8-1

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,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad \textbf{(8-1)}$$

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 that in Example 6.5, 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, ~~(that 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.

KT

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

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

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

double quote

$$\ln G = \beta_1 + \beta_2 \ln Price + w,$$

where $w = \beta_3 \ln Income + \varepsilon$. 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.

KT

Endogenous Treatment Effects: Kreuger and Dale (1999) 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

Note double quotes

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

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 estimate of δ will inappropriately

8-2

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, e.g., Imbens and Wooldridge (2009) for a survey.

for example/

AU: OK to spell out "e.g." in text?

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

(Demand) $Quantity_D = \alpha_0 + \alpha_1 Price + \alpha_2 Income + \varepsilon_D,$

(Supply) $Quantity_S = \beta_0 + \beta_1 Price + \beta_2 InputPrice + \varepsilon_S,$

(Equilibrium) $Quantity_D = Quantity_S.$

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 $Price = (\alpha_0 - \beta_0 + \alpha_2 Income - \beta_2 InputPrice + \varepsilon_D - \varepsilon_S)/(\beta_1 - \alpha_1)$. 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.

minus

double quotes

KT

Dynamic Panel Data Models: In Chapter 11, we will examine a **random effects** dynamic model of the form $y_{it} = x_{it}\beta + \gamma y_{i,t-1} + \varepsilon_{it} + u_i$ 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),

minus

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

where $\Delta \ln Y_{it}$ is the growth rate of country i in year t. [See, for example, Lee, Pesaran and Smith (1997).] 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

minus minus

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

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

double quotes

minus

minus

(minus) (minus)

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 KT

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

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}. Since schooling is correlated with w_{it}, it is endogenous, and least squares is not a suitable estimator of the earnings equation. As we will show later, in cases such as this one, the mismeasurement of a relevant variable causes a particular form of inconsistency, KT **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 full population. Merely unusually high would not be a problem save for the general finding that the explanation for the nonrandomnes is found at least in part in the variables that appear elsewhere in the model. This nonrandomness of the sample of the sample translates to a form of omitted variable bias known as **sample selection bias.** KT

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 sample of JTPA participants noted above. 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 analysis 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

double quotes

previous

KT

KT

KT

8-4

self assessed health outcomes in a long panel data set extracted from the British Household Panel Data 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.

double quotes

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 two 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 18 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 $\mathbf{x}_i$ maybe 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.

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, 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 mutiple 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.

8-5

way that will be established following) that this is, once again, a case in which the disturbance and an independent variable are correlated.

None of these models can be consistently estimated by least squares—the method of instrumental variables is the standard approach.

There is an alternative method of estimation called the method of **instrumental variables** (IV). The least squares estimator is a special case, but the IV method is far more general. The method of instrumental variables is developed around the following general extension of the estimation strategy in the classical regression model: Suppose that in the classical model of (12-1), the K variables $\mathbf{x}_i$ may be correlated with ε_i. Suppose as well that there exists a set of L variables $\mathbf{z}_i$, where L is at least as large as K, such that $\mathbf{z}_i$ is correlated with $\mathbf{x}_i$ but not with ε_i. We cannot estimate β consistently by using the familiar least squares estimator. But we can construct a consistent estimator of β by using the assumed relationships among $\mathbf{z}_i$, $\mathbf{x}_i$, and ε_i.

8.2 ~~12.2~~ ASSUMPTIONS OF THE EXTENDED MODEL

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

A1. Linearity: $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$.

A2. Full rank: The $n \times K$ sample data matrix, $\mathbf{X}$ has full column rank

A3. 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.

A4. 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 $\mathbf{X}$.

A5. Stochastic or nonstochastic data: $(x_{i1}, x_{i2}, \ldots, x_{iK})$ $i = 1, \ldots, n$.

A6. 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, A3 (no correlation between $\mathbf{x}$ and ε) is, under our new assumptions,

$$\textbf{AI3.} \quad E[\varepsilon_i \,|\, \mathbf{x}_i] = \eta_i.$$

We interpret Assumption AI3 to mean that the regressors now provide information about the expectations of the disturbances. The important implication of AI3 is that the disturbances and the regressors are now correlated. Assumption AI3 implies that

$$E[\mathbf{x}_i\varepsilon_i] = \boldsymbol{\gamma} \qquad \text{(8-2 ~~12-2~~)}$$

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

$$\text{plim}\,(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \boldsymbol{\gamma}. \qquad \text{(8-3 ~~12-3~~)}$$

Notice that the original model results if $\eta_i = 0$. The implication of (8-3 ~~12-3~~) is that the regressors, $\mathbf{X}$, are no longer exogenous.

8-6

We now assume that there is an additional set of variables, $\mathbf{Z}$, that have two properties:

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

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:

AI7. $[\mathbf{x}_i, \mathbf{z}_i, \varepsilon_i], i = 1, \ldots, n$, are an i.i.d. sequence of random variables.
AI8a. $E[x_{ik}^2] = \mathbf{Q}_{\mathbf{xx},kk} < \infty$, a finite constant, $k = 1, \ldots, K$.
AI8b. $E[z_{il}^2] = \mathbf{Q}_{\mathbf{zz},ll} < \infty$, a finite constant, $l = 1, \ldots, L$.
AI8c. $E[z_{il}x_{ik}] = \mathbf{Q}_{\mathbf{zx},lk} < \infty$, a finite constant, $l = 1, \ldots, L$, $k = 1, \ldots, K$.
AI9. $E[\varepsilon_i \mid \mathbf{z}_i] = 0$.

In later work in time series models, it will be important to relax assumption AI7. Finite means of $\mathbf{z}_l$ follows from AI8b. Using the same analysis as in Section 4.9, we have 4.4

$$\text{plim}\,(1/n)\mathbf{Z}'\mathbf{Z} = \mathbf{Q}_{\mathbf{zz}}, \text{ a finite, positive definite matrix (}\textit{well behaved data}\text{)},$$

$$\text{plim}\,(1/n)\mathbf{Z}'\mathbf{X} = \mathbf{Q}_{\mathbf{zx}}, \text{ a finite, } L \times K \text{ matrix with rank } K \text{ (}\textit{relevance}\text{)},$$

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

In our statement of the classical regression model, we have assumed thus far the special case of $\eta_i = 0$; $\boldsymbol{\gamma} = \mathbf{0}$ follows. There is no need to dispense with Assumption AI7—it may well continue to be true—but in this special case, it becomes irrelevant.

8.3 ~~12.3~~ ESTIMATION

For this more general model of (12-3), we lose most of the useful results we had for least squares. We will consider the implications for least squares, then construct an alternative estimator for $\boldsymbol{\beta}$ in this extended model.

12.3.1 ORDINARY LEAST SQUARES

The estimator $\mathbf{b}$ is no longer unbiased;

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

so the Gauss–Markov theorem no longer holds. It is also inconsistent;

$$\text{plim}\,\mathbf{b} = \boldsymbol{\beta} + \text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \text{plim}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right) = \boldsymbol{\beta} + \mathbf{Q}_{\mathbf{xx}}^{-1}\boldsymbol{\gamma} \neq \boldsymbol{\beta}. \tag{12-4}$$

(The asymptotic distribution is considered in the exercises.)

12.3.2 THE INSTRUMENTAL VARIABLES ESTIMATOR

Because $E[\mathbf{z}_i\varepsilon_i] = \mathbf{0}$ and all terms have finite variances, it follows that

$$\text{plim}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right) = \text{plim}\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) - \text{plim}\left(\frac{\mathbf{Z}'\mathbf{X}\boldsymbol{\beta}}{n}\right) = \mathbf{0}.$$

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 Kreuger 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 among the instrumental 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 the x_2 = *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,*Income*,*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.

double quotes

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 regression 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 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, then construct an alternative estimator for $\boldsymbol{\beta}$ in this extended model.

and

8.3.1 LEAST SQUARES

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

minus

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

so the Gauss-Markov theorem no longer holds. The estimator is also inconsistent;

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

(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

8-8

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

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

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

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 effects 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

$$\text{plim}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right) = \text{plim}\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) - \text{plim}\left(\frac{\mathbf{Z}'\mathbf{X}\boldsymbol{\beta}}{n}\right) = \mathbf{0}.$$

Therefore,

$$\text{plim}\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = \left[\text{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]\boldsymbol{\beta} + \text{plim}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right) = \left[\text{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]\boldsymbol{\beta}. \qquad \text{8-5 } \cancel{(12\text{-}5)}$$

We have assumed that ~~Suppose that~~ **Z** has the same number of variables as **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 assumed that the rank of $\mathbf{Z}'\mathbf{X}$ is K, so now $\mathbf{Z}'\mathbf{X}$ is a square matrix. It follows that

$$\left[\text{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]^{-1}\text{plim}\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = \boldsymbol{\beta}, \qquad (8\text{-}6)$$

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

$$\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}.$$

We have already proved that $\mathbf{b}_{\text{IV}}$ is consistent. We now turn to the **asymptotic distribution.** We will use the same method as in Section ~~4.9.2~~ 4.4.2. First,

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

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 $(1/\sqrt{n})\mathbf{X}'\boldsymbol{\varepsilon}$ in Section 4.9.2, so it follows that

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

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 ~~12.1~~ 8.1 Asymptotic Distribution of the Instrumental Variables Estimator

If Assumptions A1, A2, AI3, A4, A5, AI7, AI8a–c, and AI9 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* $\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$ *is*

$$\mathbf{b}_{\text{IV}} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}_{\mathbf{zx}}^{-1}\mathbf{Q}_{\mathbf{zz}}\mathbf{Q}_{\mathbf{xz}}^{-1}\right]. \qquad \cancel{(12\text{-}6)}\ 8\text{-}\cancel{4}\,7$$

where $\mathbf{Q}_{\mathbf{zx}} = \text{plim}(\mathbf{Z}'\mathbf{X}/n)$ *and* $\mathbf{Q}_{\mathbf{zz}} = \text{plim}(\mathbf{Z}'\mathbf{Z}/n)$.

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

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

8-10

A correction for degrees of freedom is superfluous, as all results here are asymptotic, and $\hat{\sigma}^2$ would not be unbiased in any event. (Nonetheless, it is standard practice in most software to make the degrees of freedom correction.) Write the vector of residuals as

$$\mathbf{y} - \mathbf{X}\mathbf{b}_{\text{IV}} = \mathbf{y} - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}.$$

Substitute $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$ and collect terms to obtain $\hat{\boldsymbol{\varepsilon}} = [\mathbf{I} - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}']\boldsymbol{\varepsilon}$. Now,

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n} \\ &= \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} + \left(\frac{\boldsymbol{\varepsilon}'\mathbf{Z}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{Z}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right) - 2\left(\frac{\boldsymbol{\varepsilon}'\mathbf{X}}{n}\right)\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n}\right).\end{aligned}$$

We found earlier that we could (after a bit of manipulation) apply the product result for probability limits to obtain the probability limit of an expression such as this. Without KT repeating the derivation, we find that $\hat{\sigma}^2$ is a **consistent estimator** of σ^2, by virtue of the first term. The second and third product terms converge to zero. To complete the derivation, then, we will estimate Asy. Var$[\mathbf{b}_{\text{IV}}]$ with

$$\begin{aligned}\text{Est. Asy. Var}[\mathbf{b}_{\text{IV}}] &= \frac{1}{n}\left\{\left(\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}\right)\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{Z}'\mathbf{Z}}{n}\right)\left(\frac{\mathbf{X}'\mathbf{Z}}{n}\right)^{-1}\right\} \\ &= \hat{\sigma}^2(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Z})(\mathbf{X}'\mathbf{Z})^{-1}.\end{aligned} \qquad \text{8-8 } ~~\textbf{(12-7)}~~$$

12.3.3 TWO-STAGE LEAST SQUARES

There is a remaining detail. If $\mathbf{Z}$ contains more variables than $\mathbf{X}$, then much of the preceding is unusable, because $\mathbf{Z}'\mathbf{X}$ will be $L \times K$ with rank $K < L$ and will thus not have an inverse. The crucial result in all the preceding is $\text{plim}(\mathbf{Z}'\boldsymbol{\varepsilon}/n) = \mathbf{0}$. 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 $\mathbf{Z}$. But intuition correctly suggests that throwing away the information contained in the remaining $L - K$ columns is inefficient. A better choice is the projection of the columns of $\mathbf{X}$ in the column space of $\mathbf{Z}$:

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

We will return shortly to the virtues of this choice. With this choice of instrumental variables, $\hat{\mathbf{X}}$ for $\mathbf{Z}$, we have

$$\begin{aligned}\mathbf{b}_{\text{IV}} &= (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\ &= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}.\end{aligned} \qquad \textbf{(12-8)}$$

The estimator of the asymptotic covariance matrix will be $\hat{\sigma}^2$ times the bracketed matrix in (12-8). 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

8.3.3 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

$$\mathbf{Q}_{\mathbf{Z}\mathbf{y}} = \mathbf{Q}_{\mathbf{Z}\mathbf{X}}\boldsymbol{\beta}.$$

The IV estimator is obtained by solving this set of K **moment equations**. Since 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

$$\mathbf{Q}_{\mathbf{X}\mathbf{y}} = \mathbf{Q}_{\mathbf{X}\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\gamma},$$

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 $\mathbf{b}_{\text{IV}}$.

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, i.e., 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. To summarize, then, in order for a regression model correctly to identify δ, it must be assumed that variation in T is not associated with variation in ε. If it is, then as seen in (8-4), variation in y comes about through an additional source, variation in ε that is transmitted through variation in T. That is the influence of γ in (8-4). 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, $(\Delta y/\Delta z)/(\Delta T/\Delta z)$.

KT

AU: KT "identified" not in chap. list

double quotes

that is,

AU: OK to spell out "i.e." in text?

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 (1844, 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."

8-13

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

$$\frac{A_{ikm}}{\ln E_{km}} = \beta_1 C_m + \mathbf{x}'_{ikm}\beta_2 + \overline{\mathbf{x}}'_{.km}\beta_3 + \overline{\mathbf{x}}'_{..m}\beta_4 + \varepsilon_{ikm} + \varepsilon_{km} + \varepsilon_m$$

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 and 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. [The controversial topic of the study and the unconventional choice of instruments caught the attention of the popular press, for example, http://gsppi.berkeley.edu/faculty/jrothstein/hoxby/wsj.pdf, and academic observers including Rothstein (2004).] This study is an example of a "natural experiment" as described in Angrist and Pischke (2009). URL

Example 8.4 Instrumental Variable in Regression

FIG 8.1

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 ($\varepsilon_S,\varepsilon_D$) 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 income values on the demand side produce nine regimes, punctuated 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, 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.

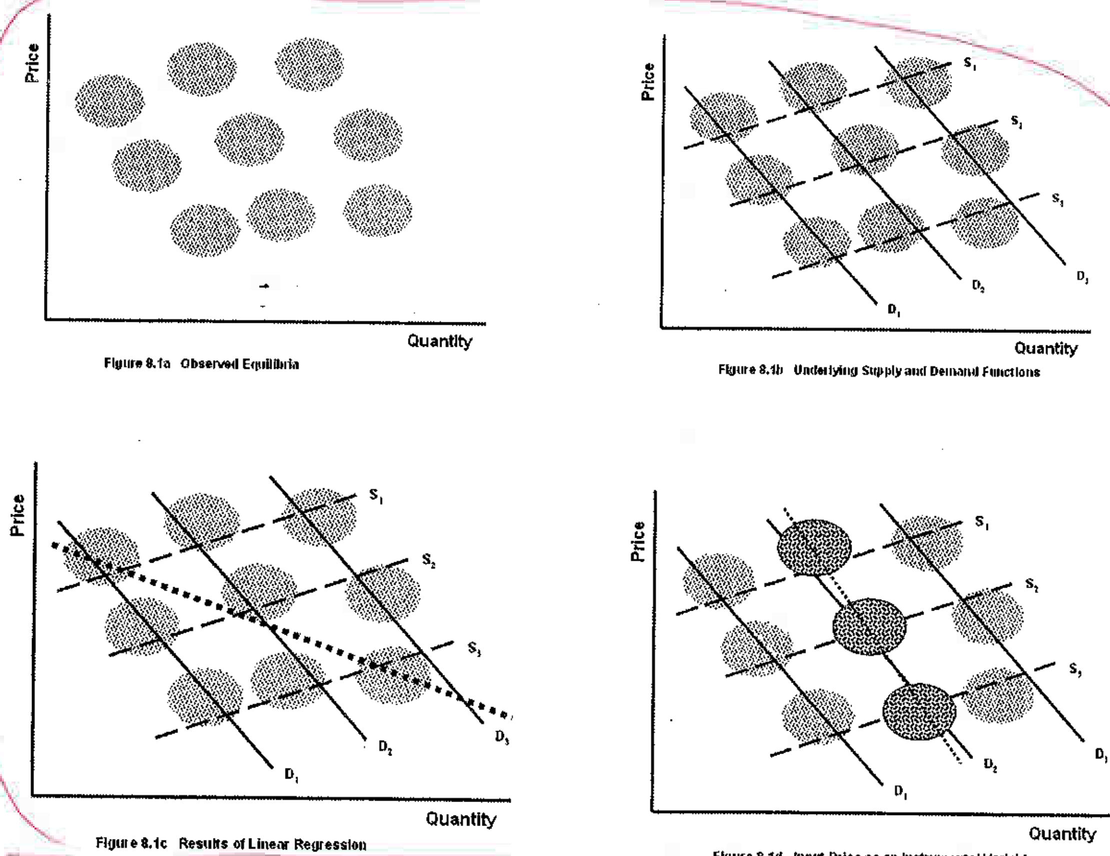

Figure 8.1a Observed Equilibria
Figure 8.1b Underlying Supply and Demand Functions
Figure 8.1c Results of Linear Regression
Figure 8.1d Input Price as an Instrumental Variable

FIGURE 8.1 Identifying a Demand Curve with an Instrumental Variable

8.3.4 TWO-STAGE LEAST SQUARES

Thus far, we have assumed that the number of instrumental variables in **Z** is the same as the number of variables (exogenous plus endogenous) in **X**. (In the typical application, the researcher provides the necessary instrumental variable for the single endogenous variable in their equation.) However, it is possible that the data contain 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

$$
\begin{aligned}
&\text{(Demand)} && Quantity_D = \alpha_0 + \alpha_1 Price + \alpha_2 Income + \varepsilon_D,\\
&\text{(Supply)} && Quantity_S = \beta_0 + \beta_1 Price + \beta_2 InputPrice + \beta_3 Rainfall + \varepsilon_S,\\
&\text{(Equilibrium)} && Quantity_D = Quantity_S.
\end{aligned}
$$

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. (We will turn to the now apparent problem of how to identify the supply equation in Section 8.4.2.) 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 **Z**, even when $L > K$. The method of two stage least squares solves the problem of how to use all the information in the sample when **Z** contains more variables than are necessary to construct an instrumental variable estimator.

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

double quotes

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

We will return shortly to the virtues of this choice. With this choice of instrumental variables, $\hat{\mathbf{X}}$ for **Z**, we have

$$\mathbf{b}_{\text{IV}} = (\hat{\mathbf{X}}'\mathbf{X})\hat{\mathbf{X}}'\mathbf{y} = [\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'X}]^{-1}\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'y}. \qquad \textbf{(8-9)}$$

The estimator of the asymptotic covariance matrix will be $\hat{\sigma}^2$ times the bracketed matrix in (8-9). 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 **X** also appears in **Z**, then that column of **X** is reproduced exactly in

$\hat{\mathbf{X}}$. 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 $\mathbf{X}$. This result is important and useful. Consider what is probably the typical application. Suppose that 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 $\mathbf{Z}$, which is obvious from the definition. 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.

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 $\mathbf{ZF}$ 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 $\mathbf{Z}$. This important result was proved in a seminal paper by Brundy and Jorgenson (1971). [See, also, Wooldridge (2002a, pp. 96–97).]

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

$$\begin{aligned}\mathbf{b}_{\mathrm{IV}} &= (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\ &= (\mathbf{X}'(\mathbf{I}-\mathbf{M}_Z)\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{M}_Z)\mathbf{y} \\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}\end{aligned} \qquad \text{8-10 } \cancel{(12\text{-}9)}$$

because $\mathbf{I}-\mathbf{M}_Z$ 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 (only logically; one need not actually do this in two steps), 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. We will revisit this form of estimator at great length at several points later, particularly in our discussion of simultaneous equations models in Section 10.5. 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

$$s_{\mathrm{IV}}^2 = \frac{(\mathbf{y}-\hat{\mathbf{X}}\mathbf{b}_{\mathrm{IV}})'(\mathbf{y}-\hat{\mathbf{X}}\mathbf{b}_{\mathrm{IV}})}{n}$$

is inconsistent for σ^2, with or without a correction for degrees of freedom.

An obvious question is where one is likely to find a suitable set of instrumental variables. 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. The asymptotic covariance matrix of the IV estimator can be rather large if $\mathbf{Z}$ is not highly correlated

The recent literature on "natural experiments" focuses on policy changes such as the Mariel Boatlift (Example 6.5) or natural outcomes such as occurrences of streams (Example 8.3) or birthdays [Angrist (1992, 1994)].

8-17

with $\mathbf{X}$; the elements of $(\mathbf{Z}'\mathbf{X})^{-1}$ grow large. (See Section 12.9 on "weak" instruments.) Unfortunately, there usually is not much choice in the selection of instrumental variables. The choice of $\mathbf{Z}$ is often ad hoc.[1] There is a bit of a dilemma in this result. It would seem to suggest that the best choices of instruments are variables that are highly correlated with $\mathbf{X}$. But the more highly correlated a variable is with the problematic columns of $\mathbf{X}$, the less defensible the claim that these same variables are *uncorrelated* with the disturbances.

FN 1

Example 12.2 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

$$\frac{A_{ikm}}{\ln E_{km}} = \beta_1 C_m + \mathbf{x}'_{ikm}\beta_2 + \overline{\mathbf{x}}'_{.km}\beta_3 + \overline{\mathbf{x}}'_{..m}\beta_4 + \cdots + \varepsilon_{ikm} + \varepsilon_{km} + \varepsilon_m$$

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. Note as well that the model specifies the nested random effects model of Section 9.7.1 as well as the Mundlak treatment discussed in Section 9.5.5. The exogenous variables are intended to capture the different sources of heterogeneity at all three levels of aggregation. 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 and 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. [The controversial topic of the study and the unconventional choice of instruments caught the attention of the popular press, for example, http://www.economicprincipals.com/issues/05.10.30.html, and academic observers, see, e.g., Rothstein (2004).]

Example 12.3 Labor Supply Model

A leading example of a model in which correlation between regressor and disturbance is likely to arise is in market equilibrium models. In Example 9.1, we built a "reduced form" wage equation,

$$\ln Wage_{it} = \beta_1 + \beta_2 Exp_{it} + \beta_3 Exp^2_{it} + \beta_4 Wks_{it} + \beta_5 Occ_{it} + \beta_6 Ind_{it} + \beta_7 South_{it} + \beta_8 SMSA_{it} + \beta_9 MS_{it} + \beta_{10} Union_{it} + \beta_{11} Ed_i + \beta_{12} Fem_i + \beta_{13} Blk_i + \varepsilon_{it}.$$

We will return to the idea of reduced forms in the setting of simultaneous equations models in Chapter 13. For the present, the implication for our estimated model is that this 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

$$Wks_{it} = \gamma_1 + \gamma_2 \ln Wage_{it} + \gamma_3 Ed_i + \gamma_4 Union_{it} + \gamma_5 Fem_i + u_{it}$$

(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, which simply contain all the

[1] Results on "optimal instruments" appear in White (2001) and Hansen (1982). In the other direction, there is a contemporary literature on "weak" instruments, such as Staiger and Stock (1997), which we will explore in Section 12.9.

8-18

Example ~~9.7~~ 8.5 Instrumental Variable Estimation of a Labor Supply Equation

A leading example of a model in which correlation between a regressor and the disturbance is likely to arise is in market equilibrium models. 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 1976-1982 from the "Non-Survey of Economic Opportunity" from the Panel Study of Income Dynamics. The estimating equation is

$$\ln Wage_{it} = \beta_1 + \beta_2\, Exp_{it} + \beta_3\, Exp_{it}^2 + \beta_4\, Wks_{it} + \beta_5\, Occ_{it} + \beta_6\, Ind_{it} + \beta_7\, South_{it} + \beta_8\, SMSA_{it} + \beta_9\, MS_{it} + \beta_{10}\, Union_{it} + \beta 11\, Ed_i + \beta_{12}\, Fem_i + \beta_1 3\, Blk_i + \varepsilon_{it}$$

where the variables are

Exp = years of full time work experience, 0 if not,
Wks = weeks worked, 0 if not,
Occ = 1 if blue collar occupation, 0 if not,
Ind = 1 if the individual works in a manufacturing industry, 0 if not,
South = 1 if the individual resides in the south, 0 if not,
SMSA = 1 if the individual resides in an SMSA, 0 if not,
MS = 1 if the individual is married, 0 if not,
Union = 1 if the individual wage is set by a union contract, 0 if not,
Ed = years of education,
Fem = 1 if the individual is female, 0 if not,
Blk = 1 if the individual is black, 0 if not.

See Appendix Table F8.1 for the data source. The main interest of the study, beyond comparing various estimation methods, is β_{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 in Section 8.3.4 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 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

$$Wks_{it} = \gamma_1 + \gamma_2 \ln Wage_{it} + \gamma_3\, Ed_i + \gamma_4\, Union_{it} + \gamma_5\, Fem_i + u_{it}.$$

(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 $\ln Wage_{it}$ 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

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

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 estimate developed in Section 8.3.2. 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 - 19

TABLE 12.1 Estimated Labor Supply Equation

	OLS		*IV with* $\mathbf{Z}_1$		*IV with* $\mathbf{Z}_2$	
Variable	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*
Constant	44.7665	1.2153	18.8987	13.0590	30.7044	4.9997
ln Wage	0.7326	0.1972	5.1828	2.2454	3.1518	0.8572
Education	−0.1532	0.03206	−0.4600	0.1578	−0.3200	0.06607
Union	−1.9960	0.1701	−2.3602	0.2567	−2.1940	0.1860
Female	−1.3498	0.2642	0.6957	1.0650	−0.2378	0.4679

variables in the model.) If the number of weeks worked and the accepted wage offer are determined jointly, then ln *Wage* 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

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

Table 12.1 presents the three sets of estimates. The OLS results 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 estimate developed in Section 12.3.2. As might be expected, the log wage coefficient becomes considerably larger. The other coefficients are, perhaps, contradictory. One have might 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.

12.4 THE HAUSMAN AND WU SPECIFICATION TESTS AND AN APPLICATION TO INSTRUMENTAL VARIABLE ESTIMATION

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

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

To compare the two matrices in the brackets, we can compare their inverses. The inverse of the first is $\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{X}'(\mathbf{I} - \mathbf{M}_\mathbf{Z})\mathbf{X} = \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{M}_\mathbf{Z}\mathbf{X}$. Because $\mathbf{M}_\mathbf{Z}$ is a nonnegative definite matrix, it follows that $\mathbf{X}'\mathbf{M}_\mathbf{Z}\mathbf{X}$ is also. So, $\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ equals $\mathbf{X}'\mathbf{X}$ minus a nonnegative definite matrix. Because $\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ is smaller, in the matrix sense, than $\mathbf{X}'\mathbf{X}$, its inverse is larger. Under the hypothesis, the asymptotic covariance

8 - 20

8.4 TWO SPECIFICATION TESTS

them

There are two aspects of the model that we would be interested in verifying if possible, rather than assuming 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. The IV estimator is robust; it is consistent whether or not $\text{plim}(\mathbf{X}'\varepsilon/n) = \mathbf{0}$. However, if not needed, that is if $\boldsymbol{\gamma} = \mathbf{0}$, then least squares would be a better estimator by virtue of its smaller variance.[2] 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 so 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 beginning 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 bio-fuels, 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 since the parameters of the demand equation can be estimated without this assumption. This section will describe two specification tests that considers these aspects of the IV estimator.

FN 2

AU: Confirm "agricultural market example" is in Sec. 8.3.4.

[2] It is possible, of course, 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. We will revisit the issue in passing at a few points later in the text.

TABLE 12.1 Estimated Labor Supply Equation

	OLS		IV with $\mathbf{Z}_1$		IV with $\mathbf{Z}_2$	
Variable	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*
Constant	44.7665	1.2153	18.8987	13.0590	30.7044	4.9997
ln Wage	0.7326	0.1972	5.1828	2.2454	3.1518	0.8572
Education	−0.1532	0.03206	−0.4600	0.1578	−0.3200	0.06607
Union	−1.9960	0.1701	−2.3602	0.2567	−2.1940	0.1860
Female	−1.3498	0.2642	0.6957	1.0650	−0.2378	0.4679

variables in the model.) If the number of weeks worked and the accepted wage offer are determined jointly, then ln *Wage* 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

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

Table 12.1 presents the three sets of estimates. The OLS results 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 estimate developed in Section 12.3.2. As might be expected, the log wage coefficient becomes considerably larger. The other coefficients are, perhaps, contradictory. One have might 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.

~~8.3.9~~ 8.4.1 ~~12.4~~ THE HAUSMAN AND WU SPECIFICATION TESTS ~~AND AN APPLICATION TO INSTRUMENTAL VARIABLE ESTIMATION~~ (LS)

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

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

To compare the two matrices in the brackets, we can compare their inverses. The inverse of the first is $\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{X}'(\mathbf{I} - \mathbf{M}_Z)\mathbf{X} = \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{M}_Z\mathbf{X}$. Because $\mathbf{M}_Z$ is a nonnegative definite matrix, it follows that $\mathbf{X}'\mathbf{M}_Z\mathbf{X}$ is also. So, $\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ equals $\mathbf{X}'\mathbf{X}$ minus a nonnegative definite matrix. Because $\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ is smaller, in the matrix sense, than $\mathbf{X}'\mathbf{X}$, its inverse is larger. Under the hypothesis, the asymptotic covariance

8-22

matrix of the LS estimator is never larger than that of the IV estimator, and it will actually be smaller unless all the columns of $\mathbf{X}$ are perfectly predicted by regressions on $\mathbf{Z}$. Thus, we have established that if $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$—that is, if LS is consistent—then it is a preferred estimator. (Of course, we knew that from all our earlier results on the virtues of least squares.)

Our interest in the difference between these two estimators goes beyond the question of efficiency. The null hypothesis of interest will usually 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 the normal equations produce $(1/n)\mathbf{X}'\mathbf{e} = \mathbf{0}$. In a seminal paper, Hausman (1978) suggested an alternative testing strategy. [Earlier work by Wu (1973) and Durbin (1954) produced what turns out to be the same test.] 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 $\mathbf{d} \neq \mathbf{0}$. Using a strategy we have used at various points before, we might test this hypothesis with a Wald statistic,

$$H = \mathbf{d}'\{\text{Est. Asy. Var}[\mathbf{d}]\}^{-1}\mathbf{d}.$$

The asymptotic covariance matrix we need for the test is

$$\begin{aligned}\text{Asy. Var}[\mathbf{b}_{\text{IV}} - \mathbf{b}_{\text{LS}}] &= \text{Asy. Var}[\mathbf{b}_{\text{IV}}] + \text{Asy. Var}[\mathbf{b}_{\text{LS}}] \\ &\quad - \text{Asy. Cov}[\mathbf{b}_{\text{IV}}, \mathbf{b}_{\text{LS}}] - \text{Asy. Cov}[\mathbf{b}_{\text{LS}}, \mathbf{b}_{\text{IV}}].\end{aligned}$$

At this point, the test is straightforward, save for the considerable complication that we do not have an expression for the covariance term. Hausman gives a fundamental result that allows us to proceed. Paraphrased slightly,

the covariance between an efficient estimator, $\mathbf{b}_E$, of a parameter vector, $\boldsymbol{\beta}$, and its difference from an inefficient estimator, $\mathbf{b}_I$, of the same parameter vector, $\mathbf{b}_E - \mathbf{b}_I$, is zero.

For our case, $\mathbf{b}_E$ is $\mathbf{b}_{\text{LS}}$ and $\mathbf{b}_I$ is $\mathbf{b}_{\text{IV}}$. By Hausman's result we have

$$\text{Cov}[\mathbf{b}_E, \mathbf{b}_E - \mathbf{b}_I] = \text{Var}[\mathbf{b}_E] - \text{Cov}[\mathbf{b}_E, \mathbf{b}_I] = \mathbf{0}$$

or

$$\text{Cov}[\mathbf{b}_E, \mathbf{b}_I] = \text{Var}[\mathbf{b}_E],$$

so,

$$\text{Asy. Var}[\mathbf{b}_{\text{IV}} - \mathbf{b}_{\text{LS}}] = \text{Asy. Var}[\mathbf{b}_{\text{IV}}] - \text{Asy. Var}[\mathbf{b}_{\text{LS}}].$$

Inserting this useful result into our Wald statistic and reverting to our empirical estimates of these quantities, we have

$$H = (\mathbf{b}_{\text{IV}} - \mathbf{b}_{\text{LS}})'\{\text{Est. Asy. Var}[\mathbf{b}_{\text{IV}}] - \text{Est. Asy. Var}[\mathbf{b}_{\text{LS}}]\}^{-1}(\mathbf{b}_{\text{IV}} - \mathbf{b}_{\text{LS}}).$$

Under the null hypothesis, we are using two different, but consistent, estimators of σ^2. If we use s^2 as the common estimator, then the statistic will be

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

8-23

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. But that method will usually be incorrect, and worse yet, *unless* **X** *and* **Z** *have no variables in common, the rank of the matrix in this statistic is less than K, and the ordinary inverse will not even exist.* In most cases, at least some of the variables in **X** will also appear in **Z**. (In almost any application, **X** and **Z** will both contain the constant term.) That is, some of the variables in **X** are known to be uncorrelated with the disturbances. For example, the usual case will involve a single variable that is thought to be problematic or that is measured with error. In this case, our hypothesis, $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$, does not really involve all K variables, because a subset of the elements in this vector, say, K_0, are known to be zero. As such, the quadratic form in the Wald test is being used to test only $K^* = K - K_0$ hypotheses. It is easy (and useful) to show that, in fact, H is a rank K^* quadratic form. Since $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ is an idempotent matrix, $(\hat{\mathbf{X}}'\hat{\mathbf{X}}) = \hat{\mathbf{X}}'\mathbf{X}$. Using this result and expanding $\mathbf{d}$, we find

$$\begin{aligned}\mathbf{d} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}[\hat{\mathbf{X}}'\mathbf{y} - (\hat{\mathbf{X}}'\hat{\mathbf{X}})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]\\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'(\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y})\\ &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{e},\end{aligned}$$

where $\mathbf{e}$ is the vector of least squares residuals. Recall that K_0 of the columns in $\hat{\mathbf{X}}$ are the original variables in $\mathbf{X}$. Suppose that these variables are the first K_0. Thus, the first K_0 rows of $\hat{\mathbf{X}}'\mathbf{e}$ are the same as the first K_0 rows of $\mathbf{X}'\mathbf{e}$, which are, of course $\mathbf{0}$. (This statement does not mean that the first K_0 elements of $\mathbf{d}$ are zero.) So, we can write $\mathbf{d}$ as

$$\mathbf{d} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\begin{bmatrix}\mathbf{0}\\ \hat{\mathbf{X}}^{*\prime}\mathbf{e}\end{bmatrix} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\begin{bmatrix}\mathbf{0}\\ \mathbf{q}^*\end{bmatrix},$$

where $\mathbf{X}^*$ is the K^* variables in $\mathbf{x}$ that are not in $\mathbf{z}$.

Finally, denote the entire matrix in H by $\mathbf{W}$. (Because that ordinary inverse may not exist, this matrix will have to be a generalized inverse; see Section A.6.12.) Then, denoting the whole matrix product by $\mathbf{P}$, we obtain

$$H = [\mathbf{0}'\ \mathbf{q}^{*\prime}](\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\mathbf{W}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\begin{bmatrix}\mathbf{0}\\ \mathbf{q}^*\end{bmatrix} = [\mathbf{0}'\ \mathbf{q}^{*\prime}]\mathbf{P}\begin{bmatrix}\mathbf{0}\\ \mathbf{q}^*\end{bmatrix} = \mathbf{q}^{*\prime}\mathbf{P}_{**}\mathbf{q}^*,$$

where $\mathbf{P}_{**}$ is the lower right $K^* \times K^*$ submatrix of $\mathbf{P}$. We now have the end result. Algebraically, H is actually a quadratic form in a K^* vector, so K^* is the degrees of freedom for the test.

The preceding Wald test requires a generalized inverse [see Hausman and Taylor (1981)], so it is going to be a bit cumbersome. In fact, one need not actually approach the test in this form, and it can be carried out with any regression program. The alternative **variable addition test** approach devised by Wu (1973) is simpler. An F 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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \hat{\mathbf{X}}^*\boldsymbol{\gamma} + \boldsymbol{\varepsilon}^*, \qquad \text{8-11 } (~~12\text{-}10~~)$$

✓ ✓ KT

8-24

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

Example ~~12.3~~ 8.6 (Continued) Labor Supply Model

For the labor supply equation estimated in Example ~~12.3,~~ 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

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

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.

Although most of the preceding results are specific to this test of correlation between some of the columns of $\mathbf{X}$ and the disturbances, $\boldsymbol{\varepsilon}$, the Hausman test is general. To reiterate, when we have a situation in which we have a pair of estimators, $\hat{\boldsymbol{\theta}}_E$ and $\hat{\boldsymbol{\theta}}_I$, such that under H_0: $\hat{\boldsymbol{\theta}}_E$ and $\hat{\boldsymbol{\theta}}_I$ are both consistent and $\hat{\boldsymbol{\theta}}_E$ is efficient relative to $\hat{\boldsymbol{\theta}}_I$, while under H_1: $\hat{\boldsymbol{\theta}}_I$ remains consistent while $\hat{\boldsymbol{\theta}}_E$ is inconsistent, then we can form a test of the hypothesis by referring the **Hausman statistic,** KT

$$H = (\hat{\boldsymbol{\theta}}_I - \hat{\boldsymbol{\theta}}_E)' \left\{ \text{Est. Asy. Var}[\hat{\boldsymbol{\theta}}_I] - \text{Est. Asy. Var}[\hat{\boldsymbol{\theta}}_E] \right\}^{-1} (\hat{\boldsymbol{\theta}}_I - \hat{\boldsymbol{\theta}}_E) \xrightarrow{d} \chi^2[J],$$

to the appropriate critical value for the chi-squared distribution. The appropriate degrees of freedom for the test, J, will depend on the context. Moreover, some sort of generalized inverse matrix may be needed for the matrix, although in at least one ~~13~~ ~~12~~ 11 common case, the random effects regression model (see Chapter 9), the appropriate approach is to extract some rows and columns from the matrix instead. The short rank issue is not general. Many applications can be handled directly in this form with a full rank quadratic form. Moreover, the Wu approach *is* specific to this application. Another applications that we will consider, the independence from irrelevant alternatives test for the multinomial logit model, does not lend itself to the regression approach and is typically handled using the Wald statistic and the full rank quadratic form. As a final note, observe that the short rank of the matrix in the Wald statistic is an algebraic result. The failure of the matrix in the Wald statistic to be positive definite, however, is sometimes a finite sample problem that is not part of the model structure. In such a case, forcing a solution by using a generalized inverse may be misleading. Hausman suggests that in this instance, the appropriate conclusion might be simply to take the result as zero and, by implication, not reject the null hypothesis.

Example ~~12.4~~ 8.7 Hausman Test for a Consumption Function

Quarterly data for 1950.1 to 2000.4 on a number of macroeconomic variables appear in Appendix Table F5.1. A consumption function of the form $C_t = \alpha + \beta Y_t + \varepsilon_t$ is estimated using the 203 observations on aggregate U.S. real consumption and real disposable personal income, omitting the first. ~~In Example 12.1,~~ This model is ~~suggested as~~ a candidate for the

F5.2

8-25

possibility of bias due to correlation between Y_t and ε_t. Consider instrumental variables estimation using Y_{t-1} and C_{t-1} as the instruments for Y_t, and, of course, the constant term is its own instrument. One observation is lost because of the lagged values, so the results are based on 203 quarterly observations. The Hausman statistic can be computed in two ways:

1. Use the Wald statistic for H with the Moore–Penrose generalized inverse. The common s^2 is the one computed by least squares under the null hypothesis of no correlation. With this computation, $H = 8.481$. There is $K^* = 1$ degree of freedom. The 95 percent critical value from the chi-squared table is 3.84. Therefore, we reject the null hypothesis of no correlation between Y_t and ε_t.
2. Using the Wu statistic based on (~~12-10~~ 8-11), we regress C_t on a constant, Y_t, and the predicted value in a regression of Y_t on a constant, Y_{t-1} and C_{t-1}. The t ratio on the prediction is 2.968, so the F statistic with 1 and 201 degrees of freedom is 8.809. The critical value for this F distribution is 4.15, so, again, the null hypothesis is rejected.

3.888 200

12.5 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.

12.5.1 LEAST SQUARES ATTENUATION

In this section, we examine some of the received results on regression analysis with badly measured data. The general assessment of the problem is not particularly optimistic. 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.[2] 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,

$$y^* = \beta x^* + \varepsilon, \tag{12-11}$$

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

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