}$$

where x_2 is the endogenous variable, so $E[x_2\varepsilon] \neq 0$. The instruments, including $\mathbf{x}_1$, are in $\mathbf{z}$. The projection of x_2 on $\mathbf{z}$ is

$$x_2 = \mathbf{z}'\boldsymbol{\pi} + u, \tag{8-13}$$

with $E[\mathbf{z}u] = 0$. We can also form the projection of ε on u,

$$\varepsilon = \rho u + w, \tag{8-14}$$

where $\rho = \sigma_{uw}/\sigma_w^2$. By construction, u and w are uncorrelated. Finally, insert (8-14) in (8-12) so that

$$y = \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda + \rho u + w. \tag{8-15}$$

This is the control function form we had earlier. The loose end, as before, is that in order to proceed, we must observe u. We cannot observe u directly, but we can estimate it using (8-13), the "reduced form" equation for x_2—this is the equation we used to check the relevance of the instrument(s) earlier. We can estimate u as the residual in (8-13), then in the second step, estimate $(\boldsymbol{\beta}, \lambda, \rho)$ by simple least squares. The estimating equation is

$$y = \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda + \rho(x_2 - \mathbf{z}'\mathbf{p}) + \widetilde{w}. \tag{8-16}$$

(The constructed disturbance $\widetilde{w}$ contains both w and the estimation error, $\mathbf{z}'\mathbf{p} - \mathbf{z}'\boldsymbol{\pi}$.) The estimated residual is a control function. The control function estimates with estimated standard errors for the model in Example 8.5 are shown in the two rightmost columns in Table 8.1.

This approach would not seem to provide much economy over 2SLS. It still requires two steps (essentially the same two steps). Surprisingly, as you can see in Table 8.1, it is actually identical to 2SLS, at least for the coefficients. (The proof of this result is pursued in the exercises.) The standard errors, however, are different. The general outcome is that control function estimators, because they contain constructed variables, require an

adjustment of the standard errors. (We will examine several applications, notably Heckman's sample selection model in Chapter 19.) Correction of the standard errors associated with control function estimators often requires elaborate post-estimation calculations (though some of them are built-in procedures in modern software).[4] The calculation for 2SLS, however, is surprisingly simple. The difference between the CF standard errors and the appropriate 2SLS standard errors is a simple scaling.[5] The 2SLS difference is the estimator of σ. Because the coefficients on $\mathbf{x}$ are identical to 2SLS, the sum of squared residuals for the CF estimator is smaller than that for the 2SLS estimator. (See Theorem 3.5.) The values are shown in the last row of Table 8.1. It follows that the only correction needed is to rescale the CF covariance matrix by $(\hat{\sigma}_{CF}/\hat{\sigma}_{2SLS})^2 = (5.1110/5.0187)^2$.

8.4.3 LIMITED INFORMATION MAXIMUM LIKELIHOOD[6]

We have considered estimation of the two equation model,

$$Wks_{it} = \beta_1 + \beta_2 \ln Wage_{it} + \beta_3 Ed_i + \beta_4 Union_{it} + \beta_5 Fem_i + \varepsilon_{it},$$
$$\ln Wage_{it} = \gamma_1 + \gamma_2 Ind_{it} + \gamma_3 Ed_i + \gamma_3 Union_{it} + \gamma_4 Fem_i + \gamma_5 SMSA_{it} + u_i,$$

using 2SLS. In generic form, the equations are

$$y = \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda + \varepsilon,$$
$$x_2 = \mathbf{z}'\boldsymbol{\gamma} + u.$$

The control function estimator is always identical to 2SLS. They use exactly the same information contained in the moments and the two conditions, relevance and exogeneity. If we add to this system an assumption that (ε, u) have a bivariate normal density, then we can construct another estimator, the limited information maximum likelihood estimator. The estimator is formed from the joint density of the two variables, $(y, x_2|\mathbf{x}_1, \mathbf{z})$. We can write this as $f(\varepsilon, u|\mathbf{x}_1, \mathbf{z})\text{abs}|\mathbf{J}|$ where $\mathbf{J}$ is the Jacobian of the transformation from (ε, u) to (y, x_2),[7] $\text{abs}|\mathbf{J}| = 1$, $\varepsilon = (y - \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda)$, and $u = (x_2 - \mathbf{z}'\boldsymbol{\gamma})$. The joint normal distribution with correlation ρ can be written $f(\varepsilon, u|\mathbf{x}_1, \mathbf{z}) = f(\varepsilon|u, \mathbf{x}_1, \mathbf{z})f(u|\mathbf{x}_1, \mathbf{z})$, where $u \sim N[0, \sigma_u^2]$ and $\varepsilon|u \sim N[(\rho\sigma_\varepsilon/\sigma_u)u, (1 - \rho^2)\sigma_\varepsilon^2]$. (See Appendix B.9.) For convenience, write the second of these as $N[\tau u, \sigma_w^2]$. Then, the log of the joint density for an observation in the sample will be

$$\ln f_i = \ln f(\varepsilon_i|u_i) + \ln f(u_i) = -(1/2)\ln \sigma_w^2 - (1/2)\{[y_i - \mathbf{x}_1'\boldsymbol{\beta} - x_{2i}\lambda - \tau(x_{2i} - \mathbf{z}_i'\gamma)]/\sigma_w\}^2 \quad \textbf{(8-17)}$$
$$- (1/2)\ln \sigma_u^2 - (1/2)\{[x_{2i} - \mathbf{z}_i'\gamma]/\sigma_u\}^2.$$

[4] See, for example, Wooldridge (2010, Appendix 6A and Chapter 12).

[5] You can see this in the results. The ratio of any two of the IV standard errors is the same as the ratio for the CF standard errors. For example, for *ED* and *Union*, $0.0661/0.1860 = 0.0649/0.1826$.

[6] Maximum likelihood estimation is developed in detail in Chapter 14. The term *Limited Information* refers to the focus on only one structural equation in what might be a larger system of equations, such as those considered in Section 10.4.

[7] $\mathbf{J} = \begin{bmatrix} \partial\varepsilon/\partial y & \partial\varepsilon/\partial x_2 \\ \partial u/\partial y & \partial u/\partial x_2 \end{bmatrix} = \begin{bmatrix} 1 & \lambda \\ 0 & 1 \end{bmatrix}$, so $\text{abs}|\mathbf{J}| = 1$.

TABLE 8.2 Estimated Labor Supply Equation

	2SLS		*LIML*	
Variable	***Estimated Parameter***	***Standard Error***[a]	***Estimated Parameter***	***Standard Error***[a]
Constant	30.7044	8.25041	30.6392	5.05118
ln *Wage*	3.15182	1.41058	3.16303	0.87325
Education	−0.31997	0.11453	−0.32074	0.06755
Union	−2.19398	0.30507	−2.19490	0.19697
Female	−0.23784	0.79781	−0.23269	0.46572
σ_w	5.01870[b]		5.01865	0.03339
Constant			5.71303	0.03316
Ind			0.08364	0.01284
Education			0.06560	0.00232
Union			0.05853	0.01448
Female			−0.46930	0.02158
SMSA			0.18225	0.01289
σ_u			0.38408	0.00384
τ			−2.57121	0.90334

[a] Standard errors are clustered at the individual level using (8-8c).
[b] Based on mean squared residual.

The log likelihood to be maximized is $\Sigma_i \ln f_i$.[8] Table 8.2 compares the 2SLS and LIML estimates for the model of Example 8.5 using instruments $\mathbf{z}_2$. The LIML estimates are only slightly different from the 2SLS results, but have substantially smaller standard errors. We can view this as the payoff to the narrower specification, that is, the additional normality assumption (though one should be careful about drawing a conclusion about the efficiency of an estimator based on one set of results). There is yet another approach to estimation. The LIML estimator could be computed in two steps, by computing the estimates of $\boldsymbol{\gamma}$ and σ_u first (by least squares estimation of the second equation), then maximizing the log likelihood over $(\boldsymbol{\beta}, \lambda, \tau, \sigma_w)$. This would be identical to the control function estimator—$(\boldsymbol{\beta}, \lambda, \tau)$ would be estimated by regressing y on $(\mathbf{x}_1, x_2, \hat{u})$, then σ_w would be estimated using the residuals. (Note that this would not estimate σ_ε. That would be done by using only the coefficients on $\mathbf{x}_1$ and x_2 to compute the residuals.)

8.5 ENDOGENOUS DUMMY VARIABLES: ESTIMATING TREATMENT EFFECTS

The leading recent application of models of sample selection and endogeneity is the evaluation of "treatment effects." The central focus is on analysis of the effect of participation in a treatment, C, on an outcome variable, y—examples include job training

[8] The parameter estimates would be computed by minimizing (8-17)using one of the methods described in Appendix E. If the equation is overidentified, the least variance ratio estimator described in Section 10.4.4 is an alternative estimation approach. The two approaches will produce the same results.

programs[9] and education.[10] Imbens and Wooldridge (2009, pp. 22–23) cite a number of labor market applications. Recent, more narrow, examples include Munkin and Trivedi's (2007) analysis of the effect of dental insurance and Jones and Rice's (2011) survey that notes a variety of techniques and applications in health economics. A simple starting point, useful for framing ideas, is the linear regression model with a "treatment dummy variable,"

$$y = \mathbf{x}'\boldsymbol{\beta} + \delta C + \varepsilon.$$

The analysis turns on whether it is possible to estimate the "treatment effect" (here, δ), and under what assumptions is δ a meaningful quantity that we are interested in measuring.

Empirical measurement of treatment effects, such as the impact of going to college or participating in a job training or agricultural extension program, presents a large variety of econometric complications. The natural, ultimate objective of an analysis of a treatment or intervention would be *the effect of treatment on the treated.* For example, what is the effect of a college education on the lifetime income of someone who goes to college? Measuring this effect econometrically encounters at least two compelling complications:

Endogeneity of the treatment: The analyst risks attributing to the treatment causal effects that should be attributed to factors that motivate both the treatment and the outcome. In our example, the individual who goes to college might well have succeeded (more) in life than his or her counterpart who did not go to college even if the individual did not attend college. Example 6.8 suggests another case in which some of the students who take the SAT a second time in hopes of improving their scores also take a test preparation course ($C = 1$),

$$\Delta SAT = (SAT_1 - SAT_0) = \mathbf{x}'\boldsymbol{\beta} + \delta C + \varepsilon.$$

The complication here would be whether it is appropriate to attach a causal interpretation to δ.

Missing counterfactual: The preceding thought experiment is not actually the effect we wish to measure. In order to measure the impact of college attendance on lifetime earnings in a pure sense, we would have to run an individual's lifetime twice, once with college attendance and once without (and with all other conditions as they were). Any individual is observed in only one of the two states, so the pure measurement is impossible. The SAT example has the same nature − the experiment can only be run once, either with $C = 1$ or with $C = 0$.

Accommodating these two problems forms the focal point of this enormous and still growing literature. Rubin's causal model (1974, 1978) provides a useful framework for the analysis. Every individual in a population has a potential outcome, y, and can be exposed to the treatment, C. We will denote by C the binary indicator of whether or not the individual receives the treatment. Thus, the potential outcomes are $y|(C = 1) = y_1$ and $y|(C = 0) = y_0$. We can combine these in

$$y = Cy_1 + (1 - C)y_0 = y_0 + C(y_1 - y_0).$$

[9]See LaLonde (1986), Business Week (2009), Example 8.6.

[10]For example, test scores, Angrist and Lavy (1999), Van der Klaauw (2002).

The *average treatment effect*, averaged across the entire population, is

$$ATE = E[y_1 - y_0].$$

The compelling complication is that the individual will exist in only one of the two states, so it is not possible to estimate ATE without further assumptions. More specifically, what the researcher would prefer to see is the average treatment effect on the treated,

$$ATET = E[y_1 - y_0 | C = 1],$$

and note that the second term is now the missing counterfactual.[11]

One of the major themes of the recent research is to devise robust methods of estimation that do not rely heavily on fragile assumptions such as identification by functional form (e.g., relying on bivariate normality) and identification by exclusion restrictions (e.g., relying on basic instrumental variable estimators). This is a challenging exercise—we will rely heavily on these assumptions in much of the rest of this book. For purposes of the general specification, we will denote by $\mathbf{x}$ the exogenous information that will be brought to bear on this estimation problem. The vector $\mathbf{x}$ may (usually will) be a set of variables that will appear in a regression model, but it is useful to think more generally than that and consider $\mathbf{x}$ rather to be an information set. Certain minimal assumptions are necessary to make any headway at all. The following appear at different points in the analysis.

Conditional independence: Receiving the treatment, C, does not depend on the outcome variable once the effect of $\mathbf{x}$ on the outcome is accounted for. In particular, $(y_0, y_1)|\mathbf{x}$ is independent of C. Completely random assignment to the treatment would certainly imply this. If assignment is completely random, then we could omit the effect of $\mathbf{x}$ in this assumption. A narrower case would be assignment based completely on observable criteria ($\mathbf{x}$), which would be "selection on observables" (as opposed to "selection on unobservables which is the foundation of models of "sample selection"). This assumption is extended for regression approaches with the **conditional mean independence assumption**: $E[y_0|\mathbf{x}, C] = E[y_0|\mathbf{x}]$ and $E[y_1|\mathbf{x}, C] = E[y_1|\mathbf{x}]$. This states that the outcome in the untreated state does not affect the participation. The assumption is also labeled *ignorability of the treatment*. As its name implies (and as is clear from the definitions), under ignorability, $ATE = ATET$.

Distribution of potential outcomes: The model that is used for the outcomes is the same for treated and nontreated, $f(y|\mathbf{x}, C = 1) = f(y|\mathbf{x}, C = 0)$. In a regression context, this would mean that the same regression applies in both states and that the disturbance is uncorrelated with T, or that T is exogenous. This is a very strong assumption that we will relax later.

[11]Imbens and Angrist (1994) define a still narrower margin, the "local average treatment effect," or *LATE*. *LATE* is defined with respect to a specific binary instrumental variable. Unlike *ATET*, the *LATE* is defined for a subpopulation related to the instrumental variable and differs with the definition of the instrument. Broadly, the *LATE* narrows the relevant subpopulation to those induced to participate by the variation of the instrument. This specification extends the function of the IV to make it part of the specification of the model to the extent that the object of estimation (*LATE*) is defined by the IV, not independently of it, as in the usual case.

Stable unit treatment value assumption (SUTVA): The treatment of individual i does not affect the outcome of any other individual, j. Without this assumption, which observations are subject to treatment becomes ambiguous. Pure random sampling of observations in a data set would be sufficient for statistical purposes.

Overlap assumption: For any value of $\mathbf{x}$, $0 < \text{Prob}(C = 1|\mathbf{x}) < 1$. The strict inequality in this assumption means that for any $\mathbf{x}$, the population will contain a mix of treated and nontreated individuals. The usefulness of the overlap assumption is that with it, we can expect to find, for any treated individual, an individual who looks like the treated individual, but is not treated. This assumption will be useful for regression approaches.

The following sections will describe three major tools used in the analysis of treatment effects: instrumental variable regression, regression analysis with control functions, and propensity score matching. A fourth, regression discontinuity design, was discussed in Section 6.4.2. As noted, this is a huge and rapidly growing literature. For example, Imbens and Wooldridge's (2009) survey paper runs to 85 pages and includes nearly 300 references, most of them since 2000 (likewise, Wooldridge (2010, Chapter 21)). Our purpose here is to provide some of the vocabulary and a superficial introduction to methods. The survey papers by Imbens and Wooldridge (2009) and Jones and Rice (2010) provide greater detail. The conference volume by Millment, Smith, and Vytlacil (2008) contains many theoretical contributions and empirical applications.[12] A *Journal of Business and Economic Statistics* symposium [Angrist (2001)] raised many of the important questions on whether and how it is possible to measure treatment effects.

Example 8.6 German Labor Market Interventions

"Germany long had the highest ratio of unfilled jobs to unemployed people in Europe. Then, in 2003, Berlin launched the so-called Hartz reforms, ending generous unemployment benefits that went on indefinitely. Now payouts for most recipients drop sharply after a year, spurring people to look for work. From 12.7% in 2005, unemployment fell to 7.1% last November. Even now, after a year of recession, Germany's jobless rate has risen to just 8.6%.

At the same time, lawmakers introduced various programs intended to make it easier for people to learn new skills. One initiative instructed the Federal Labor Agency, which had traditionally pushed the long-term unemployed into government-funded make-work positions, to cooperate more closely with private employers to create jobs. That program last year paid Dutch staffing agency Randstad to teach 15,000 Germans information technology, business English, and other skills. And at a Daimler truck factory in Wörth, 55 miles west of Stuttgart, several dozen short-term employees at risk of being laid off got government help to continue working for the company as mechanic trainees.

Under a second initiative, Berlin pays part of the wages of workers hired from the ranks of the jobless. Such payments make employers more willing to take on the costs of training new workers. That extra training, in turn, helps those workers keep their jobs after the aid expires, a study by the government-funded Institute for Employment Research found. Café Nenninger in the city of Kassel, for instance, used the program to train an unemployed single mother. Co-owner Verena Nenninger says she was willing to take a chance on her in part because the government picked up about a third of her salary the first year. 'It was very helpful, because you never know what's going to happen,' Nenninger says." [*Business Week* (2009)]

[12] In the initial essay in the volume, Goldberger (2008) reproduces Goldberger (1972), in which the author explores the endogeneity issue in detail with specific reference to the Head Start program of the 1960s.

Example 8.7 Treatment Effects on Earnings

LaLonde (1986) analyzed the results of a labor market experiment, The National Supported Work Demonstration, in which a group of disadvantaged workers lacking basic job skills were given work experience and counseling in a sheltered environment. Qualified applicants were assigned to training positions randomly. The treatment group received the benefits of the program. Those in the control group "were left to fend for themselves."[13] The training period was 1976–1977; the outcome of interest for the sample examined here was post-training 1978 earnings. We will attempt to replicate some of the received results based on these data in Example 8.10.

Example 8.8 The Oregon Health Insurance Experiment

The Oregon Health Insurance Experiment is a landmark study of the effect of expanding public health insurance on health care use, health outcomes, financial strain, and well-being of low-income adults. It uses an innovative randomized controlled design to evaluate the impact of Medicaid in the United States. Although randomized controlled trials are the gold standard in medical and scientific studies, they are rarely possible in social policy research. In 2008, the state of Oregon drew names by lottery for its Medicaid program for low-income, uninsured adults, generating just such an opportunity. This ongoing analysis represents a collaborative effort between researchers and the state of Oregon to learn about the costs and benefits of expanding public health insurance. (www.nber.org/oregon/) (Further details appear in Chapter 6.)

Example 8.9 The Effect of Counseling on Financial Management

Smith, Hochberg, and Greene (2014) examined the impact of a financial management skills program on later credit outcomes such as credit scores, debt, and delinquencies of a sample of home purchasers. From the abstract of the study:

> *. . . . [D]evelopments in mortgage products and drastic changes in the housing market have made the realization of becoming a homeowner more challenging. Fortunately, homeownership counseling is available to help navigate prospective homebuyers in their quest. But the effectiveness of such counseling over time continues to be contemplated. Previous studies have made important strides in our understanding of the value of homeownership counseling, but more work is needed. More specifically, homeownership education and counseling have never been rigorously evaluated through a randomized field experiment.*
>
> *This study is based on a long-term (five-year) effort undertaken by the Federal Reserve Bank of Philadelphia on the effectiveness of pre-purchase homeownership and financial management skills counseling. . . . [T]he study employs an experimental design, with study participants randomly assigned to a control or a treatment group. Participants completed a baseline survey and were tracked for four years after receiving initial assistance by means of an annual survey, which also tracks participants' life changes over time. To assist in the analysis, additional information was obtained annually to track changes in the participants' creditworthiness. The study considers the influence of counseling on credit scores, total debt, and delinquencies in payments.*

8.5.1 REGRESSION ANALYSIS OF TREATMENT EFFECTS

An earnings equation that purports to account for the value of a college education is

$$\ln \text{Earnings}_i = \mathbf{x}_i'\boldsymbol{\beta} + \delta C_i + \varepsilon_i,$$

[13]The demonstration was run in numerous cities in the mid-1970s. See LaLonde (1986, pp. 605–609) for details on the NSW experiments.

where C_i is a dummy variable indicating whether or not the individual attended college. The same format has been used in any number of other analyses of programs, experiments, and treatments. The question is: Does δ measure the value of a college education (assuming that the rest of the regression model is correctly specified)? The answer is no if the typical individual who chooses to go to college would have relatively high earnings whether or not he or she went to college. The problem is one of self-selection. If our observation is correct, then least squares estimates of δ will actually overestimate the treatment effect—it will likely pick up the college effect as well as effects explainable by the other latent factors (that are not in $\mathbf{x}$). The same observation applies to estimates of the treatment effects in other settings in which the individuals themselves decide whether or not they will receive the treatment.

8.5.2 INSTRUMENTAL VARIABLES

The starting point to the formulation of the earnings equation would be the familiar RCM,

$$y = \mu_0 + C(\mu_1 - \mu_0) + \varepsilon_0 + C(\varepsilon_1 - \varepsilon_0),$$

where $\mu_j = E[y_j]$. Suppose, first, that $\varepsilon_1 = \varepsilon_0$, so the final term falls out of the equation. [Though the assumption is unmotivated, we note that no sample will contain direct observations on $(\varepsilon_1 - \varepsilon_0)$—no individual will be in both states—so the assumption is a reasonable normalization.] There is no presumption at this point that ε_j is uncorrelated with $\mathbf{x}$. Suppose, as well, that there exist instrumental variables, $\mathbf{z}$, that contain at least one variable that is not in $\mathbf{x}$, such that the linear projection of ε_0 on $\mathbf{x}$ and $\mathbf{z}$, $Proj(\varepsilon_0|\mathbf{x}, \mathbf{z})$, equals $Proj(\varepsilon_0|\mathbf{x})$. That is, $\mathbf{z}$ is *exogenous*. (See Section 4.4.5 and (4-34) for definition of the linear projection. It will be convenient to assume that $\mathbf{x}$ and $\mathbf{z}$ have no variables in common.) The linear projection is $Proj(\varepsilon_0|\mathbf{x}) = \gamma_0 + \mathbf{x}'\boldsymbol{\gamma}$. Then,

$$y = (\mu_0 + \gamma_0) + \delta C + \mathbf{x}'\boldsymbol{\gamma} + w_0,$$

where $w_0 = \varepsilon_0 - (\gamma_0 + \mathbf{x}'\boldsymbol{\gamma})$. By construction, w_0 and $\mathbf{x}$ are uncorrelated. There is also no assumption that C is uncorrelated with w_0 since we have assumed that C is correlated with ε_0 at the outset. The setup would seem now to lend itself to a familiar IV approach. However, we have yet to certify $\mathbf{z}$ as a proper instrument. We assumed $\mathbf{z}$ is exogenous. We assume it is *relevant*, still using the projections, with $Proj(C|\mathbf{x}, \mathbf{z}) \neq Proj(C|\mathbf{x})$. This would be the counterpart to the relevance condition in Assumption 1 in Section 8.2. The model is, then,

$$y = \lambda_0 + \delta C + \mathbf{x}'\boldsymbol{\gamma} + w_0.$$

The parameters of this model can, in principle, be estimated by 2SLS. In the notation of Section 6.3, $\mathbf{X}_i = [1, C_i, \mathbf{x}_i']$ and $\mathbf{Z}_i = [1, \mathbf{z}_i', \mathbf{x}_i']$. Consistency and asymptotic normality of the 2SLS estimator are based on the usual results. See Theorem 8.1. Because we have not assumed anything about $\text{Var}[w_0|\mathbf{x}]$, efficiency is unclear. Consistency is the objective, however, and inference can be based on heteroscedasticity robust estimators of the asymptotic covariance matrix of the 2SLS estimator, as in (8-8h) or (8-8c).

The relevance assumption holds that in the projection of C on $\mathbf{x}$ and $\mathbf{z}$,

$$C = \gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z + w_c = \mathbf{f}'\boldsymbol{\gamma}_c + w_c,$$

$\boldsymbol{\gamma}_z$ is not zero. Strictly, the projection works. However, because C is a binary variable, w_c equals either $-\mathbf{f}'\boldsymbol{\gamma}_c$ or $1 - \mathbf{f}'\boldsymbol{\gamma}_c$, so the lack of correlation between w_c and $\mathbf{f}$ (specifically $\mathbf{z}$) is a result of the construction of the linear projection, not necessarily a characteristic of the underlying design of the real-world counterpart to the variables in the model (though one would expect $\mathbf{z}$ to have been chosen with this in mind). One might observe that the understanding of the functioning of the instrument is that its variation makes participation more (or less) likely. As such, the relevance of the instrument is to the probability of participation. A more convincing specification that is consistent with this observation, albeit one less general, can replace the relevance assumption with a formal parametric specification of the conditional probability that C equals 1, $Prob(C = 1|\mathbf{x}, \mathbf{z}) = F(\mathbf{x}, \mathbf{z}: \boldsymbol{\theta}) \neq Prob(C = 1|\mathbf{x})$. We also replace projections with expected values in the exogeneity assumption; $Proj(\varepsilon_0|\mathbf{x}, \mathbf{z}) = Proj(\varepsilon_0|\mathbf{x})$ will now be $E(\varepsilon_0|\mathbf{x}, \mathbf{z}) = Proj(\varepsilon_0|\mathbf{x}) = (\gamma_0 + \mathbf{x}'\boldsymbol{\gamma})$. This suggests an instrument of the form $F(\mathbf{x}, \mathbf{z}: \boldsymbol{\theta}) = Prob(C = 1|\mathbf{x}, \mathbf{z})$, a known function—the usual choice would be a probit model (see Section 17.2)—$\Phi(\theta_0 + \mathbf{x}'\boldsymbol{\theta}_x + \mathbf{z}'\boldsymbol{\theta}_z)$ where $\Phi(t)$ is the standard normal CDF. To reiterate, the conditional probability is correlated with $C|\mathbf{x}$ but not correlated with $w_0|\mathbf{x}$. With this additional assumption, a natural instrument in the form of $\hat{F}(\mathbf{x}, \mathbf{z}: \boldsymbol{\theta}) = \Phi(\hat{\theta}_0 + \mathbf{x}'\hat{\boldsymbol{\theta}}_x + \mathbf{z}'\hat{\boldsymbol{\theta}}_z)$ (estimated by maximum likelihood) can be used. The advantages of this approach are internally consistent specification of the treatment dummy variable and some gain in efficiency of the estimator that follows from the narrower assumptions.

This approach creates an additional issue that is not present in the previous linear approach. The approach suggested here would succeed even if there were no variables in $\mathbf{z}$. The IV estimator is $(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$ where the rows of $\mathbf{Z}$ and $\mathbf{X}$ are $[1, \hat{\Phi}, \mathbf{x}']$ and $[1, C, \mathbf{x}']$. As long as $\hat{\Phi}$ is not a linear function of $\mathbf{x}$ (and is both relevant and exogenous), then the parameters will be identified by this IV estimator. Because $\hat{\Phi}$ is nonlinear, it could meet these requirements even without any variables in $\mathbf{z}$. The parameters in this instance are identified by the nonlinear functional form of the probability model. Typically, the probability is at least reasonably highly (linearly) correlated with the variables in the model, so possibly severe problems of multicollinearity are likely to appear. But, more to the point, the entire logic of the instrumental variable approach is based on an exogenous source of variation that is correlated with the endogenous variable and not with the disturbance. The nonlinear terms in the probability model do not persuasively pass that test. Thus, the typical application does, indeed, ensure that there are excluded (from the main equation) variables in $\mathbf{z}$.[14]

Finally, note that because $\hat{F}(\mathbf{x}, \mathbf{z}: \boldsymbol{\theta})$ is not a linear function of $\mathbf{x}$ and $\mathbf{z}$, this IV estimator is not two-stage least squares. That is, $\mathbf{y}$ is not regressed on $(1, \hat{\Phi}, \mathbf{x})$ to estimate $\lambda_0, \delta, \boldsymbol{\gamma}$. Rather, the estimator is in (8-6), $(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$. Because no assumption has been made about the disturbance variance, the robust covariance matrix estimator in (8-8h) should be used.

[14]As an example, Scott, Schurer, Jensen, and Sivey (2009) state, "Although the model is formally identified by its nonlinear functional form, as long as the full rank condition of the data matrix is ensured (Heckman, 1978; Wilde, 2000), we introduce exclusion restrictions to aid identification of the causal parameter . . . The row vector I_{ij} captures the variables included in the PIP participation Equation (5) but excluded from the outcome Equation (4)." ("The Effects of an Incentive Program on Quality of Care in Diabetes Management," *Health Economics*, 19, 2009, pp. 1091–1108, Section 4.2.)

8.5.3 A CONTROL FUNCTION ESTIMATOR

The list of assumptions and implications that produced the second IV estimator above was:

Rubin Causal Model	$y = Cy_1 + (1 - C)y_0$ $= \mu_0 + C(\mu_1 - \mu_0) + \varepsilon_0 + C(\varepsilon_1 - \varepsilon_0),$
Nonignorability of the Treatment	$\text{Cov}(C, \varepsilon_0) \neq 0,$
Normalization	$\varepsilon_1 - \varepsilon_0 = 0,$
Exogeneity and Linearity	$Proj(\varepsilon_0 \mid \mathbf{x}, \mathbf{z}) = E[\varepsilon_0 \mid \mathbf{x}, \mathbf{z}] = \gamma_0 + \mathbf{x}'\boldsymbol{\gamma},$ no assumption is made about $\text{Var}[\varepsilon_0 \mid \mathbf{x}]$,
Relevance of the Instruments	$Prob(C = 1 \mid \mathbf{x}, \mathbf{z}) = F(\mathbf{x}, \mathbf{z}: \boldsymbol{\theta}) \neq Prob(C = 1 \mid \mathbf{x}),$
Reduced Form	$y = \lambda_0 + \delta C + \mathbf{x}'\boldsymbol{\gamma} + w_0,\ Cov(\mathbf{x}, w_0) = 0$ is implied,
Endogenous Treatment Dummy Variable	$Cov(C, w_0) \neq 0,$
Probit Model for Prob(C = 1|x, z)	$C* = \gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z + w_c, w_c \sim N[0, 1^2],$ $C = 1$ if $C* > 0$ and $C = 0$ if $C* \leq 0,$ $Prob(C = 1 \mid \mathbf{x}, \mathbf{z}) = \Phi(\theta_0 + \mathbf{x}'\boldsymbol{\theta}_x + \mathbf{z}'\boldsymbol{\theta}_z).$

The source of the endogeneity of the treatment dummy variable is now more explicit. Because neither $\mathbf{x}$ nor $\mathbf{z}$ is correlated with w_0, the source is the correlation of w_c and w_0. As in all such cases, the ultimate source of the endogeneity is the covariation among the unobservables in the model.

The foregoing is sufficient to produce a consistent instrumental variable estimator. We now pursue whether with the same data and assumptions, there is a regression-based estimator. Based on the assumptions, we find that

$$E[y \mid C = 1, \mathbf{x}, \mathbf{z}] = \lambda_0 + \delta + \mathbf{x}'\boldsymbol{\gamma} + E[w_0 \mid C = 1, \mathbf{x}, \mathbf{z}],$$
$$E[y \mid C = 0, \mathbf{x}, \mathbf{z}] = \lambda_0 + \mathbf{x}'\boldsymbol{\gamma} + E[w_0 \mid C = 0, \mathbf{x}, \mathbf{z}].$$

Because we have not specified the last term, the model is incomplete. Suppose the model is fully parameterized with (w_0, w_c) bivariate normally distributed with means 0, variances σ^2 and 1 and covariance $\rho\sigma$. Under these assumptions, the functional form of the conditional mean is known,

$$\begin{aligned} E[y \mid C = 1, \mathbf{x}, \mathbf{z}] &= \lambda_0 + \mathbf{x}'\boldsymbol{\gamma} + \delta + E[w_0 \mid C = 1, \mathbf{x}, \mathbf{z}] \\ &= \lambda_0 + \mathbf{x}'\boldsymbol{\gamma} + \delta + E[w_0 \mid w_c > (-\gamma_0 - \mathbf{x}'\boldsymbol{\gamma}_x - \mathbf{z}'\boldsymbol{\gamma}_z)] \\ &= \lambda_0 + \mathbf{x}'\boldsymbol{\gamma} + \delta + \rho\sigma\left[\frac{\phi(\gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z)}{\Phi(\gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z)}\right]. \end{aligned}$$

The counterpart for $C = 0$ would be

$$E[y \mid C = 0, \mathbf{x}, \mathbf{z}] = \lambda_0 + \mathbf{x}'\boldsymbol{\gamma} + \rho\sigma\left[\frac{-\phi(\gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z)}{[1 - \Phi(\gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z)]}\right].$$

By using the symmetry of the normal distribution, $\phi(t) = \phi(-t)$ and $\Phi(t) = 1 - \Phi(-t)$, we can combine these into a single regression,

$$\begin{aligned} E[y \mid C\, \mathbf{x}_i, \mathbf{z}_i] &= \lambda_0 + \mathbf{x}'\boldsymbol{\gamma} + \delta C + \rho\sigma\left[\frac{(2C - 1)\phi[(2C - 1)(\gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z)]}{\Phi[(2C - 1)(\gamma_0 + \mathbf{x}'\boldsymbol{\gamma}_x + \mathbf{z}'\boldsymbol{\gamma}_z)]}\right] \\ &= \lambda_0 + \mathbf{x}'\boldsymbol{\gamma} + \delta C + \tau G(C, \mathbf{x}, \mathbf{z}: \boldsymbol{\theta}). \end{aligned}$$

(See Theorem 19.5.) The result is a feature of the bivariate normal distribution. There are two approaches that could be taken. The conditional mean function is a nonlinear regression that can be estimated by nonlinear least squares. The bivariate normality assumption carries an implicit assumption of homoscedasticity, so there is no need for a heteroscedasticity robust estimator for the covariance matrix. Nonlinear least squares might be quite cumbersome. A simpler, two-step "control function" approach would be to fit the probit model as before, then compute the bracketed term and add it as an additional term. The estimating equation is

$$y = \lambda_0 + \delta C + \mathbf{x}'\boldsymbol{\gamma} + \boldsymbol{\tau}\hat{G} + h,$$

where $h = y - E[y|C, \mathbf{x}, \mathbf{z}]$. This can be estimated by linear least squares. As with other control function estimators, the asymptotic covariance matrix for the estimator must be adjusted for the constructed regressor. [See Heckman (1979) for results related to this model.] The result of Murphy and Topel (2002) can be used to obtain the correction. Bootstrapping can be used as well. [This turns out to be identical to Heckman's (1979) "sample selection" model developed in Section 19.5.2. A covariance matrix for the two-step estimator as well as a full information maximum likelihood estimator are developed there.]

The precision and compactness of this result has been purchased by adding the bivariate normality assumption. It has also been made much simpler with the still unmotivated assumption, $\varepsilon_1 - \varepsilon_0 = 0$. A distributional assumption can be substituted for the normalization. Wooldridge (2010, pp. 945–948) assumes that $[w_c, (\varepsilon_1 - \varepsilon_0)]$ are bivariate normally distributed, and obtains another control function estimator, again based on properties of the bivariate normal distribution.

8.5.4 PROPENSITY SCORE MATCHING

If the treatment assignment is completely ignorable, then, as noted, estimation of the treatment effects is greatly simplified. Suppose, as well, that there are observable variables that influence both the outcome and the treatment assignment. Suppose it is possible to obtain pairs of individuals matched by a common $\mathbf{x}_i$, one with $C_i = 0$, the other with $C_i = 1$. If done with a sufficient number of pairs so as to average over the population of $\mathbf{x}_i$s, then a *matching estimator*, the average value of $(y_i|C_i = 1) - (y_i|C_i = 0)$, would estimate $E[y_1 - y_0]$, which is what we seek. Of course, it is optimistic to hope to find a large sample of such matched pairs, both because the sample overall is finite and because there may be many regressors, and the "cells" in the distribution of $\mathbf{x}_i$ are likely to be thinly populated. This will be worse when the regressors are continuous, for example, with a family income variable. Rosenbaum and Rubin (1983) and others[15] suggested, instead, matching on the propensity score, $F(\mathbf{x}_i) = \text{Prob}(C_i = 1|\mathbf{x}_i)$. Individuals with similar propensity scores are paired and the average treatment effect is then estimated by the differences in outcomes. Various strategies are suggested by the authors for obtaining the necessary subsamples and for verifying the conditions under which the procedures will be valid.[16] We will examine and try to replicate a well-known application in Example 8.10.

[15]Other important references in this literature are Becker and Ichino (1999), Dehejia and Wahba (1999), LaLonde (1986), Heckman, Ichimura, and Todd (1997, 1998), Robins and Rotnitzky (1995), Heckman, Ichimura, Smith, and Todd (1998), Heckman, LaLonde, and Smith (1999), Heckman, Tobias, and Vytlacil (2003), Hirano, Imbens, and Ridder (2003), and Heckman and Vytlacil (2000).

[16]See, for example, Becker and Ichino (2002).

Example 8.10 Treatment Effects on Earnings

LaLonde (1986) analyzed the results of a labor market experiment, The National Supported Work Demonstration, in which a group of disadvantaged workers lacking basic job skills were given work experience and counseling in a sheltered environment. Qualified applicants were assigned to training positions randomly. The treatment group received the benefits of the program. Those in the control group "were left to fend for themselves." The training period was 1976–1977; the outcome of interest for the sample examined here was posttraining 1978 earnings.

LaLonde reports a large variety of estimates of the treatment effect, for different subgroups and using different estimation methods. Nonparametric estimates for the group in our sample are roughly \$900 for the income increment in the posttraining year. (See LaLonde, p. 609.) Similar results are reported from a two-step regression-based estimator similar to the control function estimator in Section 8.5.3. (See LaLonde's footnote to Table 6, p. 616.)

LaLonde's data are fairly well traveled, having been used in replications and extensions in, for example, Dehejia and Wahba (1999), Becker and Ichino (2002), Stata (2006), Dehejia (2005), Smith and Todd (2005), and Wooldridge (2010). We have reestimated the matching estimates reported in Becker and Ichino along with several side computations including the estimators developed in Sections 8.5.2 and 8.5.3. The data in the file used there (and here) contain 2,490 control observations and 185 treatment observations on the following variables:

$$
\begin{aligned}
t &= \text{treatment dummy variable,}\\
age &= \text{age in years,}\\
educ &= \text{education in years,}\\
marr &= \text{dummy variable for married,}\\
black &= \text{dummy variable for black,}\\
hisp &= \text{dummy variable for Hispanic,}\\
nodegree &= \text{dummy for no degree (not used),}\\
re74 &= \text{real earnings in 1974,}\\
re75 &= \text{real earnings in 1975,}\\
re78 &= \text{real earnings in 1978.}
\end{aligned}
$$

Transformed variables added to the equation are

$$
\begin{aligned}
age^2 &= age \text{ squared,}\\
educ^2 &= educ \text{ squared,}\\
re74^2 &= re74 \text{ squared,}\\
re75^2 &= re75 \text{ squared,}\\
blacku\ 74 &= black \text{ times } 1(re74 = 0).
\end{aligned}
$$

We also scaled all earnings variables by 10,000 before beginning the analysis. (See Appendix Table F19.3. The data are downloaded from the Website http://users.nber.org/~rdehejia/nswdata2.html. The two specific subsamples are in http://www.nber.org/~rdehejia//nsw_control.txt, and http://www.nber.org/~rdehejia/nsw_treated.txt.) (We note that Becker and Ichino report they were unable to replicate Dehejia and Wahba's results, although they could come reasonably close. We, in turn, were not able to replicate either set of results, though we, likewise, obtained quite similar results. See Table 8.3.)

To begin, Figure 8.2 describes the re78 data for the treatment group in the upper panel and the controls in the lower. Any regression- (or sample means–) based analysis of the differences of the two distributions will reflect the fact that the mean of the controls is far larger than that of the treatment group. The re74 and re75 data appear similar, so estimators that account

for the observable past values should be able to isolate the difference attributable to the treatment, if there is a difference.

Table 8.3 lists the results obtained with the regression-based methods and matching based on the propensity scores. The specification for the regression-based approaches is

TABLE 8.3 Estimates of Average Treatment Effect on the Treated

Simple difference in means: $\overline{re78}_1 - \overline{re78}_0 = 6{,}349 - 21{,}553 = -15{,}204$[a]

Estimator	δ	*Standard Error (Method)*
Regression Based		
Simple OLS	859[a]	765[a] (Robust Standard Error)
2SLS	2,021	1,690 (Robust Standard Error)
IV Using predicted probabilities	2,145	1,131 (Robust Standard Error)
2 Step Control Function	2,273	1,012 (100 Bootstrap Replications)
		1,249 (Heckman Two Step)
Propensity Score Matching[c]		
Matching	1,571	669 (25 Bootstrap Replications)
Becker and Ichino	1,537[b]	1,016[b] (100 Bootstrap Replications)

[a] See Wooldridge (2010, p. 929, Table 21.1).
[b] See Becker and Ichino (2002, p. 374) based on Kernel Matching and common support. Number of controls = 1,157 (1,155 here).
[c] Becker and Ichino employed the `pscore` and `attk` routines in *Stata*. Results here used `LOGIT` and `PSMATCH` in *NLOGIT6*.

TABLE 8.4 Empirical Distribution of Propensity Scores

Percent	*Lower*	*Upper*
0–5	0.000591	0.000783
5–10	0.000787	0.001061
10–15	0.001065	0.001377
15–20	0.001378	0.001748
20–25	0.001760	0.002321
25–30	0.002340	0.002956
30–35	0.002974	0.004057
35–40	0.004059	0.005272
40–45	0.005278	0.007486
45–50	0.007557	0.010451
50–55	0.010563	0.014643
55–60	0.014686	0.022462
60–65	0.022621	0.035060
65–70	0.035075	0.051415
70–75	0.051415	0.076188
75–80	0.076376	0.134189
80–95	0.134238	0.320638
85–90	0.321233	0.616002
90–95	0.624407	0.949418
95–100	0.949418	0.974835

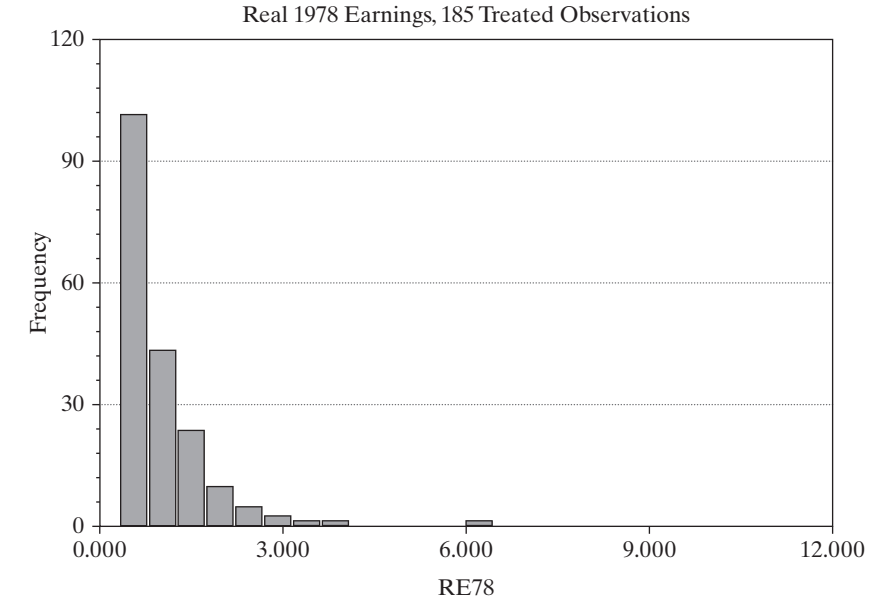
Sample size = 1,347
Average score = 0.137238
Std. Dev score = 0.274079

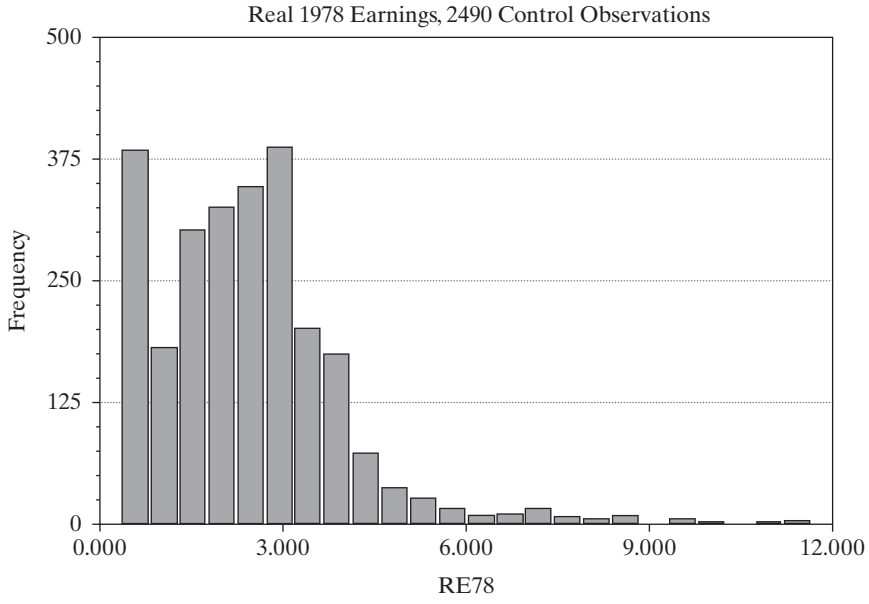
	Lower	Upper	# Obs
1	0.000591	0.098016	1041
2	0.098016	0.195440	63
3	0.195440	0.390289	65
4	0.390289	0.585138	36
5	0.585138	0.779986	32
6	0.779986	0.877411	17
7	0.877411	0.926123	7
8	0.926123	0.974835	86

$$re78 = \lambda_0 + y_1 age + y_2 educ + y_3 black + y_4 hisp + y_5 marr + y_6 re74 + y_7 re75 + \delta T + w_0.$$

The additional variables in **z** are (age^2, $educ^2$, $re74^2$, $re75^2$, *blacku*74). [Note, for consistency with Becker and Ichino, *nodegree* was not used. The specification of **x** in the regression equation follows Wooldridge (2010).] As anticipated, the simple difference in means is

FIGURE 8.2 Real 1978 Earnings, Treated Versus Controls.





uninformative. The regression-based estimates are quite consistent; the estimate of ATT is roughly $2,100. The propensity score method focuses only on the observable differences in the observations (including, crucially, *re*74 and *re*75) and produces an estimate of about $1,550.

The propensity score matching analysis proceeded as follows: A logit model in which the included variables were a *constant, age, age*^{2}, *education, education*^{2}, *marr, black, hisp, re74, re75, re742, re752*, and *blacku74* was computed for the treatment assignment. The fitted probabilities are used for the propensity scores. By means of an iterative search, the range of propensity scores was partitioned into eight regions within which, by a simple *F* test, the mean scores of the treatments and controls were not statistically different. The partitioning is shown in Table 8.4. The 1,347 observations are all the treated observations and the 1,162 control observations are those whose propensity scores fell within the range of the scores for the treated observations.

Within each interval, each treated observation is paired with a small number of the nearest control observations. We found the average difference between treated observation and control to equal $1,574.35. Becker and Ichino reported $1,537.94.

8.6 HYPOTHESIS TESTS

There are several tests to be carried out in this model.

8.6.1 TESTING RESTRICTIONS

For testing linear restrictions in H_0: $\boldsymbol{R\beta} = \boldsymbol{q}$, *the Wald statistic based on whatever form of* Asy.Var$[\mathbf{b}_{IV}]$ *has been computed will be the usual choice. The test statistic, based on the unrestricted estimator, will be*

$$\chi^2[J] = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}\,\text{Est.Asy.Var}(\hat{\boldsymbol{\beta}})\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}). \tag{8-18}$$

For testing the simple hypothesis that a coefficient equals zero, this is the square of the usual *t* ratio that is always reported with the estimated coefficient. The *t* ratio, itself, can be used instead, though the implication is that the large sample critical value, 1.96 for 95%, for example, would be used rather than the *t* distribution.

For the 2SLS estimator based on least squares regression of $\mathbf{y}$ on $\hat{\mathbf{X}}$ an asymptotic *F* statistic can be computed as follows:

$$F[J, n - K] = \frac{\left\{\sum_{i=1}^{n}(y_i - \hat{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}_{\text{Restricted}})^2 - \sum_{i=1}^{n}(y_i - \hat{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}_{\text{Unrestricted}})^2\right\}/J}{\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{Unrestricted}})^2/(n - K)}. \tag{8-19}$$

[See Wooldridge (2010, p. 105).] As in the regression model [see (5-14) and (5-15)], an approximation to the *F* statistic will be the chi-squared statistic, *JF*. Unlike the earlier case, however, *J* times the statistic in (8-19) is not equal to the result in (8-18) even if the denominator is rescaled by $(n-K)/n$. They are different approximations. The *F* statistic is computed using both resticted and unresticted estimators.

A third approach to testing the hypothesis of the restrictions can be based on the Lagrange multiplier principle. The moment equation for the 2SLS estimator is

$$\bar{\mathbf{g}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{x}}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{x}}_i\hat{\varepsilon}_i = \mathbf{0}.$$

(Note that the residuals are computed using the original $\mathbf{x}$, not the prediction.) The mean vector $\bar{\mathbf{g}}$ will equal $\mathbf{0}$ when it is computed using $\hat{\boldsymbol{\beta}}_{\text{Unrestricted}}$ to compute the residuals. It will generally not equal zero if $\hat{\boldsymbol{\beta}}_{\text{Restricted}}$ is used instead. We consider using a Wald test to test the hypothesis that $\text{E}[\bar{\mathbf{g}}] = 0$. The asymptotic variance of $\bar{\mathbf{g}}$ will be estimated using $1/n$ times the matrix in (8-8), (8-8h) or (8-8c), whichever is appropriate. The Wald statistic will be

$$\chi^2[J] = \left[\sum_{i=1}^{n} \hat{\mathbf{x}}_i (y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{\text{Restricted}}) \right]' \left(\frac{1}{n} \text{Est.Asy.Var} \left[\hat{\boldsymbol{\beta}}_{\text{Restricted}} \right] \right)^{-1} \left[\sum_{i=1}^{n} \hat{\mathbf{x}}_i (y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{\text{Restricted}}) \right].$$

A convenient way to carry out this test is the approximation $\chi^2[J] = nR^2$ where the R^2 is the uncentered R^2 in the least squares regression of $\hat{\boldsymbol{\varepsilon}}$ on $\hat{\mathbf{X}}$.

8.6.2 SPECIFICATION TESTS

There are two aspects of the model that we would be interested in verifying if possible, rather than assuming them at the outset. First, it will emerge in the derivation in Section 8.4.1 that of the two estimators considered here, least squares and instrumental variables, the first is unambiguously more efficient (i.e., has a smaller variance around its mean). The IV estimator is robust; it is consistent whether or not $\text{plim}(\mathbf{X}'\boldsymbol{\varepsilon}/n) = \mathbf{0}$. However, if IV is not needed, that is, if $\boldsymbol{\gamma} = \mathbf{0}$, then least squares would be a better estimator by virtue of its smaller variance.[17] For this reason, and possibly in the interest of a test of the theoretical specification of the model, a test that reveals information about the bias of least squares will be useful. Second, the use of two-stage least squares with $L > K$, that is, with "additional" instruments, entails $L - K$ restrictions on the relationships among the variables in the model. As might be apparent from the derivation thus far, when there are K variables in $\mathbf{X}$, some of which may be endogenous, then there must be at least K variables in $\mathbf{Z}$ in order to identify the parameters of the model, that is, to obtain consistent estimators of the parameters using the information in the sample. When there is an excess of instruments, one is actually imposing additional, arguably superfluous restrictions on the process generating the data. Consider, once again, the agricultural market example at the end of Section 8.3.4. In that structure, it is certainly safe to assume that *Rainfall* is an exogenous event that is uncorrelated with the disturbances in the demand equation. But, it is conceivable that the interplay of the markets involved might be such that the *InputPrice* is correlated with the shocks in the demand equation. In the market for biofuels, corn is both an input in the market supply and an output in other markets. In treating *InputPrice* as exogenous in that example, we would be imposing the assumption that *InputPrice* is uncorrelated with ε_D, at least by some measure unnecessarily because the parameters of the demand equation can be estimated without this assumption. This section will describe two specification tests that consider these aspects of the IV estimator.

[17] It is possible that even if least squares is inconsistent, it might still be more precise. If LS is only slightly biased but has a much smaller variance than IV, then by the expected squared error criterion, variance plus squared bias, least squares might still prove the preferred estimator. This turns out to be nearly impossible to verify empirically.

8.6.3 TESTING FOR ENDOGENEITY: THE HAUSMAN AND WU SPECIFICATION TESTS

If the regressors in the model are not correlated with the disturbances and are not measured with error, then there would be some benefit to using the least squares (LS) estimator rather than the IV estimator. Consider a comparison of the two covariance matrices *under the hypothesis that both estimators are consistent, that is, assuming* $\text{plim}\,(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$ and assuming A.4 (Section 8.2). The difference between the asymptotic covariance matrices of the two estimators is

$$\begin{aligned}\text{Asy.Var}[\mathbf{b}_{\text{IV}}] - \text{Asy.Var}[\mathbf{b}_{\text{LS}}] &= \frac{\sigma^2}{n}\text{plim}\left(\frac{\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}}{n}\right)^{-1} - \frac{\sigma^2}{n}\text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\\ &= \frac{\sigma^2}{n}\text{plim}\, n[(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]\\ &= \frac{\sigma^2}{n}\text{plim}\, n\{[\mathbf{X}'(\mathbf{I} - \mathbf{M}_{\mathbf{Z}})\mathbf{X}]^{-1} - [\mathbf{X}'\mathbf{X}]^{-1}\}\\ &= \frac{\sigma^2}{n}\text{plim}\, n\{[\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}]^{-1} - [\mathbf{X}'\mathbf{X}]^{-1}\}.\end{aligned} \tag{8-20}$$

The matrix in braces is nonnegative definite, which establishes that least squares is more efficient than IV. Our interest in the difference between these two estimators goes beyond the question of efficiency. The null hypothesis of interest will be specifically whether $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$. Seeking the covariance between $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ through $(1/n)\mathbf{X}'\mathbf{e}$ is fruitless, of course, because $(1/n)\mathbf{X}'\mathbf{e} = \mathbf{0}$. In a seminal paper, Hausman (1978) developed an alternative testing strategy. The logic of Hausman's approach is as follows. Under the null hypothesis, we have two consistent estimators of $\boldsymbol{\beta}$, $\mathbf{b}_{\text{LS}}$ and $\mathbf{b}_{\text{IV}}$. Under the alternative hypothesis, only one of these, $\mathbf{b}_{\text{IV}}$, is consistent. The suggestion, then, is to examine $\mathbf{d} = \mathbf{b}_{\text{IV}} - \mathbf{b}_{\text{LS}}$. Under the null hypothesis, plim $\mathbf{d} = \mathbf{0}$, whereas under the alternative, plim $\mathbf{d} \neq \mathbf{0}$. We will test this hypothesis with a Wald statistic,

$$\begin{aligned}H &= \mathbf{d}'\{\text{Est.Asy.Var}[\mathbf{d}]\}^{-1}\mathbf{d}\\ &= (\mathbf{b}_{\text{IV}} - \mathbf{b}_{\text{LS}})'\{\text{Est.Asy.Var}[\mathbf{b}_{\text{IV}}] - \text{Est.Asy.Var}[\mathbf{b}_{\text{LS}}]\}^{-1}(\mathbf{b}_{\text{IV}} - \mathbf{b}_{\text{LS}})\\ &= (\mathbf{b}_{\text{IV}} - \mathbf{b}_{\text{LS}})'\hat{\mathbf{H}}^{-1}(\mathbf{b}_{\text{IV}} - \mathbf{b}_{\text{LS}}),\end{aligned}$$

where $\hat{\mathbf{H}}^{-1}$ is the estimator of the covariance matrix in (8-20). Under the null hypothesis, we have two different, but consistent, estimators of σ^2. If we use s^2 as the common estimator, then the statistic will be

$$H = \frac{\mathbf{d}'[(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{d}}{s^2}.$$

It is tempting to invoke our results for the full rank quadratic form in a normal vector and conclude the degrees of freedom for this chi-squared statistic is K. However, the rank of $[(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]$ is only $K^* = K - K_0$, where K_0 is the number of exogenous variables in $\mathbf{X}$ (and the ordinary inverse will not exist), so K^* is the degrees of freedom for the test. The Wald test requires a generalized inverse [see Hausman and Taylor (1981)], so it is going to be a bit cumbersome. An alternative **variable addition test** approach devised by Wu (1973) and Durbin (1954) is simpler. An F or Wald statistic with

K^* and $n - K - K^*$ degrees of freedom can be used to test the joint significance of the elements of $\boldsymbol{\gamma}$ in the augmented regression,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \hat{\mathbf{X}}^*\boldsymbol{\gamma} + \boldsymbol{\varepsilon}^*, \tag{8-21}$$

where $\hat{\mathbf{X}}^*$ are the fitted values in regressions of the variables in $\mathbf{X}^*$ on $\mathbf{Z}$. This result is equivalent to the Hausman test for this model.[18]

Example 8.5 Labor Supply Model (Continued)

For the labor supply equation estimated in Example 8.5, we used the Wu (variable addition) test to examine the endogeneity of the ln *Wage* variable. For the first step, ln $Wage_{it}$ is regressed on $\mathbf{z}_{1,it}$. The predicted value from this equation is then added to the least squares regression of Wks_{it} on $\mathbf{x}_{it}$. The results of this regression are

$$\begin{aligned} Wks_{it} = {} & \underset{(12.3284)}{18.8987} + \underset{(0.1980)}{0.6938}\ \ln Wage_{it} - \underset{(0.1490)}{0.4600}\ Ed_i - \underset{(0.2423)}{2.3602} Union_{it} \\ & + \underset{(1.0054)}{0.6958}\ Fem_i + \underset{(2.1290)}{4.4891}\ fitted \ln Wage_{it} + u_{it}, \end{aligned}$$

where the estimated standard errors are in parentheses. The *t* ratio on the fitted log wage coefficient is 2.108, which is larger than the critical value from the standard normal table of 1.96. Therefore, the hypothesis of exogeneity of the log *Wage* variable is rejected. If $\mathbf{z}_{2,it}$ is used instead, the *t* ratio on the predicted value is 2.96, which produces the same conclusion.

The control function estimator based on (8-16),

$$y = \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda + \rho(x_2 - \mathbf{z}'\mathbf{p}) + \tilde{w},$$

resembles the estimating equation in (8-21). It is actually equivalent. If the residual in (8-16) is replaced by the prediction, $\mathbf{z}'\mathbf{p}$, the identical least squares results are obtained save for the coefficient on the residual, which changes sign. The results in the preceding example would thus be identical save for the sign of the coefficient on the prediction of ln *Wage*, which would be negative. The implication (as happens in many applications) is that the control function estimator provides a simple constructive test for endogeneity that is the same as the Hausman–Wu test. A test of the significance of the coefficient on the control function is equivalent to the Hausman test.

8.6.4 A TEST FOR OVERIDENTIFICATION

The motivation for choosing the IV estimator is not efficiency. The estimator is constructed to be consistent; efficiency is a secondary consideration. In Chapter 13, we will revisit the issue of efficient method of moments estimation. The observation that 2SLS represents the most efficient use of all L instruments establishes only the efficiency of the estimator in the class of estimators that use K linear combinations of the columns of $\mathbf{Z}$. The IV estimator is developed around the **orthogonality conditions**,

$$E[\mathbf{z}_i\varepsilon_i] = \mathbf{0}. \tag{8-22}$$

The sample counterpart to this is the **moment equation**,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i = \mathbf{0}. \tag{8-23}$$

[18]Algebraic derivations of this result can be found in the articles and in Davidson and MacKinnon (2004, Section 8.7).

The solution, when $L = K$, is $\mathbf{b}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$, as we have seen. If $L > K$, then there is no single solution, and we arrived at 2SLS as a strategy. Estimation is still based on (8-23). However, the sample counterpart is now L equations in K unknowns and (8-23) has no solution. Nonetheless, under the hypothesis of the model, (8-22) remains true. We can consider the additional restrictions as a hypothesis that might or might not be supported by the sample evidence. The excess of moment equations provides a way to test the overidentification of the model. The test will be based on (8-23), which, when evaluated at $\mathbf{b}_{\mathrm{IV}}$, will not equal zero when $L > K$, though the hypothesis in (8-22) might still be true.

The test statistic will be a Wald statistic. (See Section 5.4.) The sample statistic, based on (8-23) and the IV estimator, is

$$\overline{\mathbf{m}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i e_{\mathrm{IV},i} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(y_i - \mathbf{x}_i'\mathbf{b}_{\mathrm{IV}}).$$

The Wald statistic is

$$\chi^2[L - K] = \overline{\mathbf{m}}'[\mathrm{Var}(\overline{\mathbf{m}})]^{-1}\overline{\mathbf{m}}.$$

To complete the construction, we require an estimator of the variance. There are two ways to proceed. Under the assumption of the model,

$$\mathrm{Var}[\overline{\mathbf{m}}] = \frac{\sigma^2}{n^2}\mathbf{Z}'\mathbf{Z},$$

which can be estimated easily using the sample estimator of σ^2. Alternatively, we might base the estimator on (8-22), which would imply that an appropriate estimator would be

$$\mathit{Est}.\mathrm{Var}[\overline{\mathbf{m}}] = \frac{1}{n^2}\sum_{i=1}(\mathbf{z}_i e_{\mathrm{IV},i})(\mathbf{z}_i e_{\mathrm{IV},i})' = \frac{1}{n^2}\sum_{i=1}e_{\mathrm{IV},i}^2\mathbf{z}_i\mathbf{z}_i'.$$

These two estimators will be numerically different in a finite sample, but under the assumptions that we have made so far, both (multiplied by n) will converge to the same matrix, so the choice is immaterial. Current practice favors the second. The Wald statistic is, then,

$$LM = n\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i e_{\mathrm{IV},i}\right)'\left[\frac{1}{n}\sum_{i=1}^{n}e_{\mathrm{IV},i}^2\mathbf{z}_i\mathbf{z}_i'\right]^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i e_{\mathrm{IV},i}\right).$$

A remaining detail is the number of degrees of freedom. The test can only detect the failure of $L - K$ moment equations, so that is the rank of the quadratic form; the limiting distribution of the statistic is chi squared with $L - K$ degrees of freedom. If the equation is exactly identified, then $(1/n)\mathbf{Z}'\mathbf{e}_{\mathrm{IV}}$ will be exactly zero. As we saw in testing linear restrictions in Section 8.5.1, there is a convenient way to compute the LM statistic. The chi-squared statistic can be computed as n times the uncentered R^2 in the linear regression of $\mathbf{e}_{\mathrm{IV}}$ on $\mathbf{Z}$ that would be

$$LM = \frac{\mathbf{e}_{IV}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{e}_{IV}}{\mathbf{e}_{IV}'\mathbf{e}_{IV}}.$$

Example 8.11 Overidentification of the Labor Supply Equation

In Example 8.5, we computed 2SLS estimates of the parameters of an equation for weeks worked. The estimator is based on

$$\mathbf{x} = [1, \ln \textit{Wage}, \textit{Education}, \textit{Union}, \textit{Female}]$$

and

$$\mathbf{z} = [1, \text{Ind}, \text{Education}, \text{Union}, \text{Female}, \text{SMSA}].$$

There is one overidentifying restriction. The sample moment based on the 2SLS results in Table 8.1 is

$$(1/4165)\mathbf{Z}'\mathbf{e}_{2\,\text{SLS}} = [0, .03476, 0, 0, 0, -.01543]'.$$

The chi-squared statistic is 1.09399 with one degree of freedom. If the first suggested variance estimator is used, the statistic is 1.05241. Both are well under the 95 percent critical value of 3.84, so the hypothesis of overidentification is not rejected. Table 8.5 displays the 2SLS estimates based on the two instruments separately and the estimates based on both.

We note a final implication of the test. One might conclude, based on the underlying theory of the model, that the overidentification test relates to one particular instrumental variable and not another. For example, in our market equilibrium example with two instruments for the demand equation, *Rainfall* and *InputPrice*, rainfall is obviously exogenous, so a rejection of the overidentification restriction would eliminate *InputPrice* as a valid instrument. However, this conclusion would be inappropriate; the test suggests only that one or more of the elements in (8-22) are nonzero. It does not suggest which elements in particular these are.

8.7 WEAK INSTRUMENTS AND LIML

Our analysis thus far has focused on the "identification" condition for IV estimation, that is, the "exogeneity assumption," A.I9, which produces

$$\text{plim}\,(1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0}. \tag{8-24}$$

Taking the "relevance" assumption,

$$\text{plim}\,(1/n)\mathbf{Z}'\mathbf{X} = \mathbf{Q}_{\mathbf{ZX}}, \text{ a finite, nonzero, } L \times K \text{ matrix with rank } K, \tag{8-25}$$

TABLE 8.5 2SLS Estimates of the Labor Supply Equation

	IND		*SMSA*		*IND and SMSA*	
Variable	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***
Constant	18.8987	20.26604	33.0018	9.10852	30.7044	8.25041
LWAGE	5.18285	3.47416	2.75658	1.56100	3.15182	1.41058
ED	−0.46000	0.24352	−0.29272	0.12414	−0.31997	0.11453
UNION	−2.36016	0.43069	−2.16164	0.30395	−2.19398	0.30507
FEM	0.69567	1.66754	−0.41950	0.85547	−0.23784	0.79781
$\hat{\sigma}$	5.32268		5.08719		5.11405	

as given produces a consistent IV estimator. In absolute terms, with (8-24) in place, (8-25) is sufficient to assert consistency. As such, researchers have focused on *exogeneity* as the defining problem to be solved in constructing the IV estimator. A growing literature has argued that greater attention needs to be given to the relevance condition. While, strictly speaking, (8-25) is indeed sufficient for the asymptotic results we have claimed, the common case of "weak instruments," in which (8-25) is only barely true has attracted considerable scrutiny. In practical terms, instruments are "weak" when they are only slightly correlated with the right-hand-side variables, $\mathbf{X}$; that is, $(1/n)\mathbf{Z}'\mathbf{X}$ is *close* to zero. Researchers have begun to examine these cases, finding in some an explanation for perverse and contradictory empirical results.[19]

Superficially, the problem of weak instruments shows up in the asymptotic covariance matrix of the IV estimator,

$$\text{Asy.Var}[\mathbf{b}_{\text{IV}}] = \frac{\sigma_\varepsilon^2}{n}\left[\left(\frac{\mathbf{X}'\mathbf{Z}}{n}\right)\left(\frac{\mathbf{Z}'\mathbf{Z}}{n}\right)^{-1}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]^{-1},$$

which will be "large" when the instruments are weak, and, other things equal, larger the weaker they are. However, the problems run deeper than that. Nelson and Startz (1990a,b) and Hahn and Hausman (2003) list two implications: (i) The 2SLS estimator is badly biased toward the ordinary least squares estimator, which is known to be inconsistent, and (ii) the standard first-order asymptotics (such as those we have used in the preceding) will not give an accurate framework for statistical inference. Thus, the problem is worse than simply lack of precision. There is also at least some evidence that the issue goes well beyond "small sample problems."[20]

Current research offers several prescriptions for detecting weakness in instrumental variables. For a single endogenous variable ($\mathbf{x}$ that is correlated with $\boldsymbol{\varepsilon}$), the standard approach is based on the first-step OLS regression of 2SLS. The conventional F statistic for testing the hypothesis that all the coefficients in the regression

$$x_i = \mathbf{z}_i'\boldsymbol{\pi} + u_i$$

are zero is used to test the "hypothesis" that the instruments are weak. An F statistic less than 10 signals the problem.[21] When there are more than one endogenous variables in the model, testing each one separately using this test is not sufficient, because collinearity among the variables could impact the result but would not show up in either test. Shea (1997) proposes a four-step multivariate procedure that can be used. Godfrey (1999) derived a surprisingly simple alternative method of doing the computation. For endogenous variable k, the Godfrey statistic is the ratio of the estimated variances of the two estimators, OLS and 2SLS,

$$R_k^2 = \frac{v_k(OLS)/\mathbf{e}'\mathbf{e}(\text{OLS})}{v_k(2SLS)/\mathbf{e}'\mathbf{e}(\text{2SLS})},$$

[19]Important references are Nelson and Startz (1990a,b), Staiger and Stock (1997), Stock, Wright, and Yogo (2002), Hahn and Hausman (2002, 2003), Kleibergen (2002), Stock and Yogo (2005), and Hausman, Stock, and Yogo (2005).

[20]See Bound, Jaeger, and Baker (1995).

[21]See Nelson and Startz (1990b), Staiger and Stock (1997), and Stock and Watson (2007, Chapter 12) for motivation of this specific test.

where v_k (OLS) is the kth diagonal element of $[\mathbf{e}'\mathbf{e}(\text{OLS})/(n - K)](\mathbf{X}'\mathbf{X})^{-1}$ and v_k (2SLS) is defined likewise. With the scalings, the statistic reduces to

$$R_k^2 = \frac{(\mathbf{X}'\mathbf{X})^{kk}}{(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{kk}},$$

where the superscript indicates the element of the inverse matrix. The F statistic can then be based on this measure, $F = [R_k^2/(L - 1)]/[(1 - R_k^2)/(n - L)]$ assuming that $\mathbf{Z}$ contains a constant term.

It is worth noting that the test for weak instruments is not a specification test, nor is it a constructive test for building the model. Rather, it is a strategy for helping the researcher avoid basing inference on unreliable statistics whose properties are not well represented by the familiar asymptotic results, for example, distributions under assumed null model specifications. Several extensions are of interest. Other statistical procedures are proposed in Hahn and Hausman (2002) and Kleibergen (2002).

The stark results of this section call the IV estimator into question. In a fairly narrow circumstance, an alternative estimator is the "moment"-free LIML estimator discussed in Section 8.4.3. Another, perhaps somewhat unappealing, approach is to retreat to least squares. The OLS estimator is not without virtue. The asymptotic variance of the OLS estimator,

$$\text{Asy.Var}[\mathbf{b}_{\text{LS}}] = (\sigma^2/n)\mathbf{Q}_{\mathbf{XX}}^{-1},$$

is unambiguously smaller than the asymptotic variance of the IV estimator,

$$\text{Asy.Var}[\mathbf{b}_{\text{IV}}] = (\sigma^2/n)(\mathbf{Q}_{\mathbf{XZ}}\mathbf{Q}_{\mathbf{ZZ}}^{-1}\mathbf{Q}_{\mathbf{ZX}})^{-1}.$$

(The proof is considered in the exercises.) Given the preceding results, it could be far smaller. The OLS estimator is inconsistent, however,

$$\text{plim } \mathbf{b}_{\text{LS}} - \boldsymbol{\beta} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\gamma},$$

[see (8-4)]. By a mean squared error comparison, it is unclear whether the OLS estimator with

$$M(\mathbf{b}_{\text{LS}}|\boldsymbol{\beta}) = (\sigma^2/n)\mathbf{Q}_{\mathbf{XX}}^{-1} + \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{Q}_{\mathbf{XX}}^{-1},$$

or the IV estimator, with

$$M(\mathbf{b}_{\text{IV}}|\boldsymbol{\beta}) = (\boldsymbol{\sigma}^2/n)(\mathbf{Q}_{\mathbf{XZ}}\mathbf{Q}_{\mathbf{ZZ}}^{-1}\mathbf{Q}_{\mathbf{ZX}})^{-1},$$

is more precise. The natural recourse in the face of weak instruments is to drop the endogenous variable from the model or improve the instrument set. Each of these is a specification issue. Strictly in terms of estimation strategy within the framework of the data and specification in hand, there is scope for OLS to be the preferred strategy.

8.8 MEASUREMENT ERROR

Thus far, it has been assumed (at least implicitly) that the data used to estimate the parameters of our models are true measurements on their theoretical counterparts. In practice, this situation happens only in the best of circumstances. All sorts of measurement problems creep into the data that must be used in our analyses. Even

carefully constructed survey data do not always conform exactly to the variables the analysts have in mind for their regressions. Aggregate statistics such as GDP are only estimates of their theoretical counterparts, and some variables, such as depreciation, the services of capital, and "the interest rate," do not even exist in an agreed-upon theory. At worst, there may be no physical measure corresponding to the variable in our model; intelligence, education, and permanent income are but a few examples. Nonetheless, they all have appeared in very precisely defined regression models.

8.8.1 LEAST SQUARES ATTENUATION

In this section, we examine some of the received results on regression analysis with badly measured data. The biases introduced by measurement error can be rather severe. There are almost no known finite-sample results for the models of measurement error; nearly all the results that have been developed are asymptotic.[22] The following presentation will use a few simple asymptotic results for the classical regression model.

The simplest case to analyze is that of a regression model with a single regressor and no constant term. Although this case is admittedly unrealistic, it illustrates the essential concepts, and we shall generalize it presently. Assume that the model,

$$y^* = \beta x^* + \varepsilon, \tag{8-26}$$

conforms to all the assumptions of the classical normal regression model. If data on y^* and x^* were available, then β would be estimable by least squares. Suppose, however, that the observed data are only imperfectly measured versions of y^* and x^*. In the context of an example, suppose that y^* is ln(output/labor) and x^* is ln(capital/labor). Neither factor input can be measured with precision, so the observed y and x contain errors of measurement. We assume that

$$y = y^* + v \quad \text{with } v \sim N[0, \sigma_v^2], \tag{8-27a}$$

$$x = x^* + u \quad \text{with } u \sim N[0, \sigma_u^2]. \tag{8-27b}$$

Assume, as well, that u and v are independent of each other and of y^* and x^*. (As we shall see, adding these restrictions is not sufficient to rescue a bad situation.)

As a first step, insert (8-27a) into (8-26), assuming for the moment that only y^* is measured with error,

$$y = \beta x^* + \varepsilon + v = \beta x^* + \varepsilon'.$$

This result still conforms to the assumptions of the classical regression model. As long as the regressor is measured properly, measurement error on the dependent variable can be absorbed in the disturbance of the regression and ignored. To save some cumbersome notation, therefore, we shall henceforth assume that the measurement error problems concern only the independent variables in the model.

Consider, then, the regression of y on the observed x. By substituting (8-27b) into (8-26), we obtain

$$y = \beta x + [\varepsilon - \beta u] = \beta x + w. \tag{8-28}$$

[22] See, for example, Imbens and Hyslop (2001).

Because x equals $x^* + u$, the regressor in (8-28) is correlated with the disturbance,

$$\text{Cov}[x, w] = \text{Cov}[x^* + u, \varepsilon - \beta u] = -\beta\sigma_u^2. \tag{8-29}$$

This result violates one of the central assumptions of the classical model, so we can expect the least squares estimator,

$$b = \frac{(1/n)\sum_{i=1}^{n} x_i y_i}{(1/n)\sum_{i=1}^{n} x_i^2},$$

to be inconsistent. To find the probability limits, insert (8-26) and (8-27b) and use the Slutsky theorem,

$$\text{plim } b = \frac{\text{plim}(1/n)\sum_{i=1}^{n}(x_i^* + u_i)(\beta x_i^* + \varepsilon_i)}{\text{plim}(1/n)\sum_{i=1}^{n}(x_i^* + u_i)^2}.$$

Because x^*, ε, and u are mutually independent, this equation reduces to

$$\text{plim } b = \frac{\beta Q^*}{Q^* + \sigma_u^2} = \frac{\beta}{1 + \sigma_u^2/Q^*}, \tag{8-30}$$

where $Q^* = \text{plim}(1/n)\sum_i x_i^{*2}$. As long as σ_u^2 is positive, b is inconsistent, with a persistent bias toward zero. Clearly, the greater the variability in the measurement error, the worse the bias. The effect of biasing the coefficient toward zero is called **attenuation**.

In a multiple regression model, matters only get worse. Suppose, to begin, we assume that $\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\mathbf{X} = \mathbf{X}^* + \mathbf{U}$, allowing every observation on every variable to be measured with error. The extension of the earlier result is

$$\text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right) = \mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}, \quad \text{and} \quad \text{plim}\left(\frac{\mathbf{X}'\mathbf{y}}{n}\right) = \mathbf{Q}^*\boldsymbol{\beta}.$$

Hence,

$$\text{plim } \mathbf{b} = [\mathbf{Q}^* + \Sigma_{uu}]^{-1}\mathbf{Q}^*\boldsymbol{\beta} = \boldsymbol{\beta} - [\boldsymbol{Q}^* + \boldsymbol{\Sigma}_{uu}]^{-1}\boldsymbol{\Sigma}_{uu}\boldsymbol{\beta}. \tag{8-31}$$

This probability limit is a mixture of all the parameters in the model. In the same fashion as before, bringing in outside information could lead to identification. The amount of information necessary is extremely large, however, and this approach is not particularly promising.

It is common for only a single variable to be measured with error. One might speculate that the problems would be isolated to the single coefficient. Unfortunately, this situation is not the case. For a single bad variable—assume that it is the first—the matrix $\boldsymbol{\Sigma}_{uu}$ is of the form

$$\boldsymbol{\Sigma}_{uu} = \begin{bmatrix} \sigma_u^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

It can be shown that for this special case,

$$\text{plim } b_1 = \frac{\beta_1}{1 + \sigma_u^2 q^{*11}}, \tag{8-32a}$$

[note the similarity of this result to (8-30)], and, for $k \neq 1$,

$$\text{plim } b_k = \beta_k - \beta_1\left[\frac{\sigma_u^2 q^{*k1}}{1 + \sigma_u^2 q^{*11}}\right], \tag{8-32b}$$

where q^{*k1} is the $(k,1)$th element in $(\mathbf{Q}^*)^{-1}$.[23] This result depends on several unknowns and cannot be estimated. The coefficient on the badly measured variable is still biased toward zero. The other coefficients are all biased as well, although in unknown directions. A badly measured variable contaminates all the least squares estimates.[24] If more than one variable is measured with error, there is very little that can be said.[25] Although expressions can be derived for the biases in a few of these cases, they generally depend on numerous parameters whose signs and magnitudes are unknown and, presumably, unknowable.

8.8.2 INSTRUMENTAL VARIABLES ESTIMATION

An alternative set of results for estimation in this model (and numerous others) is built around the method of instrumental variables. Consider once again the errors in variables model in (8-26) and (8-27a,b). The parameters, β, σ_ε^2, q^*, and σ_u^2 are not identified in terms of the moments of x and y. Suppose, however, that there exists a variable z such that z is correlated with x^* but not with u. For example, in surveys of families, income is notoriously badly reported, partly deliberately and partly because respondents often neglect some minor sources. Suppose, however, that one could determine the total amount of checks written by the head(s) of the household. It is quite likely that this z would be highly correlated with income, but perhaps not significantly correlated with the errors of measurement. If $\text{Cov}[x^*, z]$ is not zero, then the parameters of the model become estimable, as

$$\text{plim } \frac{(1/n)\sum_i y_i z_i}{(1/n)\sum_i x_i z_i} = \frac{\beta \text{ Cov}[x^*, z]}{\text{Cov}[x^*, z]} = \beta. \tag{8-33}$$

The special case when the instrumental variable is binary produces a useful result. If z_i is a dummy variable such that $\bar{x}_{|z=1} - \bar{x}_{|z=0}$ is not zero—that is, the instrument is relevant (see Section 8.2), then the estimator in (8-33) is

$$\hat{\beta} = \frac{\bar{y}_{|z=1} - \bar{y}_{|z=0}}{\bar{x}_{|z=1} - \bar{x}_{|z=0}}.$$

A proof of the result is given in Example 8.2.[26] This is called the Wald (1940) estimator.

For the general case, $\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathbf{X} = \mathbf{X}^* + \mathbf{U}$, suppose that there exists a matrix of variables $\mathbf{Z}$ that is not correlated with the disturbances or the measurement error,

[23]Use (A-66) to invert $[\mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}] = [\mathbf{Q}^* + (\sigma_u \mathbf{e}_1)(\sigma_u \mathbf{e}_1)']$, where $\mathbf{e}_1$ is the first column of a $K \times K$ identity matrix. The remaining results are then straightforward.

[24]This point is important to remember when the presence of measurement error is suspected.

[25]Some firm analytic results have been obtained by Levi (1973), Theil (1961), Klepper and Leamer (1983), Garber and Klepper (1980), Griliches (1986), and Cragg (1997).

[26]The proof in Example 8.2 is given for a dependent variable that is also binary. However, the proof is generic, and extends without modification to this case.

but is correlated with regressors, $\mathbf{X}$. Then the instrumental variables estimator, based on $\mathbf{Z}$, $\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$, is consistent and asymptotically normally distributed with asymptotic covariance matrix that is estimated with

$$\text{Est.Asy.Var}[\mathbf{b}_{\text{IV}}] = \hat{\sigma}^2[\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{Z}'\mathbf{Z}][\mathbf{X}'\mathbf{Z}]^{-1}. \quad \textbf{(8-34)}$$

For more general cases, Theorem 8.1 and the results in Section 8.3 apply.

8.8.3 PROXY VARIABLES

In some situations, a variable in a model simply has no observable counterpart. Education, intelligence, ability, and like factors are perhaps the most common examples. In this instance, unless there is some observable indicator for the variable, the model will have to be treated in the framework of missing variables. Usually, however, such an indicator can be obtained; for the factors just given, years of schooling and test scores of various sorts are familiar examples. The usual treatment of such variables is in the measurement error framework. If, for example,

$$\text{income} = \beta_1 + \beta_2 \,\text{education} + \varepsilon,$$

and

$$\text{years of schooling} = \text{education} + u,$$

then the model of Section 8.8.1 applies. The only difference here is that the true variable in the model is "latent." No amount of improvement in reporting or measurement would bring the proxy closer to the variable for which it is proxying.

The preceding is a pessimistic assessment, perhaps more so than necessary. Consider a **structural model**,

$$Earnings = \beta_1 + \beta_2\, Experience + \beta_3\, Industry + \beta_4\, Ability + \varepsilon.$$

Ability is unobserved, but suppose that an indicator, say, IQ, is. If we suppose that IQ is related to *Ability* through a relationship such as

$$IQ = \alpha_1 + \alpha_2\, Ability + v,$$

then we may solve the second equation for *Ability* and insert it in the first to obtain the **reduced form equation**,

$$Earnings = (\beta_1 - \beta_4\alpha_1/\alpha_2) + \beta_2\, Experience + \beta_3\, Industry + (\beta_4/\alpha_2)IQ + (\varepsilon - v\beta_4/\alpha_2).$$

This equation is intrinsically linear and can be estimated by least squares. We do not have consistent estimators of β_1 and β_4, but we do have them for the coefficients of interest, β_2 and β_3. This would appear to solve the problem. We should note the essential ingredients; we require that the **indicator**, IQ, not be related to the other variables in the model, and we also require that v not be correlated with any of the variables. (A perhaps obvious additional requirement is that the proxy not provide information in the regression that would not be provided by the missing variable if it were observed. In the context of the example, this would require that $E[Earnings \,|\, Experience, Industry, Ability, IQ] = E[Earnings \,|\, Experience, Industry, Ability]$.) In this instance, some of the parameters of the structural model are identified in terms of observable data. Note, though, that IQ is not a proxy variable; it is an

indicator of the latent variable, *Ability*. This form of modeling has figured prominently in the education and educational psychology literature. Consider in the preceding small model how one might proceed with not just a single indicator, but say with a battery of test scores, all of which are indicators of the same latent ability variable.

It is to be emphasized that a proxy variable is not an instrument (or the reverse). Thus, in the instrumental variables framework, it is implied that we do not regress $\mathbf{y}$ on $\mathbf{Z}$ to obtain the estimates. To take an extreme example, suppose that the full model was

$$\begin{aligned}\mathbf{y} &= \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon},\\ \mathbf{X} &= \mathbf{X}^* + \mathbf{U},\\ \mathbf{Z} &= \mathbf{X}^* + \mathbf{W}.\end{aligned}$$

That is, we happen to have two badly measured estimates of $\mathbf{X}^*$. The parameters of this model can be estimated without difficulty if $\mathbf{W}$ is uncorrelated with $\mathbf{U}$ and $\mathbf{X}^*$, *but not by regressing* $\mathbf{y}$ *on* $\mathbf{Z}$. The instrumental variables technique is called for.

When the model contains a variable such as education or ability, the question that naturally arises is, If interest centers on the other coefficients in the model, why not just discard the problem variable?[27] This method produces the familiar problem of an omitted variable, compounded by the least squares estimator in the full model being inconsistent anyway. Which estimator is worse? McCallum (1972) and Wickens (1972) show that the asymptotic bias (actually, degree of inconsistency) is worse if the proxy is omitted, even if it is a bad one (has a high proportion of measurement error). This proposition neglects, however, the precision of the estimates. Aigner (1974) analyzed this aspect of the problem and found, as might be expected, that it could go either way. He concluded, however, that "there is evidence to broadly support use of the proxy."

Example 8.12 Income and Education in a Study of Twins

The traditional model used in labor economics to study the effect of education on income is an equation of the form

$$y_i = \beta_1 + \beta_2\,\text{age}_i + \beta_3\,\text{age}_i^2 + \beta_4\,\text{education}_i + \mathbf{x}_i'\boldsymbol{\beta}_5 + \varepsilon_i,$$

where y_i is typically a wage or yearly income (perhaps in log form) and $\mathbf{x}_i$ contains other variables, such as an indicator for sex, region of the country, and industry. The literature contains discussion of many possible problems in estimation of such an equation by least squares using measured data. Two of them are of interest here:

1. Although "education" is the variable that appears in the equation, the data available to researchers usually include only "years of schooling." This variable is a proxy for education, so an equation fit in this form will be tainted by this problem of measurement error. Perhaps surprisingly so, researchers also find that reported data on years of schooling are themselves subject to error, so there is a second source of measurement error. For the present, we will not consider the first (much more difficult) problem.
2. Other variables, such as "ability"—we denote these μ_i—will also affect income and are surely correlated with education. If the earnings equation is estimated in the form shown above, then the estimates will be further biased by the absence of this "omitted variable." For reasons we will explore in Chapter 19, this bias has been called the **selectivity effect** in recent studies.

[27] This discussion applies to the measurement error and latent variable problems equally.

Simple cross-section studies will be considerably hampered by these problems. But, in a study of twins, Ashenfelter and Krueger (1994) analyzed a data set that allowed them, with a few simple assumptions, to ameliorate these problems.[28]

Annual "twins festivals" are held at many places in the United States. The largest is held in Twinsburg, Ohio. The authors interviewed about 500 individuals over the age of 18 at the August 1991 festival. Using pairs of twins as their observations enabled them to modify their model as follows: Let (y_{ij}, A_{ij}) denote the earnings and age for twin j, $j = 1, 2$, for pair i. For the education variable, only self-reported "schooling" data, S_{ij}, are available. The authors approached the measurement problem in the schooling variable, S_{ij}, by asking each twin how much schooling he or she had and how much schooling his or her sibling had. Denote reported schooling *by* sibling *m of* sibling *j* by $S_{ij}\,(m)$. So, the self-reported years of schooling of twin 1 is $S_{i1}\,(1)$. When asked how much schooling twin 1 has, twin 2 reports $S_{i1}\,(2)$. The measurement error model for the schooling variable is

$$S_{ij}\,(m) = S_{ij} + u_{ij}\,(m), \quad j, m = 1, 2, \quad \text{where } S_{ij} = \text{"true' schooling for twin } j \text{ of pair } i.$$

We assume that the two sources of measurement error, $u_{ij}\,(m)$, are uncorrelated and they and S_{ij} have zero means. Now, consider a simple bivariate model such as the one in (8-26),

$$y_{ij} = \beta S_{ij} + \varepsilon_{ij}.$$

As we saw earlier, a least squares estimate of β using the reported data will be attenuated,

$$\text{plim } b = \frac{\beta \times \text{Var}[S_{ij}]}{\text{Var}[S_{ij}] + \text{Var}[u_{ij}(j)]} = \beta q.$$

(Because there is no natural distinction between twin 1 and twin 2, the assumption that the variances of the two measurement errors are equal is innocuous.) The factor q is sometimes called the **reliability ratio**. In this simple model, if the reliability ratio were known, then β could be consistently estimated. In fact, the construction of this model allows just that. Because the two measurement errors are uncorrelated,

$$\begin{aligned}\text{Corr}[S_{i1}\,(1), S_{i1}\,(2)] &= \text{Corr}[S_{i2}\,(1), S_{i2}\,(2)] \\ &= \frac{\text{Var}[S_{i1}]}{\{\{\text{Var}[S_{i1}] + \text{Var}[u_{i1}(1)]\} \times \{\text{Var}[S_{i1}] + \text{Var}[u_{i1}(2)]\}\}^{1/2}} = q.\end{aligned}$$

In words, the correlation between the two reported education attainments measures the reliability ratio. The authors obtained values of 0.920 and 0.877 for 298 pairs of identical twins and 0.869 and 0.951 for 92 pairs of fraternal twins, thus providing a quick assessment of the extent of measurement error in their schooling data.

The earnings equation is a multiple regression, so this result is useful for an overall assessment of the problem, but the numerical values are not sufficient to undo the overall biases in the least squares regression coefficients. An instrumental variables estimator was used for that purpose. The estimating equation for $y_{ij} = \ln \mathit{Wage}_{ij}$ with the least squares (OLS) and instrumental variable (IV) estimates is as follows:

$y_{ij} =$	$\beta_1 +$	β_2 age$_i +$	β_3 age$_i^2 +$	$\beta_4 S_{ij}(j) +$	$\beta_5 S_{im}(m) +$	β_6 sex$_i +$	β_7 race$_i + \varepsilon_{ij}$
LS		(0.088)	(−0.087)	(0.084)		(0.204)	(−0.410)
IV		(0.088)	(−0.087)	(0.116)	(−0.037)	(0.206)	(−0.428).

[28]Other studies of twins and siblings include Bound, Chorkas, Haskel, Hawkes, and Spector (2003). Ashenfelter and Rouse (1998), Ashenfelter and Zimmerman (1997), Behrman and Rosengweig (1999), Isacsson (1999), Miller, Mulvey, and Martin (1995), Rouse (1999), and Taubman (1976).

In the equation, S_{ij} (j) is the person's report of his or her own years of schooling and S_{im} (m) is the sibling's report of the sibling's own years of schooling. The problem variable is schooling. To obtain a consistent estimator, the method of instrumental variables was used, using each sibling's report of the other sibling's years of schooling as a pair of instrumental variables. The estimates reported by the authors are shown below the equation. (The constant term was not reported, and for reasons not given, the second schooling variable was not included in the equation when estimated by LS.) This preliminary set of results is presented to give a comparison to other results in the literature. The age, schooling, and gender effects are comparable with other received results, whereas the effect of race is vastly different, −40%, here compared with a typical value of +9% in other studies. The effect of using the instrumental variable estimator on the estimates of β_4 is of particular interest. Recall that the reliability ratio was estimated at about 0.9, which suggests that the IV estimate would be roughly 11% higher (1/0.9). Because this result is a multiple regression, that estimate is only a crude guide. The estimated effect shown above is closer to 38%.

The authors also used a different estimation approach. Recall the issue of selection bias caused by unmeasured effects. The authors reformulated their model as

$$y_{ij} = \beta_1 + \beta_2 \text{age}_i + \beta_3 \text{age}_i^2 + \beta_4 S_{ij}(j) + \beta_6 \text{sex}_i + \beta_7 \text{race}_i + \mu_i + \varepsilon_{ij}.$$

Unmeasured latent effects, such as "ability," are contained in μ_i. Because μ_i is not observable but is, it is assumed, correlated with other variables in the equation, the least squares regression of y_{ij} on the other variables produces a biased set of coefficient estimates.[29] The difference between the two earnings equations is

$$y_{i1} - y_{i2} = \beta_4[S_{i1}(1) - S_{i2}(2)] + \varepsilon_{i1} - \varepsilon_{i2}.$$

This equation removes the latent effect but, it turns out, worsens the measurement error problem. As before, β_4 can be estimated by instrumental variables. There are two instrumental variables available, $S_{i2}(1)$ and $S_{i1}(2)$. (It is not clear in the paper whether the authors used the two separately or the difference of the two.) The least squares estimate is 0.092, which is comparable to the earlier estimate. The instrumental variable estimate is 0.167, which is nearly 82% higher. The two reported standard errors are 0.024 and 0.043, respectively. With these figures, it is possible to carry out Hausman's test,

$$H = \frac{(0.167 - 0.092)^2}{0.043^2 - 0.024^2} = 4.418.$$

The 95% critical value from the chi-squared distribution with one degree of freedom is 3.84, so the hypothesis that the LS estimator is consistent would be rejected. The square root of H, 2.102, would be treated as a value from the standard normal distribution, from which the critical value would be 1.96. The authors reported a t statistic for this regression of 1.97.

8.9 NONLINEAR INSTRUMENTAL VARIABLES ESTIMATION

In Section 8.2, we extended the linear regression model to allow for the possibility that the regressors might be correlated with the disturbances. The same problem can arise in nonlinear models. The consumption function estimated in Example 7.4 is almost surely

[29] This is a "fixed effects model"—see Section 11.4. The assumption that the latent effect, *ability*, is common between the twins and fully accounted for is a controversial assumption that ability is accounted for by *nature* rather than *nurture*. A search of the Internet on the subject of the "nature versus nurture debate" will turn up millions of citations. We will not visit the subject here.

a case in point. In this section, we will extend the method of instrumental variables to nonlinear regression models.

In the nonlinear model,

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i,$$

the covariates $\mathbf{x}_i$ may be correlated with the disturbances. We would expect this effect to be transmitted to the pseudoregressors, $\mathbf{x}_i^0 = \partial h(\mathbf{x}_i, \boldsymbol{\beta})/\partial\boldsymbol{\beta}$. If so, then the results that we derived for the linearized regression would no longer hold. Suppose that there is a set of variables $[\mathbf{z}_1, \ldots, \mathbf{z}_L]$ such that

$$\operatorname{plim}(1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0} \tag{8-35}$$

and

$$\operatorname{plim}(1/n)\mathbf{Z}'\mathbf{X}^0 = \mathbf{Q}_{\mathbf{zx}}^0 \neq \mathbf{0},$$

where $\mathbf{X}^0$ is the matrix of pseudoregressors in the linearized regression, evaluated at the true parameter values. If the analysis that we used for the linear model in Section 8.3 can be applied to this set of variables, then we will be able to construct a consistent estimator for $\boldsymbol{\beta}$ using the instrumental variables. As a first step, we will attempt to replicate the approach that we used for the linear model. The linearized regression model is given in (7-30),

$$\mathbf{y} = \mathbf{h}(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon} \approx \mathbf{h}^0 + \mathbf{X}^0(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \boldsymbol{\varepsilon},$$

or

$$\mathbf{y}^0 \approx \mathbf{X}^0\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y}^0 = \mathbf{y} - \mathbf{h}^0 + \mathbf{X}^0\boldsymbol{\beta}^0.$$

For the moment, we neglect the approximation error in linearizing the model. In (8-35), we have assumed that

$$\operatorname{plim}(1/n)\mathbf{Z}'\mathbf{y}^0 = \operatorname{plim}(1/n)\mathbf{Z}'\mathbf{X}^0\boldsymbol{\beta}. \tag{8-36}$$

Suppose, as we assumed before, that there are the same number of instrumental variables as there are parameters, that is, columns in $\mathbf{X}^0$. (*Note:* This number need not be the number of variables.) Then the "estimator" used before is suggested,

$$\mathbf{b}_{\mathrm{IV}} = (\mathbf{Z}'\mathbf{X}^0)^{-1}\mathbf{Z}'\mathbf{y}^0. \tag{8-37}$$

The logic is sound, but there is a problem with this estimator. The unknown parameter vector $\boldsymbol{\beta}$ appears on both sides of (8-36). We might consider the approach we used for our first solution to the nonlinear regression model, that is, with some initial estimator in hand, iterate back and forth between the instrumental variables regression and recomputing the pseudoregressors until the process converges to the fixed point that we seek. Once again, the logic is sound, and in principle, this method does produce the estimator we seek.

If we add to our preceding assumptions

$$\frac{1}{\sqrt{n}}\mathbf{Z}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}_{\mathbf{zz}}],$$

then we will be able to use the same form of the asymptotic distribution for this estimator that we did for the linear case. Before doing so, we must fill in some gaps in the preceding. First, despite its intuitive appeal, the suggested procedure for finding the estimator is very unlikely to be a good algorithm for locating the estimates. Second, we do not wish to limit ourselves to the case in which we have the same number of instrumental variables as parameters. So, we will consider the problem in general terms. The estimation criterion for nonlinear instrumental variables is a quadratic form,

$$\begin{aligned}\text{Min}_{\boldsymbol{\beta}}\, S(\boldsymbol{\beta}) &= \tfrac{1}{2}\{[\mathbf{y} - \mathbf{h}(\mathbf{X}, \boldsymbol{\beta})]'\mathbf{Z}\}(\mathbf{Z}'\mathbf{Z})^{-1}\{\mathbf{Z}'[\mathbf{y} - \mathbf{h}(\mathbf{X}, \boldsymbol{\beta})]\} \\ &= \tfrac{1}{2}\boldsymbol{\varepsilon}(\boldsymbol{\beta})'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}(\boldsymbol{\beta}).^{30}\end{aligned} \tag{8-38}$$

The first-order conditions for minimization of this weighted sum of squares are

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\mathbf{X}^{0\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}(\boldsymbol{\beta}) = \mathbf{0}. \tag{8-39}$$

This result is the same one we had for the linear model with $\mathbf{X}^0$ in the role of $\mathbf{X}$. This problem, however, is highly nonlinear in most cases, and the repeated least squares approach is unlikely to be effective. But it is a straightforward minimization problem in the frameworks of Appendix E, and instead, we can just treat estimation here as a problem in nonlinear optimization.

We have approached the formulation of this instrumental variables estimator more or less strategically. However, there is a more structured approach. The orthogonality condition,

$$\text{plim}(1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0},$$

defines a GMM estimator. With the homoscedasticity and nonautocorrelation assumption, the resultant **minimum distance estimator** produces precisely the criterion function suggested above. We will revisit this estimator in this context in Chapter 13.

With well-behaved *pseudoregressors* and instrumental variables, we have the general result for the nonlinear instrumental variables estimator; this result is discussed at length in Davidson and MacKinnon (2004).

THEOREM 8.2 Asymptotic Distribution of the Nonlinear Instrumental Variables Estimator

With well-behaved instrumental variables and pseudoregressors,

$$\mathbf{b}_{\text{IV}} \overset{a}{\sim} N[\boldsymbol{\beta}, (\sigma^2/n)(\mathbf{Q}_{\mathbf{xz}}^0(\mathbf{Q}_{\mathbf{zz}})^{-1}\mathbf{Q}_{\mathbf{zx}}^0)^{-1}].$$

We estimate the asymptotic covariance matrix with

$$\text{Est.Asy.Var}[\mathbf{b}_{\text{IV}}] = \hat{\sigma}^2[\hat{\mathbf{X}}^{0\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{X}}^0]^{-1},$$

where $\hat{\mathbf{X}}^0$ *is* $\mathbf{X}^0$ *computed using* $\mathbf{b}_{\text{IV}}$.

[30]Perhaps the more natural point to begin the minimization would be $S^0(\boldsymbol{\beta}) = [\boldsymbol{\varepsilon}(\boldsymbol{\beta})'\mathbf{Z}][\mathbf{Z}'\boldsymbol{\varepsilon}(\boldsymbol{\beta})]$. We have bypassed this step because the criterion in (8-38) and the estimator in (8-39) will turn out (following and in Chapter 13) to be a simple yet more efficient GMM estimator.

As a final observation, note that the 2SLS interpretation of the instrumental variables estimator for the linear model still applies here, with respect to the IV estimator. That is, at the final estimates, the first-order conditions (normal equations) imply that

$$\mathbf{X}^{0\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \mathbf{X}^{0\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}^{0}\boldsymbol{\beta},$$

which says that the estimates satisfy the normal equations for a linear regression of $\mathbf{y}$ (not $\mathbf{y}^0$) on the predictions obtained by regressing the columns of $\mathbf{X}^0$ on $\mathbf{Z}$. The interpretation is not quite the same here, because to compute the predictions of $\mathbf{X}^0$, we must have the estimate of $\boldsymbol{\beta}$ in hand. Thus, this two-stage least squares approach does not show how to compute $\mathbf{b}_{\text{IV}}$; it shows a characteristic of $\mathbf{b}_{\text{IV}}$.

Example 8.13 Instrumental Variables Estimates of the Consumption Function

The consumption function in Example 7.4 was estimated by nonlinear least squares without accounting for the nature of the data that would certainly induce correlation between $\mathbf{X}^0$ and $\boldsymbol{\varepsilon}$. As done earlier, we will reestimate this model using the technique of instrumental variables. For this application, we will use the one-period lagged value of consumption and one- and two-period lagged values of income as instrumental variables. Table 8.6 reports the nonlinear least squares and instrumental variables estimates. Because we are using two periods of lagged values, two observations are lost. Thus, the least squares estimates are not the same as those reported earlier.

The instrumental variable estimates differ considerably from the least squares estimates. The differences can be deceiving, however. Recall that the MPC in the model is $\beta\gamma Y^{\gamma-1}$. The 2000.4 value for *DPI* that we examined earlier was 6634.9. At this value, the instrumental variables and least squares estimates of the MPC are 1.1543 with an estimated standard error of 0.01234 and 1.08406 with an estimated standard error of 0.008694, respectively. These values do differ a bit, but less than the quite large differences in the parameters might have led one to expect. We do note that the IV estimate is considerably greater than the estimate in the linear model, 0.9217 (and greater than one, which seems a bit implausible).

8.10 NATURAL EXPERIMENTS AND THE SEARCH FOR CAUSAL EFFECTS

Econometrics and statistics have historically been taught, understood, and operated under the credo that "correlation is not causation." But, much of the still-growing field of microeconometrics and some of what we have done in this chapter have been advanced as "causal modeling."[31] In the contemporary literature on treatment effects

TABLE 8.6 Nonlinear Least Squares and Instrumental Variable Estimates

	Instrumental Variables		*Least Squares*	
Parameter	***Estimate***	***Standard Error***	***Estimate***	***Standard Error***
α	627.031	26.6063	468.215	22.788
β	0.040291	0.006050	0.0971598	0.01064
γ	1.34738	0.016816	1.24892	0.1220
σ	57.1681	–	49.87998	–
$\mathbf{e}'\mathbf{e}$	650,369.805	–	495,114.490	–

[31]See, for example, Chapter 2 of Cameron and Trivedi (2005), which is entitled "Causal and Noncausal Models" and, especially, Angrist, Imbens, and Rubin (1996), Angrist and Krueger (2001), and Angrist and Pischke (2009, 2010).

and program evaluation, the point of the econometric exercise really is to establish more than mere statistical association—in short, the answer to the question "Does the program *work*?" requests an econometric response more committed than "the data seem to be consistent with that hypothesis." A cautious approach to econometric modeling has nonetheless continued to base its view of "causality" essentially on statistical grounds.[32]

An example of the sort of causal model considered here is an equation such as Krueger and Dale's (1999) model for earnings attainment and elite college attendance,

$$\ln \textit{Earnings} = \mathbf{x}'\boldsymbol{\beta} + \delta T + \varepsilon,$$

in which δ is the "causal effect" of attendance at an elite college. In this model, T cannot vary autonomously, outside the model. Variation in T is determined partly by the same hidden influences that determine lifetime earnings. Though a causal effect can be attributed to T, measurement of that effect, δ, cannot be done with multiple linear regression. The technique of linear instrumental variables estimation has evolved as a mechanism for disentangling causal influences. As does least squares regression, the method of instrumental variables must be defended against the possibility that the underlying statistical relationships uncovered could be due to "something else." But, when the instrument is the outcome of a "natural experiment," true exogeneity can be claimed. It is this purity of the result that has fueled the enthusiasm of the most strident advocates of this style of investigation. The power of the method lends an inevitability and stability to the findings. This has produced a willingness of contemporary researchers to step beyond their cautious roots.[33] Example 8.14 describes a controversial contribution to this literature. On the basis of a natural experiment, the authors identify a cause-and-effect relationship that would have been viewed as beyond the reach of regression modeling under earlier paradigms.[34]

Example 8.14 Does Television Watching Cause Autism?

The following is the abstract of economists Waldman, Nicholson, and Adilov's (2008) study of autism.[35]

> An extensive literature in medicine investigates the health consequences of early childhood television watching. However, this literature does not address the issue of reverse causation, i.e., does early childhood television watching cause specific health outcomes or do children more likely to have these health outcomes watch more television? This paper uses a natural experiment to investigate the health consequences of early childhood television watching and so is not subject to questions concerning reverse causation. Specifically, we use repeated cross-sectional data from 1972 through 1992 on county-level mental retardation rates, county-level autism rates, and county-level children's cable-television subscription rates to investigate how early childhood television watching affects the prevalence of mental retardation and autism. We find a strong negative correlation

[32]See, among many recent commentaries on this line of inquiry, Heckman and Vytlacil (2007).

[33]See, e.g., Angrist and Pischke (2009, 2010). In reply, Keane (2010, p. 48) opines "What has always bothered me about the 'experimentalist' school is the false sense of certainty it conveys. The basic idea is that if we have a 'really good instrument,' we can come up with 'convincing' estimates of 'causal effects' that are not 'too sensitive to assumptions.'"

[34]See the symposium in the Spring 2010 *Journal of Economic Perspectives*, Angrist and Pischke (2010), Leamer (2010), Sims (2010), Keane (2010), Stock (2010), and Nevo and Whinston (2010).

[35]Extracts from Waldman, M., Nicholson, S. and Adilov, N., "Positive and Negative Mental Health Consequences of Early Childhood Television Watching," Working Paper w17786, National Bureau of Economic Research, Cambridge, 2012.

> between average county-level cable subscription rates when a birth cohort is below three and subsequent mental retardation diagnosis rates, but a strong positive correlation between the same cable subscription rates and subsequent autism diagnosis rates. Our results thus suggest that early childhood television watching has important positive and negative health consequences.

The authors continue (at page 19),

> "We next examine the role of precipitation on autism diagnoses. One possibility concerning the autism results in Table 5 is that the positive coefficients on the main cable variable may not be due to early childhood television watching being a trigger for autism but rather to some other factor positively associated with precipitation being a trigger. That is, Waldman et al. (2008)[36] find a positive correlation between the precipitation a cohort experiences prior to age three and the cohort's subsequent autism diagnosis rate, where the interpretation put forth in that paper is that there is an environmental trigger for autism positively correlated with precipitation that drives up the autism diagnosis rate when precipitation prior to age three is high. Possibilities include any potential trigger positively associated with indoor activity such as early childhood television watching, which is the focus here, vitamin D deficiency which could be more common when children are indoors more and not exposed to the sun, and any indoor chemical where exposure will be higher when the child spends more time indoors. So one possibility concerning the results in Table 5 is that cable and precipitation are positively correlated and early childhood television watching is not a trigger for autism. In this scenario the positive and statistically significant cable coefficients found in the table would not be due to the positive correlation between cable and early childhood television watching, but rather to one of these other factors being the trigger and the positive coefficients arise because cable, through a correlation with precipitation, is also correlated with this unknown 'other' trigger."

They conclude (on p 30): "We believe our results are sufficiently suggestive of early childhood television watching decreasing mental retardation and increasing autism that clinical studies focused on the health effects of early childhood television watching are warranted. Only a clinical study can show definitively the health effects of early childhood television watching."

The authors add (at page 3), "Although consistent with the hypothesis that early childhood television watching is an important trigger for autism, our first main finding is also consistent with another possibility. Specifically, because precipitation is likely correlated with young children spending more time indoors generally, not just young children watching more television, our first main finding could be due to any indoor toxin. *Therefore, we also employ a second instrumental variable or natural experiment, that is correlated with early childhood television watching but unlikely to be substantially correlated with time spent indoors*." (Emphasis added.) They conclude (on pp. 39–40): "Using the results found in Table 3's pooled cross-sectional analysis of California, Oregon, and Washington's county-level autism rates, we find that if early childhood television watching is the sole trigger driving the positive correlation between autism and precipitation then thirty-eight percent of autism diagnoses are due to the incremental television watching due to precipitation."

---

[36]Waldman, M., S. Nicholson, N. Adilov, and J. Williams, "Autism Prevalence and Precipitation Rates in California, Oregon, and Washington Counties," *Archives of Pediatrics & Adolescent Medicine*, 162, 2008, pp. 1026–1034.

Waldman, Nicholson, and Adilov's (2008)[37] study provoked an intense and widespread response among academics, autism researchers, and the public. Whitehouse (2007), writing in the *Wall Street Journal* (http://www.wsj.com/articles/SB117131554110006323), surveyed some of the discussion, which touches upon the methodological implications of the search for "causal effects" in econometric research. The author lamented that the power of techniques involving instrumental variables and natural experiments to uncover causal relationships had emboldened economists to venture into areas far from their traditional expertise, such as the causes of autism [Waldman et al. (2008)].[38]

Example 8.15 Is Season of Birth a Valid Instrument?

Buckles and Hungerman (BH, 2008) list more than 20 studies of long-term economic outcomes that use season of birth as an instrumental variable, beginning with one of the earliest and best-known papers in the "natural experiments" literature, Angrist and Krueger (1991). The assertion of the validity of season of birth as a proper instrument is that family background is unrelated to season of birth, but it is demonstrably related to long-term outcomes such as income and education. The assertion justifies using dummy variables for season of birth as instrumental variables in outcome equations. If, on the other hand, season of birth is correlated with family background, then it will "fail the exclusion restriction in most IV settings where it has been used" (BH, page 2). According to the authors, the randomness of quarter of birth over the population[39] has been taken as a given, without scientific investigation of the claim. Using data from live birth certificates and census data, BH found a numerically modest, but statistically significant relationship between birth dates and family background. They found "women giving birth in the winter look different from other women; they are younger, less educated, and less likely to be married . . . . The fraction of children born to women without a high school degree is about 10% higher (2 percentage points) in January than in May . . . We also document a 10% decline in the fraction of children born to teenagers from January to May." Precisely why there should be such a relationship remains uncertain. Researchers differ (of course) on the numerical implications of BH's finding.[40] But, the methodological implication of their finding is consistent with the observation in Whitehouse's article, that bad instruments can produce misleading results.

8.11 SUMMARY AND CONCLUSIONS

The instrumental variable (IV) estimator, in various forms, is among the most fundamental tools in econometrics. Broadly interpreted, it encompasses most of the estimation methods that we will examine in this book. This chapter has developed the basic results for IV estimation of linear models. The essential departure point is the exogeneity and relevance assumptions that define an instrumental variable. We then analyzed linear IV estimation in the form of the two-stage least squares estimator. With only a few special exceptions related to simultaneous equations models with two variables, almost no finite-sample properties have been established for the IV estimator. (We temper that,

[37]Published as NBER working paper 12632 in 2006.

[38]Whitehouse criticizes the use of proxy variables, e.g., Waltman's use of rainfall patterns for TV viewing. As we have examined in this chapter, an instrumental variable is not a proxy and this mischaracterizes the technique. It remains true, as emphasized by some prominent researchers quoted in the article, that a bad instrument can produce misleading results.

[39]See, for example, Kleibergen (2002).

[40]See Lahart (2009).

however, with the results in Section 8.7 on weak instruments, where we saw evidence that whatever the finite-sample properties of the IV estimator might be, under some well-discernible circumstances, these properties are not attractive.) We then examined the asymptotic properties of the IV estimator for linear and nonlinear regression models. Finally, some cautionary notes about using IV estimators when the instruments are only weakly relevant in the model are examined in Section 8.7.

Key Terms and Concepts

- AttenuationAsymptotic covariance matrix
- Asymptotic distribution
- Attenuation bias
- Attrition bias
- Attrition
- Consistent estimator
- Effect of the treatment on the treated
- Endogenous treatment effect
- Endogenous
- Exogenous
- Identification
- Indicator
- Instrumental variable estimator
- Instrumental variables (IV)
- Limiting distribution
- Minimum distance estimator
- Moment equations
- Natural experiment
- Nonrandom sampling
- Omitted parameter heterogeneity
- Omitted variable bias
- Omitted variables
- Orthogonality conditions
- Overidentification
- Panel data
- Proxy variable
- Random effects
- Reduced form equation
- Relevance
- Reliability ratio
- Sample selection bias
- Selectivity effect
- Simultaneous equations bias
- Simultaneous equations
- Smearing
- Structural equation system
- Structural model
- Structural specification
- Survivorship bias
- Truncation bias
- Two-stage least squares (2SLS)
- Variable addition test
- Weak instruments

Exercises

1. In the discussion of the instrumental variable estimator, we showed that the least squares estimator, $\mathbf{b}_{LS}$, is biased and inconsistent. Nonetheless, $\mathbf{b}_{LS}$ does estimate something—plim $\mathbf{b} = \boldsymbol{\theta} = \boldsymbol{\beta} + \mathbf{Q}^{-1}\boldsymbol{\gamma}$. Derive the asymptotic covariance matrix of $\mathbf{b}_{LS}$ and show that $\mathbf{b}_{LS}$ is asymptotically normally distributed.
2. For the measurement error model in (8-26) and (8-27), prove that when only x is measured with error, the squared correlation between y and x is less than that between y^* and x^*. (Note the assumption that $y_* = y$.) Does the same hold true if y^* is also measured with error?
3. Derive the results in (8-32a) and (8-32b) for the measurement error model. Note the hint in Footnote 4 in Section 8.5.1 that suggests you use result (A-66) when you need to invert

$$[\mathbf{Q}^* + \boldsymbol{\Sigma}_{uu}] = [\mathbf{Q}^* + (\sigma_u\mathbf{e}_1)(\sigma_u\mathbf{e}_1)'].$$

4. At the end of Section 8.7, it is suggested that the OLS estimator could have a smaller mean squared error than the 2SLS estimator. Using (8-4), the results of Exercise 1, and Theorem 8.1, show that the result will be true if

$$\mathbf{Q}_{\mathbf{XX}} - \mathbf{Q}_{\mathbf{XZ}}\mathbf{Q}_{\mathbf{ZZ}}^{-1}\mathbf{Q}_{\mathbf{ZX}} \gg \frac{1}{(\sigma^2/n) + \boldsymbol{\gamma}'\mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\gamma}}\boldsymbol{\gamma}\boldsymbol{\gamma}'.$$

How can you verify that this is at least possible? The right-hand side is a rank one, nonnegative definite matrix. What can be said about the left-hand side?

5. Consider the linear model, $y_i = \alpha + \beta x_i + \varepsilon_i$, in which $\text{Cov}[x_i, \varepsilon_i] = \gamma \neq 0$. Let z be an exogenous, relevant instrumental variable for this model. Assume, as well, that z is binary—it takes only values 1 and 0. Show the algebraic forms of the LS estimator and the IV estimator for both α and β.
6. This is easy to show. In the expression for $\hat{\mathbf{X}}$, if the kth column in $\mathbf{X}$ is one of the columns in $\mathbf{Z}$, say the lth, then the kth column in $(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ will be the lth column of an $L \times L$ identity matrix. This result means that the kth column in $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ will be the lth column in $\mathbf{Z}$, which is the kth column in $\mathbf{X}$.
7. Prove that the control function approach in (8-16) produces the same estimates as 2SLS.
8. Prove that in the control function estimator in (8-16), you can use the predictions, $\mathbf{z}'\mathbf{p}$, instead of the residuals to obtain the same results apart from the sign on the control function itself, which will be reversed.

Applications

1. In Example 8.5, we have suggested a model of a labor market. From the "reduced form" equation given first, you can see the full set of variables that appears in the model—that is the "endogenous variables," ln $Wage_{it}$, and Wks_{it}, and all other exogenous variables. The labor supply equation suggested next contains these two variables and three of the exogenous variables. From these facts, you can deduce what variables would appear in a labor "demand" equation for ln $Wage_{it}$. Assume (for purpose of our example) that ln $Wage_{it}$ is determined by Wks_{it} and the remaining appropriate exogenous variables. (We should emphasize that this exercise is purely to illustrate the computations—the structure here would not provide a theoretically sound model for labor market equilibrium.)
   a. What is the labor demand equation implied?
   b. Estimate the parameters of this equation by OLS and by 2SLS and compare the results. (Ignore the panel nature of the data set. Just pool the data.)
   c. Are the instruments used in this equation relevant? How do you know?