Chapter 7. Endogeneity and Instrumental Variables*

This chapter covers endogeniety and the two-stage least squares estimation. Related materials can be found in Chapter 3 of Hayashi (2000), Chapter 4 of Cameron and Trivedi (2005), Chapter 9 of Hansen (2007), and Chapter 5 of Wooldrige (2010).

1 Endogeneity

In linear regression,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i, \tag{1}$$

where y_i is the dependent variable, $\mathbf{x}_i \in \mathbb{R}^k$ is a vector of explanatory variables, $\boldsymbol{\beta}$ contains the unknown coefficients, u_i is the unobservable component of y_i, and $E[u_i|\mathbf{x}_i] = 0$. A regression is designed to carry out statistical inferences on causal effects of $\mathbf{x}_i$ on y_i. But in practice, it often happens that $\mathbf{x}_i$ and u_i are correlated. When $E[\mathbf{x}_i u_i] \neq \mathbf{0}$, there is *endogeneity*. In this case, the LSE will be asymptotically biased. The analysis of data with endogenous regressors is arguably the main contribution of econometrics to statistical science. There are five commonly encountered situations where endogeneity exists.

(i) Simultaneous causality. For example, does computer usage increase the income? Do Cigarette taxes reduce smoking? Does putting criminals in jail reduce crime? Example 1 below shows the simultaneous causality induced by a system of equations. Solutions to this problem include using instrumental variables (IVs),[1] and designing and implementing a randomizing controlled experiment in which the reverse causality channel is nullified (see references cited in the Introduction). The first solution will be discussed in this chapter.

(ii) Omitted variables. For example, in the model on returns to schooling, ability is an important variable that is correlated to years of education, but is not observable so is included in the error term. Solutions to this problem include using IVs, using panel data (see, Chamberlain (1984), Arellano and Honoré (2001) and Hsiao (2003) for an introduction to panel data analysis), and using randomizing controlled experiments.

(iii) Errors in variables. This term refers to the phenomenon that an otherwise exogenous regressor becomes endogenous when measured with error. For example, in the returns-to-schooling

*Email: pingyu@hku.hk

[1] See Stock and Trebbi (2003) for who invented instrumental variable regression.

model, the records for years of education are fraught with errors owing to lack of recall, typographical mistakes, or other reasons. The basic solution to this problem is to use IVs (e.g., exogenous determinants of the error ridden explanatory variables, or multiple indicators of the same outcome). See Bound et al. (2001) for an introduction to measurement errors in survey data.

(iv) Sample selection. For example, in the analysis of returns to schooling, only wages for employed workers are available, but we want to know the effect of education for the general population. We will discuss how to handle such an endogeneity problem in Chapter 9.

(v) Functional form misspecification. $E[y|\mathbf{x}]$ may not be linear in $\mathbf{x}$. This problem can be handled by nonparametric methods. See related chapters in Handbook of Econometrics, Härdle and Linton (1994) (or its extended version Härdle (1990)), Chen (2007) and Ichimura and Todd (2007), for an introduction.

Example 1 (Simultaneous Causality) *Wright (1928) considered estimating the elasticity of butter demand, which is critical in the policy decision on the tariff of butter. In the economic language, he considered a linear Marshallian stochastic demand/supply system. Define* $p_i = \ln P_i$ *and* $q_i = \ln Q_i$*, and the demand equation is*

$$q_i = \alpha_0 + \alpha_1 p_i + u_i, \tag{2}$$

where u_i *represents other factors besides price that affect demand, such as income and consumer taste. But the supply equation is in the same form as (2):*

$$q_i = \beta_0 + \beta_1 p_i + v_i, \tag{3}$$

where v_i *represents the factors that affect supply, such as weather conditions (see, e.g., Angrist et al. (2000)), factor prices, and union status. So* p_i *and* q_i *are determined "within" the model, and they are endogenous. Rigorously, note that*

$$\begin{aligned} p_i &= \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1}, \\ q_i &= \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 v_i - \beta_1 u_i}{\alpha_1 - \beta_1}, \end{aligned}$$

by solving two simultaneous equations (2) and (3). Suppose $Cov(u_i, v_i) = 0$*, then*

$$Cov(p_i, u_i) = -\frac{Var(u_i)}{\alpha_1 - \beta_1}, Cov(p_i, v_i) = \frac{Var(v_i)}{\alpha_1 - \beta_1},$$

which are not zero. If $\alpha_1 < 0$ *and* $\beta_1 > 0$*, then* $Cov(p_i, u_i) > 0$ *and* $Cov(p_i, v_i) < 0$*. This is intuitively correct: if a demand (supply) shifter shifts the demand (supply) curve right or* $u_i > 0$ *(*$v_i > 0$*), the price increases (decreases).*

If we regress q_i on p_i, then the slope estimator converges to

$$\frac{Cov(p_i,q_i)}{Var(p_i)} = \alpha_1 + \frac{Cov(p_i,u_i)}{Var(p_i)} = \beta_1 + \frac{Cov(p_i,v_i)}{Var(p_i)}$$
$$\overset{why?}{=} \frac{\alpha_1 Var(v_i)+\beta_1 Var(u_i)}{Var(v_i)+Var(u_i)} \in (\alpha_1, \beta_1),$$

so the LSE is neither α_1 nor β_1, but a weighted average of them. Such a bias is called the simultaneous equations bias. *The LSE cannot consistently estimate α_1 or β_1 because both curves are shifted by other factors besides price, and we cannot tell from data whether the change in price and quantity is due to a demand shift or a supply shift. If $u_i = 0$ (that is, the demand curve stays still), then the equilibrium prices and quantities will trace out the demand curve and the LSE is consistent to α_1. Figure 1 illustrates the discussion above intuitively.*

From the discussion above, p_i has one part correlated with u_i $\left(-\frac{u_i}{\alpha_1-\beta_1}\right)$ and one part uncorrelated with u_i $\left(\frac{v_i}{\alpha_1-\beta_1}\right)$. If we can isolate the second part, then we can focus on those variations in p_i that are uncorrelated with u_i and disregard the variations in p_i that bias the LSE. Take a supply shifter z_i (e.g., weather), which can be considered to be uncorrelated with the demand shifter u_i such as consumer's tastes; then

$$Cov(z_i, u_i) = 0, \text{ and } Cov(z_i, p_i) \neq 0.$$

So

$$Cov(z_i, q_i) = \alpha_1 \cdot Cov(z_i, p_i),$$

and

$$\alpha_1 = \frac{Cov(z_i, q_i)}{Cov(z_i, p_i)}.$$

A natural estimator is

$$\widehat{\alpha}_1 = \frac{\widehat{Cov}(z_i, q_i)}{\widehat{Cov}(z_i, p_i)},$$

which is the IV estimator. Another method to estimate α_1 as suggested above is to run regression

$$q_i = \alpha_0 + \alpha_1 \widehat{p}_i + \widetilde{u}_i,$$

where $\widehat{p}_i$ is the predicted value from the following regression:

$$p_i = \gamma_0 + \gamma_1 z_i + \eta_i,$$

and $\widetilde{u}_i = \alpha_1 (p_i - \widehat{p}_i) + u_i$. It is easy to show that $Cov(\widehat{p}_i, \widetilde{u}_i) = 0$, so the estimation is consistent. Such a procedure is called two-stage least squares *(2SLS) for an obvious reason. In this case, the IV estimator and the 2SLS estimator are numerically equivalent as shown in Exercise 9 below.* □

Example 2 (Omitted Variables) *Mundlak (1961) considered the production function estimation, where the error term includes factors that are observable to the economic agent under study but unobservable to the econometrician, and endogeneity arises when regressors are decisions made*

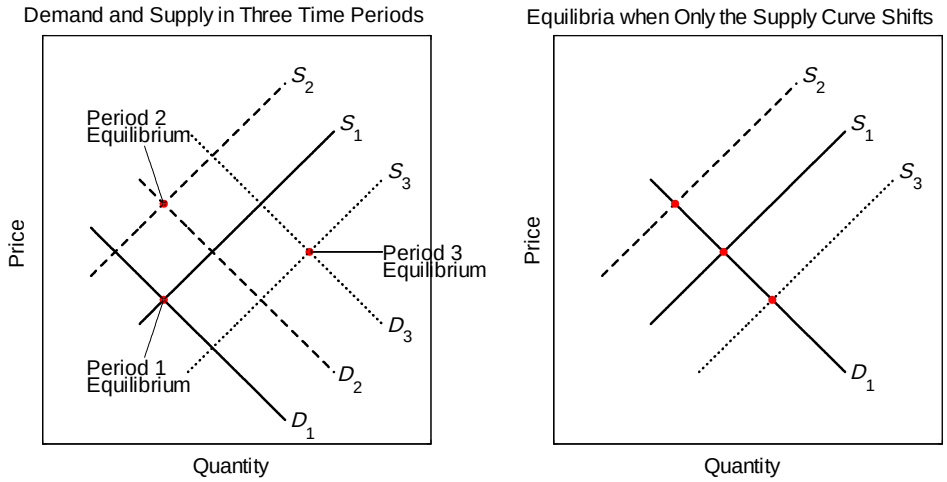

Figure 1: Endogeneity and Identification by Instrument Variables

by the agent on the basis of such factors.

Suppose that a farmer is producing a product using the Cobb-Douglas technology:

$$Q_i = A_i \cdot (L_i)^{\phi_1} \cdot \exp(\nu_i),\ 0 < \phi_1 < 1, \tag{4}$$

where Q_i is the output of the ith farm, L_i is a variable input (labor), A_i represents an input that is fixed over time (e.g., soil quality), and ν_i represents a stochastic input (e.g., rainfall) which is not under the farmer's control. We shall assume that the farmer knows the product price p and input price w, which do not depend on his decisions, and that he knows A_i but econometricians do not. The factor input decision is made before knowing ν_i, and so L_i is chosen to maximize expected profits. The factor demand equation is

$$L_i = \left(\frac{w}{p}\right)^{\frac{1}{\phi_1 - 1}} (A_i B \phi_1)^{\frac{1}{1-\phi_1}}, \tag{5}$$

so a better farm induces more labors on it. We assume that (A_i, ν_i) is i.i.d. over farms, and A_i is independent of ν_i for each i. Therefore, $B = E[\exp(\nu_i)]$ is the same for all i, and the level of output the farm expects when it chooses L_i is $A_i \cdot (L_i)^{\phi_1} \cdot B$.

Taking logarithm on both sides of (4), we have a log-linear production function:

$$\log Q_i = \log A_i + \phi_1 \cdot \log(L_i) + \nu_i,$$

where $\log A_i$ is an omitted variable. Equivalently, each farm has a different intercept. The LSE of ϕ_1 will converge to $\frac{Cov(\log Q_i, \log(L_i))}{Var(\log(L_i))} = \phi_1 + \frac{Cov(\log A_i, \log(L_i))}{Var(\log(L_i))}$, which is not ϕ_1 since there is correlation between $\log A_i$ and $\log(L_i)$ as shown in (5). Figure 2 shows the effect of $\log A_i$ on ϕ_1 by drawing $E[\log Q | \log L, \log A]$ for two farms. In Figure 2, the OLS regression line passes through points AB

with slope $\frac{\log Q_1 - \log Q_2}{\log L_1 - \log L_2}$*, but the true* ϕ_1 *is* $\frac{D-C}{\log L_1 - \log L_2}$*. Their difference is* $\frac{A-D}{\log L_1 - \log L_2} = \frac{\log A_1 - \log A_2}{\log L_1 - \log L_2}$*, which is the bias introduced by the endogeneity of* $\log A_i$*.*

Rigorously, let $u_i = \log(A_i) - E[\log(A_i)]$*, and* $\phi_0 = E[\log(A_i)]$*; then* $E[u_i] = 0$ *and* $A_i = \exp(\phi_0 + u_i)$*. (4) and (5) can be written as*

$$\log Q_i = \phi_0 + \phi_1 \cdot \log(L_i) + \nu_i + u_i, \tag{6}$$

$$\log L_i = \beta_0 + \frac{1}{1-\phi_1} u_i, \tag{7}$$

where $\beta_0 = \frac{1}{1-\phi_1}\left(\phi_0 + \log(B\phi_1) - \log\left(\frac{w}{p}\right)\right)$ *is a constant for all farms. Now, it is obvious that* $\log L_i$ *is correlated with* $(\nu_i + u_i)$*. Thus, the LSE of* ϕ_1 *in the estimation of log-linear production function confounds the contribution of* u_i *with the contribution of labor. Actually,*

$$\widehat{\phi}_{1,OLS} \xrightarrow{p} 1,$$

because substituting (7) into (6), we get

$$\log Q_i = \phi_0 - (1-\phi_1)\beta_0 + \mathbf{1} \cdot \log(L_i) + \nu_i.$$

The lesson from this example is that a variable chosen by the agent taking into account some information unobservable to the econometrician can induce endogeneity. □

Exercise 1 *Suppose the farmer could observe* ν_i *as well as* A_i *before deciding on labor input, how does the demand equation for labor (5) change? Show that* $\log Q_i$ *and* $\log(L_i)$ *are perfectly correlated.*

Example 3 (Errors in Variables) *The cross-section version of M. Friedman's (1957) Permanent Income Hypothesis can be formulated as an errors-in-variables problem. The hypothesis states that "permanent consumption"* C_i^* *for household* i *is proportional to "permanent income"* Y_i^**:*

$$C_i^* = kY_i^* \text{ with } 0 < k < 1.$$

Assume both measured consumption C_i *and income* Y_i *are contaminated by measurement error:*

$$C_i = C_i^* + c_i \text{ and } Y_i = Y_i^* + y_i,$$

where c_i *and* y_i *are independent of* C_i^* *and* Y_i^* *and are independent of each other; then*

$$C_i = kY_i + u_i \text{ with } u_i = c_i - ky_i. \tag{8}$$

It is easy to see that $E[Y_i u_i] = -kE\left[y_i^2\right] < 0$*, so the LSE of* k *converges to* $\frac{E[Y_i C_i]}{E[Y_i^2]} = \frac{kE\left[\left(Y_i^*\right)^2\right]}{E\left[\left(Y_i^*\right)^2\right] + E\left[y_i^2\right]} < k$*, that is, the OLS using measured data underestimates* k*. Taking expectation on both sides of (8),*

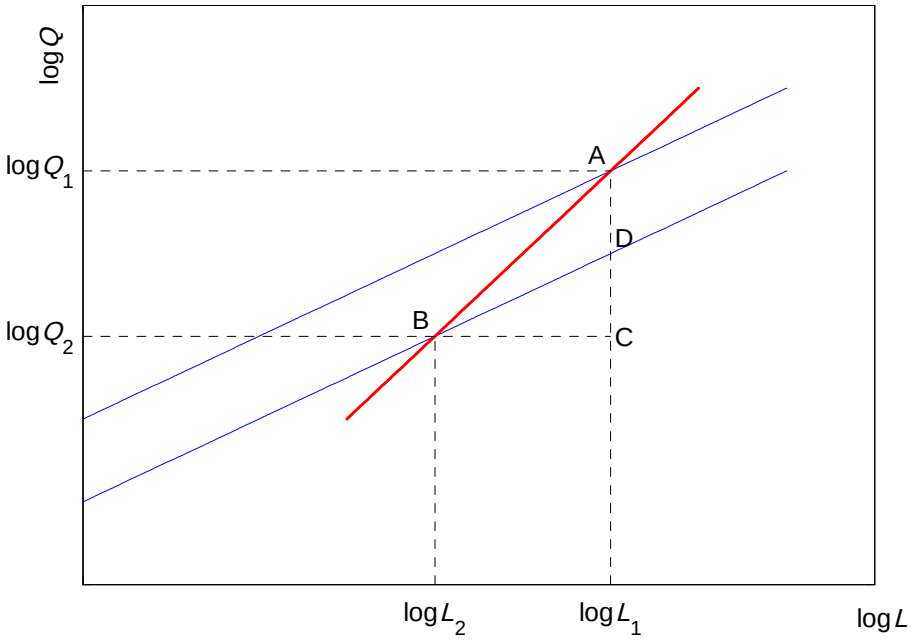

Figure 2: Effect of Soil Quality on Labor Input

we have $E[C_i] = kE[Y_i] + E[u_i]$. *So* $z = 1$ *is a valid IV if* $E[y_i] = E[c_i] = 0$ *and* $E[Y_i^*] = E[Y_i] \neq 0$.. *The IV estimation using* z *as the instrument is* $\frac{\overline{C}_i}{\overline{Y}_i}$, *which is how Friedman estimated* k.

Actually, measurement errors are embodied in regression analysis from the beginning. Galton (1889) analyzed the relationship between the height of sons and the height of fathers. Specifically, suppose the true model is

$$S_i = \alpha + \beta F_i + u_i, \tag{9}$$

where S_i *and* F_i *are the heights of sons and fathers, respectively. Even if* S_i *should perfectly match* F_i *(that is,* $\alpha_0 = 0$, $\beta_0 = 1$, *and* $u_i = 0$*), the OLS estimator would be smaller than* 1 *if there are environmental factors or measurement errors that affect* S_i. *Suppose* $S_i = F_i + f_i$, *where* f_i *is the environmental factor; then our regression becomes*

$$S_i = \alpha + \beta (S_i - f_i) + u_i = \alpha + \beta S_i + u_i - \beta f_i,$$

where the regressor S_i *should be interpreted as the mismeasured* F_i *in (9). The OLS estimator of* β *will converge to* $\frac{Cov(\alpha_0 + \beta_0 S_i + u_i - \beta_0 f_i, S_i)}{Var(S_i)} = \frac{Var(F_i)}{Var(F_i) + Var(f_i)} < 1$, *where* $\frac{Var(F_i)}{Var(F_i) + Var(f_i)} \equiv \rho$ *is called the* reliability coefficient *or* Heritability coefficient. *In Galton's analysis, this coefficient is about* $2/3$. *Similarly, the OLS estimator of* α *converges to* $(1 - \rho) E[F_i] \neq 0$ *(why?). The regression line and the true line intersect at* $E[F_i]$, *and both are shown in Figure 3. Galton wrote "the average regression of the offspring is a constant fraction of their respective mid-parental deviations" and termed this phenomenon as "regression towards mediocrity".*

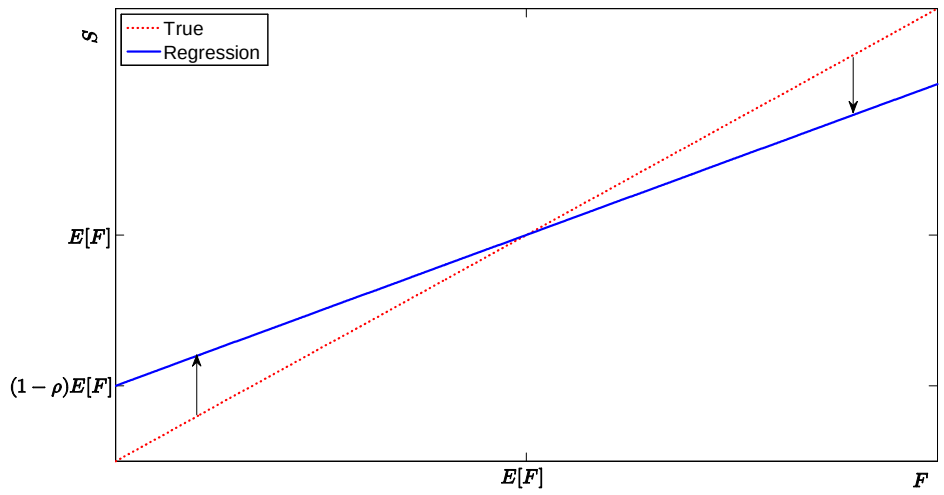

Figure 3: Relationship Between the Height of Sons and Fathers

Finally, note from Exercise 2 in Chapter 5 that measure errors in y_i will not affect the consistency of the LSE but may affect its asymptotic distribution. □

2 Instrumental Variables

We call (1) the *structural equation* or *primary equation*. In matrix notation, it can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{10}$$

Any solution to the problem of endogeneity requires additional information which we call *instrumental variables* (or simply *instruments*). The $l \times 1$ random vector $\mathbf{z}_i$ is an instrument for (1) if $E[\mathbf{z}_i u_i] = \mathbf{0}$. This condition cannot be tested in practice since u_i cannot be observed.

In a typical set-up, some regressors in $\mathbf{x}_i$ will be uncorrelated with u_i (for example, at least the intercept). Thus we make the partition

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{x}_{2i} \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix}, \tag{11}$$

where $E[\mathbf{x}_{1i} u_i] = \mathbf{0}$ yet $E[\mathbf{x}_{2i} u_i] \neq \mathbf{0}$. We call $\mathbf{x}_{1i}$ *exogenous* and $\mathbf{x}_{2i}$ *endogenous*. By the above definition, $\mathbf{x}_{1i}$ is an instrumental variable for (1), so should be included in $\mathbf{z}_i$, giving the partition

$$\mathbf{z}_i = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{z}_{2i} \end{pmatrix} \begin{matrix} k_1 \\ l_2 \end{matrix}, \tag{12}$$

where $\mathbf{x}_{1i} = \mathbf{z}_{1i}$ are the *included exogenous variables,* and $\mathbf{z}_{2i}$ are the *excluded exogenous variables.*

In other words, $\mathbf{z}_{2i}$ are variables which could be included in the equation for y_i (in the sense that they are uncorrelated with u_i) yet can be excluded, as they would have true zero coefficients in the equation which means that certain directions of causation are ruled out a priori.

The model is *just-identified* if $l = k$ (i.e., if $l_2 = k_2$) and *over-identified* if $l > k$ (i.e., if $l_2 = k_2$). We have noted that any solution to the problem of endogeneity requires instruments. This does not mean that valid instruments actually exist.

3 Reduced Form

The reduced form relationship between the variables or "regressors" $\mathbf{x}_i$ and the instruments $\mathbf{z}_i$ is found by linear projection. Let

$$\mathbf{\Gamma} = E\left[\mathbf{z}_i\mathbf{z}_i'\right]^{-1} E\left[\mathbf{z}_i\mathbf{x}_i'\right]$$

be the $l \times k$ matrix of coefficients from a projection of $\mathbf{x}_i$ on $\mathbf{z}_i$, and define

$$\mathbf{v}_i = \mathbf{x}_i - \mathbf{\Gamma}'\mathbf{z}_i$$

as the projection error. Then the reduced form linear relationship between $\mathbf{x}_i$ and $\mathbf{z}_i$ is the instrumental equation

$$\mathbf{x}_i = \mathbf{\Gamma}'\mathbf{z}_i + \mathbf{v}_i. \tag{13}$$

In matrix notation,

$$\mathbf{X} = \mathbf{Z\Gamma} + \mathbf{V}, \tag{14}$$

where $\mathbf{V}$ is a $n \times k$ matrix. By construction, $E\left[\mathbf{z}_i\mathbf{v}_i'\right] = \mathbf{0}$, so (13) is a projection and can be estimated by OLS:

$$\begin{aligned} \mathbf{X} &= \mathbf{Z}\widehat{\mathbf{\Gamma}} + \widehat{\mathbf{V}}, \\ \widehat{\mathbf{\Gamma}} &= \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\mathbf{Z}'\mathbf{X}\right). \end{aligned}$$

Substituting (14) into (10), we find

$$\mathbf{y} = \left(\mathbf{Z\Gamma} + \mathbf{V}\right)\boldsymbol{\beta} + \mathbf{u} = \mathbf{Z}\boldsymbol{\lambda} + \mathbf{e} \tag{15}$$

where $\boldsymbol{\lambda} = \mathbf{\Gamma}\boldsymbol{\beta}$ and $\mathbf{e} = \mathbf{V}\boldsymbol{\beta} + \mathbf{u}$. Observe that

$$E\left[\mathbf{z}e\right] = E\left[\mathbf{z}\mathbf{v}'\right]\boldsymbol{\beta} + E\left[\mathbf{z}u\right] = \mathbf{0}. \tag{16}$$

Thus (15) is a projection equation and may be estimated by OLS. This is

$$\begin{aligned} \mathbf{y} &= \mathbf{Z}\widehat{\boldsymbol{\lambda}} + \widehat{\mathbf{e}}, \\ \widehat{\boldsymbol{\lambda}} &= \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\mathbf{Z}'\mathbf{y}\right). \end{aligned}$$

The equation (15) is the reduced form for $\mathbf{y}$. (14) and (15) together are the reduced form equations for the system

$$\begin{aligned} \mathbf{y} &= \mathbf{Z}\boldsymbol{\lambda} + \mathbf{e}, \\ \mathbf{X} &= \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{V}. \end{aligned}$$

As we showed above, OLS yields the reduced-form estimates $\left(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\Gamma}}\right)$.

The system of equations

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \\ \mathbf{X} &= \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{V}, \end{aligned}$$

are called *triangular simultaneous equations* because the second part of equations do not depend on $\mathbf{y}$. This system of equations rules out full simultaneity and includes the same information as an "incomplete" linear system

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, E\left[\mathbf{Z}'\mathbf{u}\right] = \mathbf{0}.$$

However, in a nonlinear system, they are not equivalent in general.

Exercise 2 *(i) Show that* $E\left[\mathbf{v}_i u_i\right] \neq \mathbf{0}$. *(ii) Suppose* $k = k_2 = l = 1$, *and all variables are demeaned. Can the correlation between* u_i *and* v_i *be* 1*?*

Exercise 3 *Suppose* $y = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ *and* $\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{V}$, *where* $E[\mathbf{u}|\mathbf{Z}] = \mathbf{0}$ *and* $E[\mathbf{V}|\mathbf{Z}] = \mathbf{0}$ *but* $E[\mathbf{V}'\mathbf{u}] \neq \mathbf{0}$. *Derive the plim of* $\widehat{\boldsymbol{\beta}}_{OLS}$. *When* $\boldsymbol{\Gamma} = \mathbf{0}$, *what will plim*$\left(\widehat{\boldsymbol{\beta}}_{OLS}\right)$ *degenerate to?*

Exercise 4 *In the reduced form between the regressors* $\mathbf{x}_i$ *and instruments* $\mathbf{z}_i$ *(13), the parameter* $\boldsymbol{\Gamma}$ *is defined by the population moment condition*

$$E\left[\mathbf{z}_i \mathbf{v}_i'\right] = \mathbf{0}.$$

Show that the MoM estimator for $\boldsymbol{\Gamma}$ *is* $\widehat{\boldsymbol{\Gamma}} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}$.

4 Identification

The structural parameter $\boldsymbol{\beta}$ in triangular simultaneous equations relates to $(\boldsymbol{\lambda}, \boldsymbol{\Gamma})$ by $\boldsymbol{\lambda} = \boldsymbol{\Gamma}\boldsymbol{\beta}$. This relation can be derived directly by using the orthogonal condition $E\left[\mathbf{z}_i\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)\right] = \mathbf{0}$ which is equivalent to

$$E\left[\mathbf{z}_i y_i\right] = E\left[\mathbf{z}_i \mathbf{x}_i'\right]\boldsymbol{\beta}. \tag{17}$$

Multiplying each side by an invertible matrix $E\left[\mathbf{z}_i\mathbf{z}_i'\right]^{-1}$, we have $\boldsymbol{\lambda} = \boldsymbol{\Gamma}\boldsymbol{\beta}$. The parameter is identified, meaning that it can be uniquely recovered from the reduced form, if the *rank condition*[2]

$$\text{rank}\left(\boldsymbol{\Gamma}\right) = k \tag{18}$$

holds. This condition can be tested in practice (think about the case $k = 1$). If $\text{rank}(E\left[\mathbf{z}_i\mathbf{z}_i'\right]) = l$ (this is trivial), and $\text{rank}(E\left[\mathbf{z}_i\mathbf{x}_i'\right]) = k$ (this is crucial), this condition is satisfied. Assume that (18) holds. If $l = k$, then $\boldsymbol{\beta} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\lambda}$. If $l > k$, then for any $\mathbf{A} > 0$, $\boldsymbol{\beta} = \left(\boldsymbol{\Gamma}'\mathbf{A}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\mathbf{A}\boldsymbol{\lambda}$. If (18) is not satisfied, then $\boldsymbol{\beta}$ cannot be uniquely recovered from $(\boldsymbol{\lambda}, \boldsymbol{\Gamma})$. Note that a necessary (although not sufficient) condition for (18) is the *order condition* $l \geq k$. Since $\mathbf{Z}$ and $\mathbf{X}$ have the common variables $\mathbf{X}_1$, we can rewrite some of the expressions. Using (11) and (12) to make the matrix partitions $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2]$ and $\mathbf{X} = [\mathbf{Z}_1, \mathbf{X}_2]$, we can partition $\boldsymbol{\Gamma}$ as

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \boldsymbol{\Gamma}_{12} \\ \mathbf{0} & \boldsymbol{\Gamma}_{22} \end{pmatrix} \begin{matrix} k_1 \\ l_2 \end{matrix} .$$
$$\begin{matrix} k_1 & k_2 \end{matrix}$$

(14) can be rewritten as

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{Z}_1 \\ \mathbf{X}_2 &= \mathbf{Z}_1\boldsymbol{\Gamma}_{12} + \mathbf{Z}_2\boldsymbol{\Gamma}_{22} + \mathbf{V}_2. \end{aligned}$$

$\boldsymbol{\beta}$ is identified if $\text{rank}(\boldsymbol{\Gamma}) = k$, which is true if and only if $\text{rank}(\boldsymbol{\Gamma}_{22}) = k_2$ (by the upper-diagonal structure of $\boldsymbol{\Gamma}$). Thus the key to identification of the model rests on the $l_2 \times k_2$ matrix $\boldsymbol{\Gamma}_{22}$.

Exercise 5 *In the structural model*

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \\ \mathbf{X} &= \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{V}, \end{aligned}$$

with $\boldsymbol{\Gamma}$ $l \times k$, $l \geq k$, we claim that $\boldsymbol{\beta}$ is identified (can be recovered from the reduced form) if rank$(\boldsymbol{\Gamma}) = k$. Explain why this is true. That is, show that if rank$(\boldsymbol{\Gamma}) < k$ then $\boldsymbol{\beta}$ cannot be identified.

Example 4 (What Variable Is Qualified to Be An IV?) *It is often suggested to select an instrumental variable that is*

$$\textit{(i) uncorrelated with } u\textit{; (ii) correlated with endogenous variables.} \tag{19}$$

(i) is the instrument exogeneity condition, which says that the instruments can correlate with the dependent variable only indirectly through the endogenous variable. (ii) intends to repeat the in-

[2] The precise condition should be $\text{rank}([\lambda, \Gamma])$ =$\text{rank}(\Gamma) = k$. The first equality guarantees the existence, and the second guarantees the uniqueness. Usually, the existence is assumed, so we only write out the second equality.

strument relevance condition which says that $\mathbf{X}_1$ *and the predicted value of* $\mathbf{X}_2$ *from the regression of* $\mathbf{X}_2$ *on* $\mathbf{Z}$ *and* $\mathbf{X}_1$ *are not perfectly multicollinear; in other words, there must be "enough" extra variation in* $\widehat{\mathbf{x}}_2$ *that can not be explained by* $\mathbf{x}_1$*. Such a condition is required in the second stage regression. Scrutinizing (19) is important to practioners.*

Check the following example with only one endogenous variable:

$$\begin{aligned} y &= x_1\beta_1 + x_2\beta_2 + u, \\ E[x_1u] &= 0,\ E[x_2u] \neq 0,\ Cov(x_1, x_2) \neq 0. \end{aligned}$$

One may suggest the following instrument for x_2*, say,* $z = x_1 + \varepsilon$*, where* ε *is some computer-generated random variable independent of the system. Now,* $E[zu] = 0$ *and* $Cov(z, x_2) = Cov(x_1, x_2) \neq 0$*. It seems that* z *is a valid instrument, but intuition tells us that it is NOT, since it includes the same useful information as* x_1*. What is missing? We know the right conditions for a random variable to be a valid instrument are*

$$\begin{aligned} E[zu] &= 0, \\ x_2 &= x_1\gamma_1 + z\gamma_2 + v \text{ with } \gamma_2 \neq 0. \end{aligned} \tag{20}$$

In this example, $x_2 = x_1\gamma_1 + z\gamma_2 + v = x_1(\gamma_1 + \gamma_2) + (\varepsilon\gamma_2 + v)$*,* γ_2 *is not identified!*

The arguments above indicate that (19) is not sufficient. Is it necessary? The answer is still NO! The question can be formulated as follows: in (20), can we find some z *such that*

$$\gamma_2 \neq 0 \text{ but } Cov(z, x_2) = 0?$$

Observe that $Cov(z, x_2) = Cov(z, x_1\gamma_1 + z\gamma_2 + v) = Cov(z, x_1)\gamma_1 + Var(z)\gamma_2$*, so if* $\frac{Cov(z,x_1)}{Var(z)} = -\frac{\gamma_2}{\gamma_1}$*, this could happen. That is, although* z *is not correlated with* x_2*, it is correlated with* x_1*, and* x_1 *is correlated with* x_2*. In mathematical language,* $Cov(z, x_1) \neq 0$*,* $\gamma_1 \neq 0$*. In such a case,* z *is related to* x_2 *only indirectly through* x_1*. If we assume* $Cov(z, x_1) = 0$*, or* $\gamma_1 = 0$*, then the assumption* $Cov(z, x_2) \neq 0$ *is the right condition for* z *to be a valid instrument. So the right condition should be that* z *is partially correlated with* x_2 *after netting out the effect of* x_1*.*

In general, a necessary condition for a set of qualified instruments is that at least one instrument appears in each of the first-stage regression. Here "appear" means the coefficient of the variable is not zero. When $k = \ell$*, each instrument must appear in at least one endogenous regression.*

Given the cautions above, how to select instruments? Generally speaking, good instruments are not selected based on mathematics, but based on economic theory. For example, Angrist and Krueger (1991) propose using quarter of birth as an IV for education in the analysis of returns to schooling because of a mechanical interaction between compulsory school attendance laws and age at school entry; Card (1995) uses college proximity[3] *as an instrument to identify the returns to schooling, noting that living close to a college during childhood may induce some children to go to*

[3] Parental education is another popular IV to identify the returns to schooling.

college but is unlikely to directly affect the wages earned in their adulthood; Acemoglu et al. (2001) use the mortality rates (of soldiers, bishops, and sailors) as an IV to estimate the effect of property rights and institutions on economic development. □

Exercise 6 *Consider the linear demand and supply system:*

$$\begin{aligned} \textit{Demand:} \qquad & q_i = \alpha_0 + \alpha_1 p_i + \alpha_2 y_i + u_i, \\ \textit{Supply:} \qquad & q_i = \beta_0 + \beta_1 p_i + \beta_2 w_i + v_i. \end{aligned}$$

where income (y) *and wage* (w) *are determined outside the market. In this model, are the parameters identified?*

5 Estimation: Two-Stage Least Squares

If $l = k$, then the moment condition is $E\left[\mathbf{z}_i\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)\right] = \mathbf{0}$, and the corresponding IV estimator is a MoM estimator:

$$\widehat{\boldsymbol{\beta}}_{IV} = \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\mathbf{Z}'\mathbf{y}\right).$$

Another interpretation stems from the fact that since $\boldsymbol{\beta} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\lambda}$, we can construct the *Indirect Least Squares* (ILS) estimator of Haavelmo (1943):

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\lambda}} = \left(\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{y}\right) = \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\mathbf{Z}'\mathbf{y}\right).$$

Exercise 7 *In the linear model,*

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + u_i, \\ E[u_i|\mathbf{x}_i] &= 0. \end{aligned}$$

Suppose $\sigma_i^2 = E[u_i^2|\mathbf{x}_i]$ *is known. Show that the GLS estimator of* $\boldsymbol{\beta}$ *with the weight matrix* $diag\left\{\sigma_1^{-2}, \cdots, \sigma_n^{-2}\right\}$ *can be written as an IV estimator using some instrument* $\mathbf{z}_i$*. (Find an expression for* $\mathbf{z}_i$*.)*

Exercise 8 *Suppose* $y = x\beta + u$*,* $x = \varepsilon + \lambda^{-1}u$*, and* $z = v + \gamma\varepsilon$*, where* ε, u *and* v *are independent. Find the probability limits of* $\widehat{\beta}_{OLS}$ *and* $\widehat{\beta}_{IV}$*. Show that if* $\gamma = 0$*,* $\frac{1}{n}\sum_{i=1}^n v_i\varepsilon_i = 0$*, and* σ_u^2 *is large, the two probability limits are the same.*

When $l > k$, the two-stage least squares (2SLS) estimator can be used. It was originally proposed by Theil (1953) and Basmann (1957), and is the classic estimator for linear equations with instruments. Given any k instruments out of $\mathbf{z}$ or its linear combinations can be used to identify $\boldsymbol{\beta}$, the 2SLS chooses those that are most highly (linearly) correlated with $\mathbf{x}$. Namely, it is the sample analog of the following implication of $E[\mathbf{z}u] = \mathbf{0}$:

$$\mathbf{0} = E\left[E^*\left[\mathbf{x}|\mathbf{z}\right]u\right] = E\left[\boldsymbol{\Gamma}'\mathbf{z}u\right] = E\left[\boldsymbol{\Gamma}'\mathbf{z}(y - \mathbf{x}'\boldsymbol{\beta})\right], \tag{21}$$

where $E^*[\mathbf{x}|\mathbf{z}]$ is the linear projection of $\mathbf{x}$ on $\mathbf{z}$. Replacing population expectations with sample averages in (21) yields

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \left(\widehat{\mathbf{X}}'\mathbf{X}\right)^{-1}\widehat{\mathbf{X}}'\mathbf{y},$$

where $\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\boldsymbol{\Gamma}} \equiv \mathbf{P}\mathbf{X}$ with $\widehat{\boldsymbol{\Gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})$ and $\mathbf{P} = \mathbf{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. In other words, the 2SLS estimator is an IV estimator with the IVs being $\widehat{\mathbf{x}}_i$.

Exercise 9 *Show that if $l = k$, then $\widehat{\boldsymbol{\beta}}_{2SLS} = \widehat{\boldsymbol{\beta}}_{IV}$, that is, no matter we use $\widehat{\mathbf{x}}_i$ or $\mathbf{z}_i$ as IVs, we get the same results.*

The source of the name "two-stage" is from Theil (1953)'s formulation of 2SLS. From (16),

$$\mathbf{0} = E\left[E^*[\mathbf{x}|\mathbf{z}](u + \mathbf{v}'\boldsymbol{\beta})\right] = E\left[\left(\boldsymbol{\Gamma}'\mathbf{z}\right)\left(y - \mathbf{z}'\boldsymbol{\Gamma}\boldsymbol{\beta}\right)\right],$$

i.e., $\boldsymbol{\beta}$ is the least squares regression coefficients of the regression of y on fitted values of $\boldsymbol{\Gamma}'\mathbf{z}$, so this method is often called the *fitted-value* method. The sample analogue is the following two-step procedure:

- First, regress $\mathbf{X}$ on $\mathbf{Z}$ to get $\widehat{\mathbf{X}}$.
- Second, regress $\mathbf{y}$ on $\widehat{\mathbf{X}}$ to get

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \left(\widehat{\mathbf{X}}'\widehat{\mathbf{X}}\right)^{-1}\widehat{\mathbf{X}}'\mathbf{y} = \left(\mathbf{X}'\mathbf{P}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{P}\mathbf{y}\right). \tag{22}$$

Exercise 10 *Show that $\widehat{\boldsymbol{\beta}}_{2SLS}$ satisfies*

$$\left(\overline{\mathbf{S}} - \widehat{\lambda}\mathbf{e}_1\mathbf{e}_1'\right)\begin{pmatrix} 1 \\ -\widehat{\boldsymbol{\beta}}_{2SLS} \end{pmatrix} = \mathbf{0},$$

where $\mathbf{e}_1 = (1, 0, \cdots, 0)'$ is the first $(k+1) \times 1$ unit vector, $\overline{\mathbf{S}} = (\mathbf{y}, \mathbf{X})'\mathbf{P}(\mathbf{y}, \mathbf{X})$, and $\widehat{\lambda} = (1, -\widehat{\boldsymbol{\beta}}_{2SLS}')\overline{\mathbf{S}}(1, -\widehat{\boldsymbol{\beta}}_{2SLS}')'$. Further show that when the model is just identified, $\widehat{\lambda} = 0$ and $\overline{\mathbf{S}}$ is singular.

Another closely related formulation of 2SLS is Basmann (1957)'s version of 2SLS. It is motivated by observing that $E[\mathbf{z}u] = \mathbf{0}$ implies

$$\mathbf{0} = E^*[u|\mathbf{z}] = E^*[y|\mathbf{z}] - E^*[\mathbf{x}|\mathbf{z}]'\boldsymbol{\beta}, \tag{23}$$

so

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \left(\widehat{\mathbf{X}}'\widehat{\mathbf{X}}\right)^{-1}\widehat{\mathbf{X}}'\widehat{\mathbf{y}}.$$

Equivalently, $\widehat{\boldsymbol{\beta}}_{2SLS} = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{P}_{\mathbf{Z}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, which is a GLS estimator. Yet another algebraically equivalent formulation of 2SLS is the control function formulation of Telser (1964)[4]:

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}_{2SLS} \\ \widehat{\boldsymbol{\rho}}_{2SLS} \end{pmatrix} = \left(\widehat{\mathbf{W}}'\widehat{\mathbf{W}}\right)^{-1} \widehat{\mathbf{W}}'\mathbf{y}, \tag{24}$$

where $\widehat{\mathbf{W}} = [\mathbf{X}, \widehat{\mathbf{V}}]$. This construction exploits another implication of $E[\mathbf{z}u] = \mathbf{0}$:

$$E^*[u|\mathbf{x}, \mathbf{z}] = E^*[u|\boldsymbol{\Gamma}'\mathbf{z} + \mathbf{v}, \mathbf{z}] = E^*[u|\mathbf{v}, \mathbf{z}] = E^*[u|\mathbf{v}] \equiv \mathbf{v}'\boldsymbol{\rho}$$

for some coefficient vector $\boldsymbol{\rho}$, where the third equality follows from the orthogonality of both error terms u and $\mathbf{v}$ with $\mathbf{z}$. Thus, this particular linear combination of the first-stage errors $\mathbf{v}$ is a function that controls for the endogeneity of the regressors $\mathbf{x}$. Basmann (1957)'s formulation (23) and the control function formulation of 2SLS can be extended to more general (especially nonlinear) models discussed in Chapter 1, but the fitted-value method seems hard to extend; see Blundell and Powell (2003).

Exercise 11 *(i) Show that* $E^*[u|\mathbf{v}, \mathbf{z}] = E^*[u|\mathbf{v}]$ *if* $E[\mathbf{z}u] = \mathbf{0}$ *and* $E[\mathbf{z}\mathbf{v}'] = \mathbf{0}$. *(ii) Show that (24) generates the same formula of* $\widehat{\boldsymbol{\beta}}_{2SLS}$ *as (22).*

Exercise 12 *Take the linear model*

$$y_i = x_i\beta + u_i, E[u_i|x_i] = 0,$$

where x_i *and* β *are scalars.*

(i) *Show that* $E[x_i u_i] = 0$ *and* $E[x_i^2 u_i] = 0$. *Is* $\mathbf{z}_i = (x_i, x_i^2)'$ *a valid instrumental variable for estimation of* β*?*

(ii) *Define the 2SLS estimator of* β*, using* $\mathbf{z}_i$ *as an instrument for* x_i. *How does this differ from OLS?*

Exercise 13 *Suppose* $y = m(\mathbf{x}) + u$, $E[\mathbf{x}u] \neq \mathbf{0}$ *and* $E[\mathbf{z}u] = \mathbf{0}$.

(i) *Derive the probability limit of* $\widehat{\boldsymbol{\beta}}_{2SLS}$.

(ii) *Show that* $\widehat{\boldsymbol{\beta}}_{2SLS}$ *is not the best linear predictor of* $m(\mathbf{x})$ *in the sense that*

$$plim\left(\widehat{\boldsymbol{\beta}}_{2SLS}\right) \neq \arg\min_{\boldsymbol{\beta}} E\left[\left(m(\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta}\right)^2\right].$$

It is useful to scrutinize the projection $\widehat{\mathbf{X}}$. Recall that $\mathbf{Z} = [\mathbf{X}_1, \mathbf{Z}_2]$ and $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, so

$$\widehat{\mathbf{X}} = [\mathbf{P}\mathbf{X}_1, \mathbf{P}\mathbf{X}_2] = [\mathbf{X}_1, \mathbf{P}\mathbf{X}_2] = \left[\mathbf{X}_1, \widehat{\mathbf{X}}_2\right],$$

[4]It is difficult to locate a definitive reference to the control function version of 2SLS. Dhrymes (1970, equation 4.3.57) formally discussed this formulation. Heckman (1978) attributed it to Telser (1964).

since $\mathbf{X}_1$ lies in the span of $\mathbf{X}$. Thus in the second stage, we regress y on $\mathbf{X}_1$ and $\widehat{\mathbf{X}}_2$. So only the endogenous variables $\mathbf{X}_2$ are replaced by their fitted values:

$$\widehat{\mathbf{X}}_2 = \mathbf{Z}_1\widehat{\mathbf{\Gamma}}_{12} + \mathbf{Z}_2\widehat{\mathbf{\Gamma}}_{22}.$$

Note that as a linear combination of $\mathbf{z}$, $\widehat{\mathbf{x}}_2$ is not correlated with u and it is often interpreted as the part of $\mathbf{x}_2$ that is uncorrelated with u.

Exercise 14 *In the structural model*

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}, \\ \mathbf{X}_2 &= \mathbf{X}_1\mathbf{\Gamma}_{12} + \mathbf{Z}_2\mathbf{\Gamma}_{22} + \mathbf{V}, \end{aligned}$$

(i) show that $\widehat{\boldsymbol{\beta}}_{2,2SLS} = \left(\mathbf{X}_2'\mathbf{P}_{\widetilde{\mathbf{Z}}_2}\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{P}_{\widetilde{\mathbf{Z}}_2}\mathbf{y}$*, where* $\widetilde{\mathbf{Z}}_2 = \mathbf{M}_{\mathbf{X}_1}\mathbf{Z}_2$*; (ii) Given that* $\mathbf{x}_{1i}$ *are the included exogenous variables with* $E[\mathbf{x}_{1i}u_i] = \mathbf{0}$*, does* $\mathbf{X}_1'\widehat{\mathbf{u}}$ *equal* $\mathbf{0}$*? where* $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{2SLS}$*.*

Example 5 (Wald Estimator) *The Wald estimator is a special IV estimator when the single instrument* z *is binary. Suppose we have the model*

$$\begin{aligned} y &= \beta_0 + \beta_1 x + u,\ Cov(x,u) \neq 0, \\ x &= \gamma_0 + \gamma_1 z + u. \end{aligned}$$

The identification conditions are

$$Cov(z,x) \neq 0, Cov(z,u) = 0.^5 \tag{25}$$

From Exercise 14, the IV estimator is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \overline{z})(y_i - \overline{y})}{\sum_{i=1}^{n}(z_i - \overline{z})(x_i - \overline{x})}.$$

If z *is binary that takes the value* 1 *for* n_1 *of the* n *observations and* 0 *for the remaining* n_0 *observations, then* $\widehat{\beta}_1$ *is equivalent to*

$$\widehat{\beta}_{Wald} = \frac{\overline{y}_1 - \overline{y}_0}{\overline{x}_1 - \overline{x}_0},$$

where $\overline{y}_1$ *is mean of* y *across the* n_1 *observations with* $z = 1$*,* $\overline{y}_0$ *is the mean of* y *across the* n_0

[5] In view of Example 4, why is $Cov(z,x) \neq 0$ the right identification condition? This is because $x_1 = 1$ in this example, and $Cov(z,x) \neq 0$ is not only necessary but also sufficient for identification.

observations with $z = 0$, and analogously for x. Why?

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} y_i z_i - \overline{y} z_i}{\sum_{i=1}^{n} x_i z_i - \overline{x} z_i} = \frac{\sum_{z_i=1} (y_i z_i - \overline{y} z_i) + \sum_{z_i=0} (y_i z_i - \overline{y} z_i)}{\sum_{z_i=1} (x_i z_i - \overline{x} z_i) + \sum_{z_i=0} (x_i z_i - \overline{x} z_i)} \\ &= \frac{\sum_{z_i=1} (y_i z_i - \overline{y} z_i) / n_1}{\sum_{z_i=1} (x_i z_i - \overline{x} z_i) / n_1} = \frac{\overline{y}_1 - \overline{y}}{\overline{x}_1 - \overline{x}} = \frac{\overline{y}_1 - \overline{y}_0}{\overline{x}_1 - \overline{x}_0},\end{aligned}$$

where the last equality is from $\overline{y} = \frac{n_1 \overline{y}_1 + n_0 \overline{y}_0}{n}$. This estimator is called the Wald estimator *first proposed in Wald (1940) and converges in probability to*

$$\frac{E[y|z=1] - E[y|z=0]}{E[x|z=1] - E[x|z=0]}. \tag{26}$$

A simple interpretation of this estimator is to take the effect of z on y and divide by the effect of z on x. Figure 4 provides some intuition for the identification scheme of the Wald estimator in the linear demand/supply system - the shift in p by z devided by the shift in q by z is indeed a reasonable slope estimator of the demand curve.

The Wald estimator has many applications. In Card (1995), y is the log weekly wage, x is years of schooling S, and z is a dummy which equals 1 if born in the neighborhood of an university and 0 otherwise. In studying the returns to schooling in China, someone ever used a dummy indicator of living through the Cultural Revolution or not as z. Angrist and Evans (1998) use the dummy of whether the sexes of the first two children are the same, which indicates the parental preferences for a mixed sibling-sex composition, (and also a twin second birth) as the instrument to study the effect of a third child on employment, hours worked and labor income. Hearst et al. (1986) and Angrist (1990) use the Vietnam era draft lottery as an instrument for veteran status to identify the effects of mandatory military conscription on subsequent civilian mortality and earnings. Imbens et al. (2001) use "winning a prize in the lottery" as an instrument to identify the effects of unearned income on subsequent labor supply, earnings, savings and consumption behavior. □

Exercise 15 *Consider the single equation model*

$$y_i = x_i \beta + u_i,$$

where y_i and x_i are both real-valued. Let $\widehat{\beta}$ denote the IV estimator of β using as instrument a dummy variable d_i (takes only the values 0 and 1). Find a simple expression for the IV estimator in this context and derive its probability limit. What is the difference between this probability limit and the probability limit of the Wald estimator?

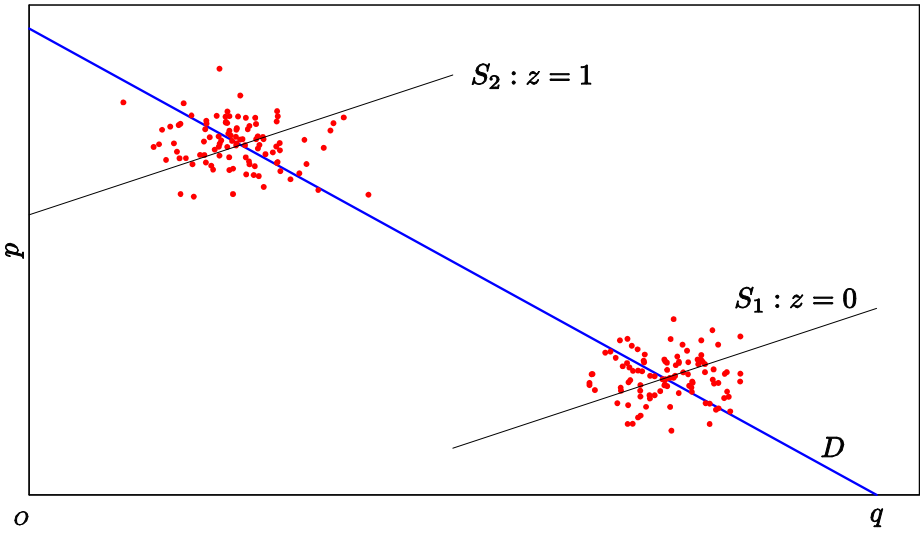

Figure 4: Intuition for the Wald Estimator in the Linear Demand/Supply System

Exercise 16 (*) *Suppose*

$$\begin{aligned} y &= \beta_0 + \beta_1 x + u,\ Cov(x,u) \neq 0, \\ x &= \gamma_0 + \gamma_1 z + v, E[u|z] = 0, E[zv] = 0, \end{aligned}$$

where x is binary. Unless z is binary, $E[x|z]$ cannot be a linear function. Suppose we run a Probit regression in the first stage and get $\widehat{x} = \Phi\left(\widehat{\gamma}_0 + \widehat{\gamma}_1 z\right)$.

(i) *Show that if $E[x|z] = \Phi\left(\gamma_0 + \gamma_1 z\right)$, then $\widehat{\boldsymbol{\beta}} \equiv \left(\widehat{\beta}_0, \widehat{\beta}_1\right)'$ based on regressing y on $1, \widehat{x}$ is consistent.*

(ii) *Show that if $E[x|z] \neq \Phi\left(\gamma_0 + \gamma_1 z\right)$, then $\widehat{\boldsymbol{\beta}}$ based on regressing y on $1, \widehat{x}$ is not consistent.*

(iii) *Show that if $E[x|z] = \Phi\left(\gamma_0 + \gamma_1 z\right)$, $plim\left(\widehat{\boldsymbol{\beta}}\right)$ is the same as the plim of the IV estimator using $(1, \widehat{x})$ as the instrumental variables, but if $E[x|z] \neq \Phi\left(\gamma_0 + \gamma_1 z\right)$, they are generally different.*

(iv) *Show that the IV estimator using $(1, \widetilde{x})$ as the instrumental variables is consistent, where $\widetilde{x}$ is the linear projection of x on $(1, z)$.*

(Hint: if $E[x|z] = \Phi\left(\gamma_0 + \gamma_1 z\right)$, $\widehat{\boldsymbol{\gamma}} = \left(\widehat{\gamma}_0, \widehat{\gamma}_1\right)'$ is consistent; otherwise, it is inconsistent.)

6 Interpretation of the IV Estimator

In this section, we intend to answer two questions: (i) How to interpret the IV estimator (and the 2SLS estimator) in the projection language of Chapter 2? (ii) What is the IV estimator estimating?

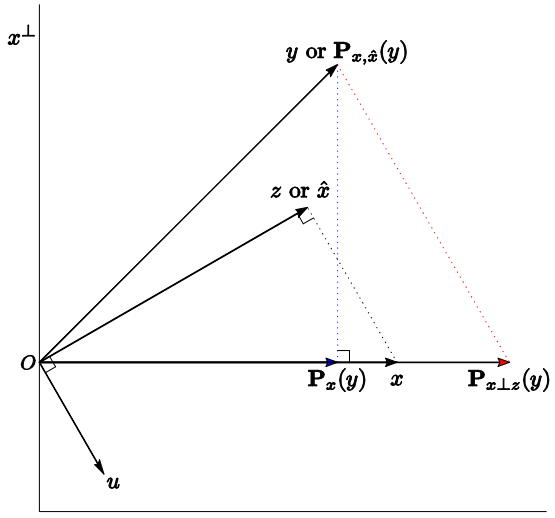

Figure 5: Projection Interpretation of the IV Estimator

6.1 Geometric Interpretation of the IV Estimator

Figure 5 illustrates the geometric meaning of the IV estimator. For simplicity, we assume $k = k_2 = 1$ and $l = l_2 = 1$; also, we discuss the population version of the IV estimator instead of the sample version and denote $\text{plim}\left(\widehat{\beta}_{IV}\right)$ as β_{IV}. In this simple case, $x\beta_{IV}$ is the projection of y onto $span(x)$ along $span^{\perp}(z)$; this can be easily seen from $x\beta_{IV} = xE[zx]^{-1}E[zy] \equiv \mathbf{P}_{x\perp z}(y)$ (compare to $\mathbf{P_{X\perp Z}(y)}$ in Section 4.1 of Chapter 2). Since $z \perp u$, this is also the projection of y onto $span(x)$ along u. In the figure, $\mathbf{P}_{x\perp z}(y)$ is very different from the orthogonal projection of y onto $span(x)$ - $\mathbf{P}_x(y) \equiv xE[x^2]^{-1}E[xy]$, because z is different from x (otherwise, $E[zu] \neq 0$ since $E[xu] > 0$ in the figure). On the other hand, z cannot be orthogonal to x in the figure (which corresponds to the rank condition); otherwise, $\mathbf{P}_{x\perp z}(y)$ is not well defined. So z must be between x and $x^{\perp}$, just as shown in the figure.

If there are more than one instruments, or $l > 1$, then the 2SLS estimator first orthogonally projects x onto $span(\mathbf{z})$ to get $\widehat{x}$ in the figure, and then projects y onto $span(x)$ along $span^{\perp}(\widehat{x})$. Now, $span(\mathbf{z})$ determines the direction of $\widehat{x}$. Also, y in the figure should be replaced by $\mathbf{P}_{x,\widehat{x}}(y)$ by noting that $\mathbf{P}_{x\perp\widehat{x}}(y) = xE[\widehat{x}x]^{-1}E[\widehat{x}y] = xE[\widehat{x}x]^{-1}E[\widehat{x}\mathbf{P}_{x,\widehat{x}}(y)]$, where $\mathbf{P}_{x,\widehat{x}}(y)$ is the projection of y on $span\,(x, \widehat{x})$.

6.2 What is the IV Estimator Estimating? (*)

What is the IV estimator estimating? This is an interesting question. To be specific, consider the example of Angrist and Krueger (1991). In this example, z is a dummy variable equal to 0 if born in the first quarter of the year and 1 otherwise, $x = S$ is also a dummy indicating high school

graduates versus nongraduates, and y is the log weekly wage. Since z is dummy, the IV estimator is the Wald estimator. To acknowledge the dependence of y on S, we use y_j to denote the log weekly wage with jth level of schooling, $j = 0, 1$, and to acknowledge the dependence of S on z, we use S_i to denote the schooling level when $z = i$, $i = 0, 1$.

Now, $S = z \cdot S_1 + (1 - z) \cdot S_0$, and $y = y_0 + (y_1 - y_0)S$. The numerator of (26)

$$\begin{aligned} & E[y|z = 1] - E[y|z = 0] \\ = \; & E[y_0 + (y_1 - y_0)S_1|z = 1] - E[y_0 + (y_1 - y_0)S_0|z = 0] \\ = \; & E[y_0 + (y_1 - y_0)S_1] - E[y_0 + (y_1 - y_0)S_0] \\ = \; & E[(y_1 - y_0) \cdot (S_1 - S_0)] \end{aligned}$$

where the second equality is from the exclusion restriction which requires that z affects y only through S, e.g., the independence of z with (y_0, y_1, S_0, S_1) can guarantee this. Using the more familiar notations, we write the system as

$$\begin{aligned} y &= \beta_0 + \beta_1 S + u \\ S &= \gamma_0 + \gamma_1 z + v, \end{aligned}$$

where $y_j = \beta + j \cdot \beta_1 + u$, and $S_i = \gamma_0 + i \cdot \gamma_1 + v$. If z is independent of (u, v), the second equality follows.

We classify the possibility of S_1 and S_0 in the following table.

		S_0	
		0	1
S_1	0	$y_0 - y_0 = 0$ Never-taker	$y_0 - y_1 = -(y_1 - y_0)$ Defier
	1	$y_1 - y_0$ Complier	$y_1 - y_1 = 0$ Always-taker

Table: Causal Effect of z on y, $y_{S_1} - y_{S_0}$ Classified by S_0 and S_1

The name "complier" in the table is because this group of individuals always comply with their assignment z; other names can be similarly understood. $S_1 - S_0$ could be $-1, 0$, or 1, where 0 indicates those whose schooling status is unchanged, and 1 and -1 could be similarly understood. Therefore,

$$\begin{aligned} & E[(y_1 - y_0) \cdot (S_1 - S_0)] \\ = \; & E[(y_1 - y_0)|S_1 - S_0 = 1] P(S_1 - S_0 = 1) + E[(y_1 - y_0)|S_1 - S_0 = -1] P(S_1 - S_0 = -1). \end{aligned}$$

If for everyone, we always have $S_1 - S_0 = 1$ or 0, that is, compulsory attendance laws cannot reduce

schooling, then

$$E\left[(y_1 - y_0) \cdot (S_1 - S_0)\right] = E\left[(y_1 - y_0) \left| S_1 - S_0 = 1\right.\right] P\left(S_1 - S_0 = 1\right).$$

In the table, we exclude the possibility of defiers. This is the *monotonicity* assumption in Imbens and Angrist (1994).[6] The denominator of (26)

$$E\left[S|z=1\right] - E\left[S|z=0\right] = E\left[S_1 - S_0\right] = P\left(S_1 - S_0 = 1\right),$$

so we have

$$\widehat{\beta}_1 \xrightarrow{p} E\left[(y_1 - y_0) \left| S_1 - S_0 = 1\right.\right].$$

$E\left[(y_1 - y_0) \left| S_1 - S_0 = 1\right.\right]$ is called the local average treatment effect in Imbens and Angrist (1994), that is, the average treatment effect for those individuals whose schooling decision is affected by the law. This set of individuals is only implicitly defined and cannot be observed. $\widehat{\beta}_1$ is called the **local average treatment effect estimator** (LATE). For different z, this set of individuals is different, so different from the usual estimators (such as the LSE) whose interpretations are invariant, the interpretation of the IV estimator depends on the choice of instruments.

When S takes $J > 2$ levels, Angrist and Imbens (1995) show that

$$\widehat{\beta}_1 \xrightarrow{p} \sum_{j=1}^{J} \omega_j \cdot E\left[y_i - y_{j-1} | S_1 \geq j > S_0\right] \equiv \beta$$

where $\omega_j = \frac{P(S_1 \geq j > S_0)}{\sum_{j=1}^{J} P(S_1 \geq j > S_0)}$. That is, β_1 is a weighted average of per-unit treatment effect. For the case with continuous S, see Angrist et al. (2000).

7 LIML (*)

Simultaneous equations models can be estimated by the MLE, which is called the **full-information maximum likelihood** (FIML) **estimator**. Sometimes, we are only interested in the parameters of a single equation. The corresponding MLE is called the **limited-information maximum likelihood** (LIML) **estimator**. This estimator is proposed by Anderson and Rubin (1949, 1950), and is the ML counterpart of the 2SLS estimator. The LIML estimator predates the 2SLS estimator and is asymptotically equivalent to the 2SLS estimator given homoskedastic errors. The LIML estimator is less efficient than the FIML estimator, but more robust (invariant to the normalization used in a simultaneous equations system). This section is based on Section 8.6 of Hayashi (2000).

If only one-equation is of interest, we come back to the setup of Section 2, 3 and 4. The notations

[6] Balke and Pearl (1997) refer to it as the "no-defiance" assumption, and Heckman and Vytlacil (2005) call it the uniformity assumption because *all* individuals respond to z in the same direction.

there can be applied in this case. Define

$$\underset{((k_2+1)\times(k_2+1))}{\mathbf{B}} = \begin{pmatrix} 1 & \mathbf{0} \\ -\boldsymbol{\beta}_2 & \mathbf{I}_{k_2} \end{pmatrix}, \underset{((k_2+1)\times l)}{\mathbf{\Gamma}} = \begin{pmatrix} \boldsymbol{\beta}_1 & \mathbf{\Gamma}_{12} \\ 0 & \mathbf{\Gamma}_{22} \\ 1 & k_2 \end{pmatrix} \begin{matrix} k_1 \\ l_2 \end{matrix},$$

$$\boldsymbol{\theta} = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{\Gamma}_{12}, \mathbf{\Gamma}_{22}\right), \mathsf{y}_i = \left(y_i, \mathbf{x}_{2i}'\right)', \mathbf{e}_i = \begin{pmatrix} u_i \\ \mathbf{v}_{2i} \end{pmatrix},$$

where y_i collects endogenous variables. If $\mathbf{e}_i|\mathbf{z}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$, then the average log-likelihood

$$\ell_n(\boldsymbol{\theta}, \mathbf{\Sigma}) = -\frac{k_2+1}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{\Sigma}|) - \frac{1}{2n}\sum_{i=1}^{n}\left(\mathbf{B}'\mathsf{y}_i - \mathbf{\Gamma}'\mathbf{z}_i\right)'\mathbf{\Sigma}^{-1}\left(\mathbf{B}'\mathsf{y}_i - \mathbf{\Gamma}'\mathbf{z}_i\right),$$

where $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$, and note that the Jacobian of the transformation from y_i to $\mathbf{e}_i$ is $\mathbf{B}$ whose determinant is 1. The average log likelihood function concentrated with respect to the parameter $\mathbf{\Sigma}, \boldsymbol{\beta}_1, \mathbf{\Gamma}_{12}, \mathbf{\Gamma}_{22}$ (see the technical appendix) is

$$\ell_n(\boldsymbol{\beta}_2) = -\frac{k_2+1}{2}\log(2\pi) - \frac{1}{2}\log\kappa(\boldsymbol{\beta}_2) - \frac{1}{2}\log\left|\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\right|,$$

where

$$\kappa(\boldsymbol{\beta}_2) = \frac{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_1\mathbf{Y}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\boldsymbol{\gamma}} = \frac{\left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}_2\boldsymbol{\beta}_2\right)'\left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}_2\boldsymbol{\beta}_2\right)}{\left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}_2\boldsymbol{\beta}_2\right)'\mathbf{M}_{\widetilde{\mathbf{Z}}_2}\left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}_2\boldsymbol{\beta}_2\right)} \tag{27}$$

with $\gamma = (1, -\boldsymbol{\beta}_2')'$, $Y = [\mathbf{y}, \mathbf{X}_2]$, $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'$ and for any random matrix $\mathbf{A}$, $\widetilde{\mathbf{A}} = \mathbf{M}_1\mathbf{A}$. Maximizing $\ell_n(\boldsymbol{\beta}_2)$ is equivalent to minimizing $\kappa(\boldsymbol{\beta}_2)$. Because $\boldsymbol{\beta}_2$ can be obtained in this way, LIML estimates are sometimes referred to as **least variance ratio** estimates. First of all, $\widehat{\kappa} \equiv \kappa\left(\widehat{\boldsymbol{\beta}}_2\right) \geq 1$, since $span(\mathbf{Z}_1) \subset span(\mathbf{Z})$ and the numerator of $\kappa(\boldsymbol{\beta}_2)$ cannot be smaller than the denominator for any possible $\boldsymbol{\gamma}$. In fact, for any equation that is overidentified, $\widehat{\kappa}$ will always be greater than 1 in finite samples. For an equation that is just identified, $\widehat{\kappa}$ will be exactly equal to 1 because the number of free parameters to be estimated is then just equal to k, the rank of $\mathbf{Z}$. Thus, in this case, it is possible to choose $\boldsymbol{\gamma}$ so that the numerator and denominator of (27) are equal.

Thanks to the special form of $\mathbf{B}$ and no exclusion restrictions in the endogenous variable regression, there is a closed-form solution to the LIML estimator (see the technical appendix):

$$\left(\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2'\right)' = \left[\mathbf{X}'\left(\mathbf{I}_n - \widehat{\kappa}\mathbf{M}_{\mathbf{Z}}\right)\mathbf{X}\right]^{-1}\mathbf{X}'\left(\mathbf{I}_n - \widehat{\kappa}\mathbf{M}_{\mathbf{Z}}\right)\mathbf{y}, \tag{28}$$

where $\widehat{\kappa}$ is the smallest characteristic root of $\mathbf{W}_1\mathbf{W}^{-1}$ with

$$\underset{((k_2+1)\times(k_2+1))}{\mathbf{W}_1} = \mathbf{Y}'\mathbf{M}_1\mathbf{Y}, \underset{((k_2+1)\times(k_2+1))}{\mathbf{W}} = \mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y},$$

or the smallest root of the determinantal equation $|\mathbf{W}_1 - \kappa\mathbf{W}| = 0$. The covariance matrix for the LIML estimator $\widehat{\boldsymbol{\beta}}$ can be estimated by

$$\widehat{\sigma}^2 \left[\mathbf{X}' \left(\mathbf{I}_n - \widehat{\kappa}\mathbf{M}\right) \mathbf{X}\right]^{-1},$$

where $\widehat{\sigma}^2 = n^{-1} \left(\mathbf{y} - \mathbf{X}'\widehat{\boldsymbol{\beta}}\right)' \left(\mathbf{y} - \mathbf{X}'\widehat{\boldsymbol{\beta}}\right)$. The likelihood ratio statistic for testing overidentifying restrictions reduces to

$$LR = n \log \widehat{\kappa},$$

which converges to χ^2_{l-k}. When $l = k$, $\widehat{\kappa} = 1$ and $LR = 0$ as expected. This test statistic was first proposed by Anderson and Rubin (1950).

The LIML estimator (28) is a **K-class estimator**; see Theil (1961) and Nagar (1959). Inspection of the 2SLS formula (22) shows that the 2SLS estimator is a K-class estimator with $\widehat{\kappa} = 1$, and the OLS estimator is a K-class estimator with $\widehat{\kappa} = 0$. It follows that the LIML and 2SLS are numerically the same when the equation is just identified. When the errors are normally distributed, the 2SLS estimator has the p-th moment when $p \leq l - k$, that is, the number of finite moments for 2SLS equals the number of overidentifying restrictions; see Mariano and Sawa (1972), Sawa (1969) and also Richardson (1968) and Kinal (1980). This implies that 2SLS does not even have a mean if the equation is just identified. On the other hand, the LIML estimator has no finite moments; see Mariano (1982) and Phillips (1983). Nevertheless, Anderson et al. (1982) present analytical results that show that LIML approaches its asymptotic normal distribution much more rapidly than 2SLS. They also show that the LIML has a less median bias than the 2SLS estimator. Sargan (1958) reports that the bias of the 2SLS is of the order of the inverse of the minimum population canonical correlation between $\mathbf{X}_2$ and $\mathbf{Z}_2$.[7] Hillier (1990) criticizes the 2SLS estimator by arguing that the object that is identified is the direction $(1, -\boldsymbol{\beta}')'$ but not its magnitude. He then shows that the 2SLS estimator of direction is distorted by its dependence on normalization of the parameter. On the other hand, the LIML is less sensitive to the normalization and is a better estimator of direction. Bekker (1994) presents small-sample results for LIML and a generalization of LIML. Hahn and Hausman (2002) justify the use of the LIML estimator in some cases with weak instruments.

There are many other K-class estimators. For example, Sawa (1973) has suggested a way of modifying the 2SLS estimator to reduce bias, and Fuller (1977) and Morimune (1978, 1983) have suggested modified versions of the LIML estimator. Fuller's estimator, which is the simplest of these, uses $\widehat{\kappa} = \widehat{\kappa}_{LIML} - \alpha/(n - l)$ with α being a fixed number. One good choice is $\alpha = 1$, since it yields estimates that are approximately unbiased. In contract to the LIML estimator, which has no finite moments, Fuller's modified estimator has all moments finite provided the sample size is large enough.

It has been shown that all members of the K-class for which K converges to 1 at a rate faster than $n^{-1/2}$ have the same asymptotic distribution as the 2SLS estimator. These are largely of

[7] He also gave a minimax instrumental-variable interpretation to the original LIML estimator. For a minimum distance interpretation of the LIML estimator, see Goldberger (1971).

theoretical interest, given the pervasive use of 2SLS or OLS. The large sample properties of all K-class estimators are the same, but the finite sample properties are possibly very different. Mariano (1982) discusses a number of analytical results and provides some guidance as to when LIML is likely to perform better than 2SLS. He suggests some evidence favors LIML when the sample size is not large while the number of overidentifying restrictions is. However, much depends on the particular model and data set.

LIML is rarely used as it is more difficult to implement and harder to explain than 2SLS. Nevertheless, Pagan (1979) shows that the LIML estimator can be computed by treating the system of equations as a **seemingly unrelated regressions** (SUR) models, ignoring both the constraints on the reduced form and the correlation between $\mathbf{x}_{2i}$ and u_i, and using the iterative GLS method.

Exercise 17 (Empirical) *The data file card.dat is taken from David Card "Using Geographic Variation in College Proximity to Estimate the Return to Schooling" in Aspects of Labour Market Behavior (1995). There are 2215 observations with 29 variables, listed in card.pdf. We want to estimate a wage equation*

$$\log(Wage) = \beta_0 + \beta_1 Educ + \beta_2 Exper + \beta_2 Exper^2 + \beta_4 South + \beta_5 Black + u,$$

where Educ = Education (Years), Exper = Experience (Years), and South and Black are regional and racial dummy variables.

(a) *Estimate the model by OLS. Report estimates and standard errors.*

(b) *Now treat Education as endogenous, and the remaining variables as exogenous. Estimate the model by 2SLS, using the instrument near4, a dummy indicating that the observation lives near a 4-year college. Report estimates and standard errors.*

(c) *Re-estimate by 2SLS (report estimates and standard errors) adding three additional instruments: near2 (a dummy indicating that the observation lives near a 2-year college), fatheduc (the education, in years, of the father) and motheduc (the education, in years, of the mother).*

Technical Appendix: Concentrated Likelihood in the LIML

The derivation here is based on Section 18.5 of Davidson and MacKinnon (1993). First concentrate out $\mathbf{\Sigma}$. Taking derivative with respect to $\mathbf{\Sigma}^{-1}$, we have

$$\frac{\partial \ell_n(\boldsymbol{\theta}, \mathbf{\Sigma})}{\partial \mathbf{\Sigma}^{-1}} = \frac{1}{2}\mathbf{\Sigma} - \frac{1}{2n}\sum_{i=1}^{n}\left(\mathbf{B}'\mathsf{y}_i - \mathbf{\Gamma}'\mathbf{z}_i\right)'\left(\mathbf{B}'\mathsf{y}_i - \mathbf{\Gamma}'\mathbf{z}_i\right),$$

so

$$\widehat{\mathbf{\Sigma}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{B}'\mathsf{y}_i - \mathbf{\Gamma}'\mathbf{z}_i\right)'\left(\mathbf{B}'\mathsf{y}_i - \mathbf{\Gamma}'\mathbf{z}_i\right) = \frac{1}{n}\left(\mathbf{YB} - \mathbf{Z\Gamma}\right)'\left(\mathbf{YB} - \mathbf{Z\Gamma}\right).$$

As a result, the concentrated average log-likelihood is

$$\begin{aligned}\ell_n(\boldsymbol{\theta}) &= -\frac{k_2+1}{2}(\log(2\pi)+1) - \frac{1}{2}\log\left|\frac{1}{n}(\mathbf{YB}-\mathbf{Z\Gamma})'(\mathbf{YB}-\mathbf{Z\Gamma})\right| \\ &= -\frac{k_2+1}{2}(\log(2\pi)+1) - \frac{1}{2}\log\left|\frac{1}{n}\left(\mathbf{Y}-\mathbf{Z\Gamma B}^{-1}\right)'\left(\mathbf{Y}-\mathbf{Z\Gamma B}^{-1}\right)\right|,\end{aligned}$$

where the second equality is due to $|\mathbf{B}|=1$. Note that

$$\mathbf{\Gamma B}^{-1} = \begin{pmatrix} \boldsymbol{\beta}_1 & \mathbf{\Gamma}_{12} \\ 0 & \mathbf{\Gamma}_{22} \end{pmatrix}\begin{pmatrix} 1 & \mathbf{0} \\ \boldsymbol{\beta}_2 & \mathbf{I}_{k_2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1+\mathbf{\Gamma}_{12}\boldsymbol{\beta}_2 & \mathbf{\Gamma}_{12} \\ \mathbf{\Gamma}_{22}\boldsymbol{\beta}_2 & \mathbf{\Gamma}_{22} \end{pmatrix},$$

which is the (restricted) reduced-from coefficient matrix, the top part corresponds to $\mathbf{Z}_1$ and the bottom part to $\mathbf{Z}_2$. Since $\boldsymbol{\beta}_1$ does not appear in the bottom part, it is clear that for whatever value of $\boldsymbol{\beta}_2$, we can find values of $\boldsymbol{\beta}_1$ and $\mathbf{\Gamma}_{12}$ such that the top part is equal to anything at all. In other words, the structural equations do not impose any restrictions on the (unrestricted) reduced-form coefficients corresponding to $\mathbf{Z}_1$. In general, however, they do impose restrictions on the coefficients corresponding to $\mathbf{Z}_2$.

It is obvious that minimizing $\ell_n(\boldsymbol{\theta})$ is equivalent to minimizing $\left|\left(\mathbf{Y}-\mathbf{Z\Gamma B}^{-1}\right)'\left(\mathbf{Y}-\mathbf{Z\Gamma B}^{-1}\right)\right|$. If there are no restrictions on the coefficients of $\mathbf{Z}$, then the minimizer (corresponding to $\mathbf{\Gamma B}^{-1}$) should be the OLS estimate $\widehat{\mathbf{\Pi}}$ which minimizes $\left|(\mathbf{Y}-\mathbf{Z\Pi})'(\mathbf{Y}-\mathbf{Z\Pi})\right|$. To see why, let $\widetilde{\mathbf{\Pi}} = \widehat{\mathbf{\Pi}}+\mathbf{A}$ be another candidate. Then

$$\begin{aligned}&\left|\left(\mathbf{Y}-\mathbf{Z}\widetilde{\mathbf{\Pi}}\right)'\left(\mathbf{Y}-\mathbf{Z}\widetilde{\mathbf{\Pi}}\right)\right| \\ =\ &\left|\left(\mathbf{Y}-\mathbf{Z}\widehat{\mathbf{\Pi}}-\mathbf{ZA}\right)'\left(\mathbf{Y}-\mathbf{Z}\widehat{\mathbf{\Pi}}-\mathbf{ZA}\right)\right| \\ =\ &\left|(\mathbf{M_Z Y}-\mathbf{ZA})'(\mathbf{M_Z Y}-\mathbf{ZA})\right| \\ =\ &\left|\mathbf{Y}'\mathbf{M_Z Y}+\mathbf{A}'\mathbf{Z}'\mathbf{ZA}\right|.\end{aligned}$$

Because the determinant of the sum of two positive definite matrices is always greater than the determinants of either of those matrix, we can see $\widehat{\mathbf{\Pi}}$ is indeed the minimizer. Since there are no restrictions on the rows of $\mathbf{\Pi}$ that corresponds to $\mathbf{Z}_1$, we can use OLS to estimate those parameters, and then concentrate them out of the determinant. When we do this, the determinant of $\left|(\mathbf{YB}-\mathbf{Z\Gamma})'(\mathbf{YB}-\mathbf{Z\Gamma})\right|$, which equals that of $\left|\left(\mathbf{Y}-\mathbf{Z\Gamma B}^{-1}\right)'\left(\mathbf{Y}-\mathbf{Z\Gamma B}^{-1}\right)\right|$, becomes

$$\left|(\mathbf{YB}-\mathbf{Z\Gamma})'\mathbf{M}_1(\mathbf{YB}-\mathbf{Z\Gamma})\right|,$$

which can be rewritten as

$$\begin{vmatrix} \left(\widetilde{\mathbf{Y}}\boldsymbol{\gamma}\right)'\left(\widetilde{\mathbf{Y}}\boldsymbol{\gamma}\right) & \left(\widetilde{\mathbf{Y}}\boldsymbol{\gamma}\right)'\left(\widetilde{\mathbf{X}}_2-\widetilde{\mathbf{Z}}_2\mathbf{\Gamma}_{22}\right) \\ \left(\widetilde{\mathbf{X}}_2-\widetilde{\mathbf{Z}}_2\mathbf{\Gamma}_{22}\right)'\left(\widetilde{\mathbf{Y}}\boldsymbol{\gamma}\right) & \left(\widetilde{\mathbf{X}}_2-\widetilde{\mathbf{Z}}_2\mathbf{\Gamma}_{22}\right)'\left(\widetilde{\mathbf{X}}_2-\widetilde{\mathbf{Z}}_2\mathbf{\Gamma}_{22}\right) \end{vmatrix} \tag{29}$$

This determinant depends only on $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}_{22}$; we further concentrate out $\boldsymbol{\Gamma}_{22}$. Using the result that for any matrix $\mathbf{A}$ and $\mathbf{B}$,

$$\left| \begin{array}{cc} \mathbf{A}'\mathbf{A} & \mathbf{A}'\mathbf{B} \\ \mathbf{B}'\mathbf{A} & \mathbf{B}'\mathbf{B} \end{array} \right| = \left|\mathbf{A}'\mathbf{A}\right| \left|\mathbf{B}'\mathbf{M}_{\mathbf{A}}\mathbf{B}\right|, \tag{30}$$

we have our target (29) equal to

$$\left(\widetilde{\mathbf{Y}}\boldsymbol{\gamma}\right)' \left(\widetilde{\mathbf{Y}}\boldsymbol{\gamma}\right) \left| \left(\widetilde{\mathbf{X}}_2 - \widetilde{\mathbf{Z}}_2\boldsymbol{\Gamma}_{22}\right)' \mathbf{M}_{\mathbf{v}} \left(\widetilde{\mathbf{X}}_2 - \widetilde{\mathbf{Z}}_2\boldsymbol{\Gamma}_{22}\right) \right|, \tag{31}$$

where $\mathbf{v} = \widetilde{\mathbf{Y}}\boldsymbol{\gamma}$, and note that $\left(\widetilde{\mathbf{Y}}\boldsymbol{\gamma}\right)' \left(\widetilde{\mathbf{Y}}\boldsymbol{\gamma}\right)$ is scalar so its determinant is itself. The parameters $\boldsymbol{\Gamma}_{22}$ appears only in the second factor of (31). This factor is the determinant of the matrix of sums of squares and cross-products of the residuals from regressions of $\mathbf{M}_{\mathbf{v}}\widetilde{\mathbf{X}}_2$ on $\mathbf{M}_{\mathbf{v}}\widetilde{\mathbf{Z}}_2$. From the discussion above, the minimizer (corresponding to $\boldsymbol{\Gamma}_{22}$) should be the OLS estimate, and the matrix of residuals is $\mathbf{M}_{\mathbf{M}_{\mathbf{v}}\widetilde{\mathbf{Z}}_2}\mathbf{M}_{\mathbf{v}}\widetilde{\mathbf{X}}_2 = \mathbf{M}_{\mathbf{v},\widetilde{\mathbf{Z}}_2}\widetilde{\mathbf{X}}_2$. Consequently, the second factor of (31), minimized with respect to $\boldsymbol{\Gamma}_{22}$, is

$$\left|\widetilde{\mathbf{X}}_2'\mathbf{M}_{\mathbf{v},\widetilde{\mathbf{Z}}_2}\widetilde{\mathbf{X}}_2\right|. \tag{32}$$

The fact that $\mathbf{v}$ and $\widetilde{\mathbf{Z}}_2$ appear in a symmetrical fashion in (32) can be exploited in order to make (32) depend on $\boldsymbol{\gamma}$ only through a scalar factor. Consider the determinant

$$\left| \begin{array}{cc} \mathbf{v}'\mathbf{M}_{\widetilde{\mathbf{Z}}_2}\mathbf{v} & \mathbf{v}'\mathbf{M}_{\widetilde{\mathbf{Z}}_2}\widetilde{\mathbf{X}}_2 \\ \widetilde{\mathbf{X}}_2'\mathbf{M}_{\widetilde{\mathbf{Z}}_2}\mathbf{v} & \widetilde{\mathbf{X}}_2'\mathbf{M}_{\widetilde{\mathbf{Z}}_2}\widetilde{\mathbf{X}}_2 \end{array} \right|. \tag{33}$$

By use of (30), this determinant can be factorized just as (29) was. We obtain

$$\left(\mathbf{v}'\mathbf{M}_{\widetilde{\mathbf{Z}}_2}\mathbf{v}\right) \left|\widetilde{\mathbf{X}}_2'\mathbf{M}_{\mathbf{v},\widetilde{\mathbf{Z}}_2}\widetilde{\mathbf{X}}_2\right|.$$

Using the facts that $\mathbf{M}_1\mathbf{M}_{\widetilde{\mathbf{Z}}_2} = \mathbf{M}_{\mathbf{Z}}$ and that $\mathbf{v} = \mathbf{M}_1\mathbf{Y}\boldsymbol{\gamma}$, (33) can be rewritten as

$$\left| \begin{array}{cc} \boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\boldsymbol{\gamma} & \boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\boldsymbol{\gamma} & \mathbf{X}_2'\mathbf{M}_{\mathbf{Z}}\mathbf{X}_2 \end{array} \right| = \left|\mathbf{B}'\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\mathbf{B}\right| = \left|\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\right|, \tag{34}$$

where the first equality is from the definition of $\mathbf{B}$, and the second equality is from the fact that $|\mathbf{B}| = 1$. It implies that (34) does not depend on $\mathbf{B}$ at all.

In summary, minimizing the concentrated log-likelihood is equivalent to minimizing

$$\left(\widetilde{\mathbf{Y}}\boldsymbol{\gamma}\right)' \left(\widetilde{\mathbf{Y}}\boldsymbol{\gamma}\right) \frac{|\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}|}{\mathbf{v}'\mathbf{M}_{\widetilde{\mathbf{Z}}_2}\mathbf{v}} = \frac{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_1\mathbf{Y}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\boldsymbol{\gamma}} \left|\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\right| = \kappa\left(\boldsymbol{\beta}_2\right) \left|\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\right|,$$

or minimizing $\kappa\left(\boldsymbol{\beta}_2\right)$ since $|\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}|$ is free of parameters. Differentiating $\kappa\left(\boldsymbol{\beta}_2\right)$ with respect to

$\boldsymbol{\gamma}$, we have the FOCs:

$$2\mathbf{Y}'\mathbf{M}_1\mathbf{Y}\boldsymbol{\gamma}\left(\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_\mathbf{Z}\mathbf{Y}\boldsymbol{\gamma}\right) - 2\mathbf{Y}'\mathbf{M}_\mathbf{Z}\mathbf{Y}\boldsymbol{\gamma}\left(\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_1\mathbf{Y}\boldsymbol{\gamma}\right) = \mathbf{0}.$$

Dividing both sides by $2\left(\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_\mathbf{Z}\mathbf{Y}\boldsymbol{\gamma}\right)$, we have

$$\mathbf{Y}'\mathbf{M}_1\mathbf{Y}\boldsymbol{\gamma} - \kappa\mathbf{Y}'\mathbf{M}_\mathbf{Z}\mathbf{Y}\boldsymbol{\gamma} = \mathbf{0}, \tag{35}$$

or

$$\left(\mathbf{W}_1 - \kappa\mathbf{W}\right)\boldsymbol{\gamma} = \mathbf{0},$$

where $\mathbf{W}_1$ and $\mathbf{W}$ are defined in the main text. In other words, $\widehat{\kappa}$ is the smallest eigenvalue of $\mathbf{W}_1\mathbf{W}^{-1}$, and $\widehat{\boldsymbol{\gamma}}$ is the corresponding eigenvector. To find $\widehat{\boldsymbol{\gamma}}$ or $\widehat{\boldsymbol{\beta}}_2$, expanding (35) as

$$\left[\begin{pmatrix} \mathbf{y}'\mathbf{M}_1\mathbf{y} & \mathbf{y}'\mathbf{M}_1\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{M}_1\mathbf{y} & \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2 \end{pmatrix} - \widehat{\kappa}\begin{pmatrix} \mathbf{y}'\mathbf{M}_\mathbf{Z}\mathbf{y} & \mathbf{y}'\mathbf{M}_\mathbf{Z}\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{M}_\mathbf{Z}\mathbf{y} & \mathbf{X}_2'\mathbf{M}_\mathbf{Z}\mathbf{X}_2 \end{pmatrix}\right]\begin{pmatrix} 1 \\ -\widehat{\boldsymbol{\beta}}_2 \end{pmatrix} = \mathbf{0}.$$

When the rows corresponding to $\mathbf{X}_2$ are multiplied out, this becomes

$$\mathbf{X}_2'\left(\mathbf{M}_1 - \widehat{\kappa}\mathbf{M}_\mathbf{Z}\right)\mathbf{y} - \mathbf{X}_2'\left(\mathbf{M}_1 - \widehat{\kappa}\mathbf{M}_\mathbf{Z}\right)\mathbf{X}_2\widehat{\boldsymbol{\beta}}_2 = \mathbf{0},$$

which implies

$$\widehat{\boldsymbol{\beta}}_2 = \left(\mathbf{X}_2'\left(\mathbf{M}_1 - \widehat{\kappa}\mathbf{M}_\mathbf{Z}\right)\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\left(\mathbf{M}_1 - \widehat{\kappa}\mathbf{M}_\mathbf{Z}\right)\mathbf{y}.$$

Given $\widehat{\boldsymbol{\beta}}_2$, $\widehat{\boldsymbol{\beta}}_1$ can be obtained by regressing $\mathbf{y} - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2$ on $\mathbf{X}_1$. Combining the formulas for $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$, we can show (28).