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 20 of Ruud (2000), Chapter 4 of Cameron and Trivedi (2005), Chapter 5 of Wooldrige (2010), and Chapter 12 of Hansen (2022).

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. Note here that $\boldsymbol{\beta}$ in (1) is the structural parameter rather than the linear projection coefficient of y on *span* $(\mathbf{x})$ since from Chapter 2 we can always find a $\boldsymbol{\beta}$ such that $E[\mathbf{x}_i u_i] = \mathbf{0}$. 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, do higher hotel prices decrease occupancy rates? 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 *randomizied controlled trial* (RCT)[2] in which the reverse causality channel is nullified (see references cited in the Introduction). The first solution will be discussed in this chapter.

*Email: pingyu@hku.hk

[1] See Stock and Trebbi (2003) for who invented instrumental variable regression. The current evidence shows that both the father Philip Green Wright and his oldest son Sewall Green Wright contributed to this remarkable identification idea. Philip Wright (1861–1934) was an American mathematician and economist at Lombard College (now defunct). He was also a special expert at the United States Tariff Commission. Sewall Wright (1889-1988) was an American geneticist known for his influential work on evolutionary theory and also for his invention of path analysis (a key step for causal effects evaluation). He worked at the Department of Zoology at the University of Chicago until retirement in 1955 and then moved to the University of Wisconsin–Madison.

[2] Banerjee, Abhijit (1961-, MIT), Esther Duflo (1972-, MIT) and Michael Kremer (1964-, Harvard) won the Nobel Prize in 2019 because they successfully applies RCT to improve our ability to fight global poverty.

(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 two chapters in Handbook of Econometris, Chamberlain (1984) and Arellano and Honoré (2001), and some popular books, Diggle et al. (2002), Arellano (2003), Pesaran (2015), Baltagi (2021), and Hsiao (2022), for an introduction to panel data analysis), and using RCTs.

(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 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, i.e., repeated measurements) or auxiliary data-set that contains information about the conditional distribution of the true variables given the mismeasured variables. See the chapter in Handbook of Econometrics, Bound et al. (2001), and Chen et al. (2011) for an introduction to measurement errors in survey data. A review of the commonly used techniques in statistics can be found in Fuller (1987) and Carroll et al. (2006).

(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; see Winship and Mare (1992), Vella (1998) and Heckman (2008) for an introduction.

(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) *Philip 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*

$$
\begin{aligned}
p_i &= \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1}, \\
q_i &= \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 v_i - \beta_1 u_i}{\alpha_1 - \beta_1},
\end{aligned}
\tag{4}
$$

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

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

where the last equality is from substituting in the formulas of p_i and q_i in (4), 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.*[3] *Figure 1 illustrates the discussion above intuitively. The identification problem in simultaneous equations dates back to Philip Wright (1915) and Working (1927).*

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

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

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 implicitly defined in Appendix B of Philip (1928). Another method to

[3] Note that the supply curve is still not identifiable because essentially, only one point on the supply curve is observed.

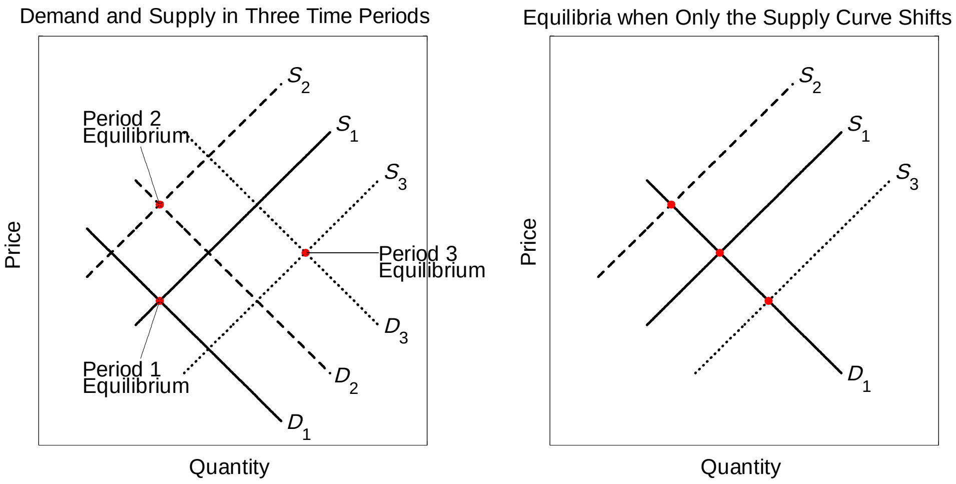

Figure 1: Endogeneity and Identification by Instrument Variables

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 10 below.* □

Example 2 (Omitted Variables) *Mundlak (1961)*[4] *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 by the agent on the basis of such factors.*

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

$$Q_i = A_i \cdot (L_i)^{\phi_1} \cdot \exp(\nu_i),\ 0 < \phi_1 < 1, \tag{5}$$

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*

[4] Yair Mundlak (1927-2015) was an agricultural economist at the Economics Department of the University of Chicago. He contributed significantly to the early literature on the econometrics of panel data. His 1978 *Econometrica* paper is cited in all relevant textbooks and in close to 1,000 articles.

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

$$L_i = \left(\frac{w}{p}\right)^{\frac{1}{\phi_1 - 1}} (A_i B \phi_1)^{\frac{1}{1-\phi_1}}\,, \tag{6}$$

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 (5), 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 (6). Figure 2 shows the effect of $\log A_i$ on ϕ_1 by drawing $E\left[\log Q \middle| \log L, \log A\right]$ 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 $A_i = \exp(\phi_0 + u_i)$. (5) and (6) can be written as

$$\log Q_i = \phi_0 + \phi_1 \cdot \log(L_i) + \nu_i + u_i, \tag{7}$$

$$\log L_i = \beta_0 + \frac{1}{1-\phi_1} u_i, \tag{8}$$

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 (8) into (7), 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 (6) change? Show that $\log Q_i$ and $\log(L_i)$ are perfectly correlated.*

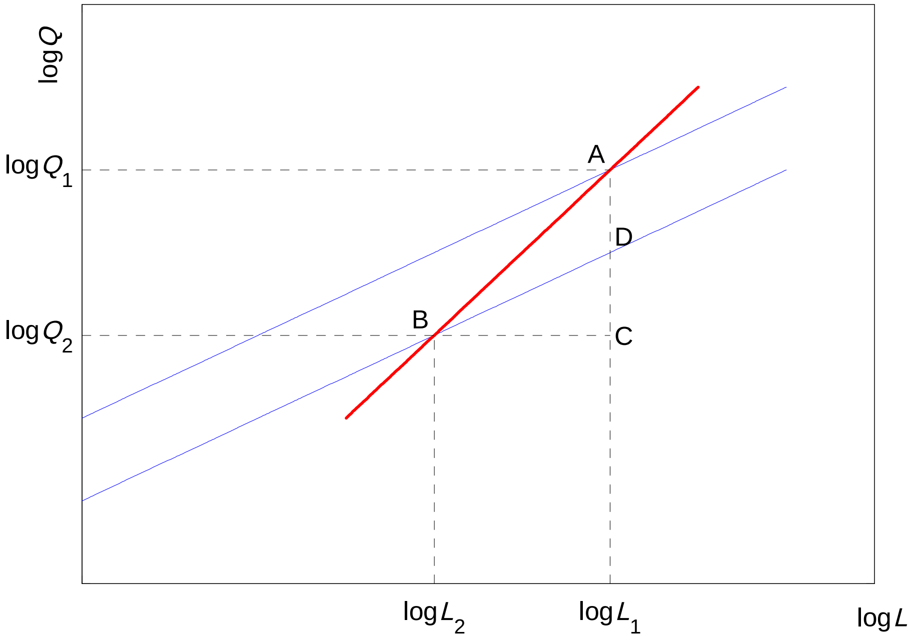

Figure 2: Effect of Soil Quality on Labor Input

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

$$C_i = kY_i + u_i \text{ with } u_i = c_i - ky_i. \tag{9}$$

It is easy to see that $E\left[Y_i u_i\right] = -kE\left[y_i^2\right] < 0$, so the LSE of k converges to $\frac{E[Y_iC_i]}{E[Y_i^2]} = \frac{kE\left[\left(Y_i^*\right)^2\right]}{E\left[\left(Y_i^*\right)^2\right]+E\left[y_i^2\right]} < k$, that is, the OLS using measured data underestimates k. Essentially, the effect of the measurement error in the covariate ends up being "netted out" in the error term, so the mismeasured covariate is negatively correlated with the error term, which makes the OLS underestimate the slope. More intuitively, with measurement error, the covariate is more spreading, which makes the OLS estimate of slope smaller. Taking expectation on both sides of (9), 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 $E\left[Y_i^*\right] = E[Y_i] \neq 0$.. The IV estimation using z as the *instrument is* $\frac{\overline{C}}{\overline{Y}}$*, which is how Friedman estimated* k*.*

Actually, measurement errors are embodied in regression analysis from the beginning. Galton (1889)[5] *analyzed the relationship between the height of sons and the height of fathers. Specifically, suppose the true model is*

$$S_i^* = \alpha + \beta F_i^* + u_i, \tag{10}$$

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^* *and* F_i^**. Suppose the observables are* $S_i = S_i^* + s_i$*, and* $F_i = F_i^* + f_i$*, where* s_i *and* f_i *are the mean-zero environmental factors; then our regression becomes*

$$S_i = \alpha + \beta\left(F_i - f_i\right) + s_i = \alpha + \beta F_i + s_i - \beta f_i.$$

The OLS estimator of β *will converge to* $\frac{Cov(F_i,S_i)}{Var(F_i)} = \frac{Var(F_i^*)}{Var(F_i^*)+Var(f_i)} < 1$*, where* $\frac{Var(F_i^*)}{Var(F_i^*)+Var(f_i)} \equiv \rho$ *is called the* reliability coefficient *or* Heritability coefficient.[6] *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".*

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{11}$$

Any solution to the problem of endogeneity requires additional information which we call *instrumental variables* (or simply *instruments*). The label "instrumental variables" was introduced by Reiersøl (1945). The $l \times 1$ random vector $\mathbf{z}_i$ is an instrument for (1) if $E\left[\mathbf{z}_i u_i\right] = \mathbf{0}$. 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

[5] Sir Francis Galton (1822-1911) was an English statistician. In addition to inventing the concept of regression, he is credited with introducing the concepts of correlation, the standard deviation, and the bivariate normal distribution. He was Charles Darwin (1809-1882)'s half-cousin, sharing the common grandparent Erasmus Darwin, and was also the advisor of Karl Pearson and the inventor of fingerprinting.

[6] Since S_i and F_i have the same distribution, the LSE converges to $\frac{Cov(S_i,F_i)}{Var(F_i)} = \frac{Cov(S_i,F_i)}{\sqrt{Var(F_i)Var(S_i)}}$ which is the correlation between S_i and F_i and is less than 1.

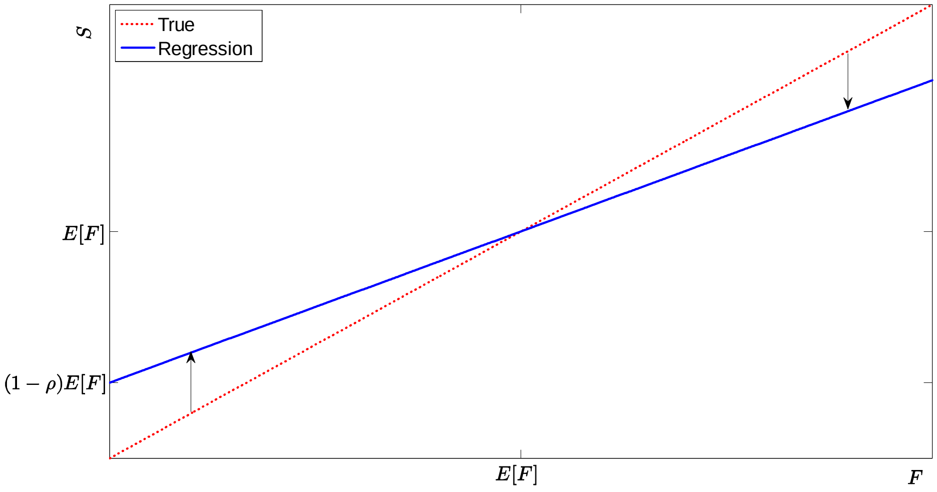

Figure 3: Relationship Between the Height of Sons and Fathers

intercept). Thus we make the partition

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{x}_{2i} \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix}, \tag{12}$$

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

$$\mathbf{z}_i = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{z}_{2i} \end{pmatrix} \begin{matrix} k_1 \\ l_2 \end{matrix}, \tag{13}$$

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

$$\mathbf{x}_i = \mathbf{\Gamma}'\mathbf{z}_i + \mathbf{v}_i. \tag{14}$$

In matrix notation,

$$\mathbf{X} = \mathbf{Z\Gamma} + \mathbf{V}, \tag{15}$$

where $\mathbf{V}$ is a $n \times k$ matrix. By construction, $E\left[\mathbf{z}_i\mathbf{v}_i'\right] = \mathbf{0}$, so (14) 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 (15) into (11), we find

$$\mathbf{y} = \left(\mathbf{Z\Gamma} + \mathbf{V}\right)\boldsymbol{\beta} + \mathbf{u} = \mathbf{Z}\boldsymbol{\lambda} + \mathbf{e} \tag{16}$$

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

$$E\left[\mathbf{z}e\right] = E\left[\mathbf{z}\mathbf{v}'\right]\boldsymbol{\beta} + E\left[\mathbf{z}u\right] = \mathbf{0}. \tag{17}$$

Thus (16) 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 (16) is the reduced form for $\mathbf{y}$. (15) and (16) together are the reduced form equations for the system

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

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

The system of equations

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

are called *triangular* (or *recursive*) *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.

To see how a recursive model can originate in practice, consider the return-to-schooling example again. Our statement of the example follows from Imbens and Newey (2009). Let Y denote individual lifetime earnings and X denote level of education, where we use the capital letters such as X to denote random variables and the corresponding lower case letters such as x denote the potential values they may take. The value of X is chosen first, as a function of expected but not of realized Y. The value of Y is determined next, as a function of X, as well as of other observable and unobservable variables. In the simple version of such a model, the typical individual's lifetime earnings are a function of the level of education and productivity (or ability), U. Although productivity is unobserved to the agent at the time of deciding the length of education, the value of a signal, V, correlated with productivity, U, is known to the agent at such time. A measure, Z, determining the cost of a unit of education is observed as well. Denote the cost of education by the value of a function $c(X, Z)$. The agent chooses X to maximize expected lifetime earnings given the signal, V, and the cost of education:

$$X = \arg\max_{x} \left\{E\left[m_1(x, U)\,|Z, V\right] - c(x, Z)\right\}.$$

The solution is then a function, m_2, of Z and V. The recursive model becomes

$$\begin{aligned} Y &= m_1(X, U), \\ X &= m_2(Z, V). \end{aligned} \tag{18}$$

If m_1 and m_2 are linear, then we obtain the recursive linear model above. If X is an input and Y is output, then this example is related to Example 2 in Section 1; see Mundlak (1963) for more discussion.

Exercise 2 *(i) Show that $E[\mathbf{v}_i u_i] \neq \mathbf{0}$. (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 $\mathbf{z}_i$ (14), 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

$$E\left[\mathbf{z}_i y_i\right] = E\left[\mathbf{z}_i \mathbf{x}_i'\right]\boldsymbol{\beta}. \tag{19}$$

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*[7]

$$\operatorname{rank}(\boldsymbol{\Gamma}) = k \tag{20}$$

holds. Intuitively, the rank condition requires that $\mathbf{z}$ can perturb $\mathbf{x}$ in all directions. This condition can be tested (think about the case $k = 1$); see, e.g., Cragg and Donald (1993, 1996, 1997), Robin and Smith (2000), Kleibergen and Paap (2006) and Chen and Fang (2019). If $\operatorname{rank}(E\left[\mathbf{z}_i\mathbf{z}_i'\right]) = l$ (this is trivial), and $\operatorname{rank}(E\left[\mathbf{z}_i\mathbf{x}_i'\right]) = k$ (this is crucial), this condition is satisfied. Assume that (20) 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 (20) is not satisfied, then $\boldsymbol{\beta}$ cannot be uniquely recovered from $(\boldsymbol{\lambda}, \boldsymbol{\Gamma})$. Note that a necessary (although not sufficient) condition for (20) 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 (12) and (13) 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

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

(15) 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 $\operatorname{rank}(\boldsymbol{\Gamma}) = k$, which is true if and only if $\operatorname{rank}(\boldsymbol{\Gamma}_{22}) = k_2$ (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 $\boldsymbol{\Gamma}_{22}$. Alternatively, rewrite (16) as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\lambda}_1 + \mathbf{Z}_2\boldsymbol{\lambda}_2 + \mathbf{e},$$

where $\boldsymbol{\lambda}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}_{12}\boldsymbol{\beta}_2$ and $\boldsymbol{\lambda}_2 = \boldsymbol{\Gamma}_{22}\boldsymbol{\beta}_2$. So $\boldsymbol{\beta}_2$ can be identified if and only if $\operatorname{rank}(\boldsymbol{\Gamma}_{22}) = k_2$. If $\boldsymbol{\beta}_2$ can be identified, then $\boldsymbol{\beta}_1$ can be identified as $\boldsymbol{\beta}_1 = \boldsymbol{\lambda}_1 - \boldsymbol{\Gamma}_{12}\boldsymbol{\beta}_2$. Thus the key identification condition is indeed $\operatorname{rank}(\boldsymbol{\Gamma}_{22}) = k_2$.

[7] The precise condition should be $\operatorname{rank}([\lambda, \Gamma]) = \operatorname{rank}(\Gamma) = k$. 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.

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*

$$\text{(i) uncorrelated with } u\text{; (ii) correlated with endogenous variables.}^{8} \tag{21}$$

(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 instrument relevance condition which says that $\mathbf{X}_1$ and the predicted value of $\mathbf{X}_2$ (from the regression of $\mathbf{X}_2$ on $\mathbf{Z}_2$ 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 (21) 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.[9] *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*

$$\begin{aligned}E[zu] &= 0, \qquad (22)\\ x_2 &= x_1\gamma_1+z\gamma_2+v \text{ with } \gamma_2\neq 0.\end{aligned}$$

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! Actually, from Exercise 6, γ_2 can be identified as 0.

The arguments above indicate that (21) is not sufficient. Is it necessary? The answer is still NO! The question can be formulated as follows: in (22), can we find some z such that

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

[8] Of course, we also require the instrument to be excluded from the outcome equation.

[9] WLOG, assume $E[x_1]=E[\varepsilon]=0$ so that $E[zu]=Cov(z,u)$.

*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 (need not be the same one) instrument appears in each of the first-stage regression. Here "appear" means the coefficient of the variable is not zero. When $k=l$, each instrument must appear in at least one endogenous regression. □

Exercise 6 *Show that the true value of γ_2 in (22) is zero.*

Given the cautions in the above example, how to select instruments? Generally speaking, good instruments are not selected based on mathematics, but based on economic theory.[10] In the return to schooling example, the usual practice in the literature is to seek instruments which proxy, or are correlated with, costs of schooling. 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;[11] Butcher and Case (1994) use the sex of siblings, in particular whether a girl has any sisters, as an IV to estimate the schooling return to women because the gender of siblings may affect the cost of investing in a child's human capital through the existence of borrowing constraints if there are exogenous gender differences in the return to human capital;[12] Card (1995) uses college proximity 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 college but is unlikely to directly affect the wages earned in their adulthood; Blundell et al. (2005) use three instruments - a dummy variable for whether the parents reported an adverse financial shock at either age 11 or age 16 of the child, a dummy variable for whether the child's teacher ranked the parent's "interest in education" high or low when the child was 7, and

[10] Murray (2006) provides nine strategies to justify the validity of an instrument.

[11] Children born earlier in the year enter school at an older age (children turning six by January 1 can enter the primary school on September 1) and are therefore allowed to drop out (on their 16th or 17th birthday) after having completed less schooling than children born later in the year. The exclusion condition may fail because children born in the first quarter are a few months older than other children, and at vey young ages a difference of a few months might be an advantage in performance in school. This indicates that the estimator based on this IV may underestimate the return to schooling.

[12] According to Butcher and Case, in the presence of borrowing constraints and assuming that boys receive a higher return to each level of schooling, "we should expect to see not only that boys receive more education, but also that the presence of sons reduces the educational attainment of daughters." They also considered other justifications of this IV, e.g., the costs of raising girls are different from boys. On the contrary, they find that girls who have any sisters, conditional on the number of siblings, have lower school attainment than do girls with no sisters; on the other hand, the school attainment of boys is found to be unrelated to gender composition. This may be because parents prefer a "gender mix".

the number of older siblings of the child[13].[14] In development economics, 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. In political economics, Levitt (1997) uses the timing of mayoral and gubernatorial elections as an IV to identify the causal effect of police on crime by arguing that after controlling some economic variables such as state unemployment rates and spending on public welfare or education this IV does not affect the crime rate but will affect the number of police officers.

As emphasized in Deaton (2010)[15], exogeneity is different from externality. Exogeneity of a variable means that it is orthogonal to the error term, while externality of a variable means that it is not set or caused by the variables in the model. The former is not guaranteed by the latter. The instruments mentioned above are external, but their exogeneity still need careful arguments.

Exercise 7 *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 Tinbergen[16] (1930) and Haavelmo[17] (1943):

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

[13] This variable can take three values in the sample, 0, 1 and 2. It is argued that number of *older* siblings is more relevant to schooling than number of *total* siblings.

[14] Other popular instruments in the return to schooling analysis include the local unemployment rate at age 18, a state-level tuition variable, the dummy for residence in an urban area at age 18, parental education (e.g., dummies for attending college or not), etc.

[15] Angus Deaton (1945-) is a British-American economist, Dwight D. Eisenhower professor at Princeton University, and 2015 Nobel Prize winner. He is best known for his work on consumption theory, welfare and inequality. In the consumption theory, Deaton and Muellbauer (1980) developed the Almost Ideal Demand System (AIDS).

[16] Jan Tinbergen (1903-1994) was a Dutch economist who shared the first Nobel Prize in Economics with Ragnar Frisch in 1969. He is widely considered to be one of the most influential economists of the 20th century and one of the founding fathers of econometrics. It has been argued that the development of the first macro econometric models, the solution of the identification problem, and the understanding of dynamic models are his three most important legacies to econometrics.

[17] Trygve M. Haavelmo (1911-1999) was a Professor at the University of Oslo. He won the Nobel prize in 1989 "for his clarification of the probability theory foundations of econometrics and his analyses of simultaneous economic structures" especially in Haavelmo (1943, 1944)

Exercise 8 *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 9 *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 $E[\mathbf{z}u] = \mathbf{0}$:

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

where $E^*\left[\mathbf{x}|\mathbf{z}\right]$ is the linear projection of $\mathbf{x}$ on $\mathbf{z}$. Replacing population expectations with sample averages in (23) yields

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

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

Exercise 10 *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 (17),

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

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

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \left(\widehat{\mathbf{X}}'\widehat{\mathbf{X}}\right)^{-1}\widehat{\mathbf{X}}'\mathbf{y} = \left(\mathbf{X}'\mathbf{PX}\right)^{-1}\left(\mathbf{X}'\mathbf{Py}\right). \tag{24}$$

Exercise 11 *Show that $\widehat{\boldsymbol{\beta}}_{2SLS}$ satisfies*

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

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.[18] It is motivated by observing that $E[\mathbf{z}u]=\mathbf{0}$ implies

$$0=E^*\left[u|\mathbf{z}\right]=E^*\left[y|\mathbf{z}\right]-E^*[\mathbf{x}|\mathbf{z}]'\boldsymbol{\beta}, \tag{25}$$

so

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

Equivalently, $\widehat{\boldsymbol{\beta}}_{2SLS}=\arg\min_{\boldsymbol{\beta}}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\mathbf{P}_{\mathbf{Z}}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)$, which is a GLS estimator. Intuitively, $\mathbf{P}_{\mathbf{Z}}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)$ should converge in probability to zero because $E[\mathbf{z}u]=\mathbf{0}$, so we try to find some $\boldsymbol{\beta}$ value such that the length of $\mathbf{P}_{\mathbf{Z}}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)$ is as close to zero as possible. Another algebraically equivalent formulation of 2SLS is the control function formulation of Telser (1964)[19]:

$$\left(\begin{array}{c}\widehat{\boldsymbol{\beta}}_{2SLS}\\ \widehat{\boldsymbol{\rho}}_{2SLS}\end{array}\right)=\left(\widehat{\mathbf{W}}'\widehat{\mathbf{W}}\right)^{-1}\widehat{\mathbf{W}}'\mathbf{y}, \tag{26}$$

where $\widehat{\mathbf{W}}=[\mathbf{X},\widehat{\mathbf{V}}]$. This estimator is the OLS estimator of y on $\mathbf{x}$ and $\widehat{\mathbf{v}}$. This construction exploits another implication of $E[\mathbf{z}u]=\mathbf{0}$:

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

for some coefficient vector $\boldsymbol{\rho}$, where the third equality follows from the orthogonality of both error terms u and $\mathbf{v}$ with $\mathbf{z}$. So

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

Thus, this particular linear combination of the first-stage errors $\mathbf{v}$ is a function that controls for the endogeneity of the regressors $\mathbf{x}$; one can think of $\mathbf{v}$ as proxying for the factors in u that are

[18] Robert L. Basmann (1926-) is an American econometrician. He earned a Ph.D. in economics from Iowa State University in 1955, and was a Professor of Econometrics at Texas A&M University until his retirement. He served as a lecturer at Binghamton University after his retirement.

[19] 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). Lester G. Telser (1931-) is an American economist and Professor Emeritus in Economics at the University of Chicago. He received his Ph.D. from the University of Chicago in 1956 under the supervision of Milton Friedman. His first name is an anagram of his surname.

correlated with $\mathbf{x}$. From the FWL theorem, $\widehat{\boldsymbol{\beta}}_{2SLS}$ is the effect of the net variation in $\mathbf{x}$ on y after excluding the variation in $\widehat{\mathbf{v}}$, while the net variation in $\mathbf{x}$ comes from $\mathbf{z}$ because $\mathbf{x} = \widehat{\boldsymbol{\Gamma}}'\mathbf{z} + \widehat{\mathbf{v}}$. Basmann (1957)'s formulation (25) 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 12 *(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 (26) generates the same formula of $\widehat{\boldsymbol{\beta}}_{2SLS}$ as (24).*

Exercise 13 *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 $E[x_iu_i] = 0$ and $E[x_i^2u_i] = 0$. 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 14 *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) *Is $\widehat{\boldsymbol{\beta}}_{2SLS}$ is the best linear predictor of $m(\mathbf{x})$ in the sense that*

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

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

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

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

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

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. Second, $\widehat{\mathbf{v}}$. In the control function formulation of 2SLS, only $\widehat{\mathbf{v}}_2 = \mathbf{x}_2 - \widehat{\mathbf{x}}_2$ should be added to the regression since $\widehat{\mathbf{v}}_1 = \mathbf{x}_1 - \widehat{\mathbf{x}}_1 = \mathbf{x}_1 - \mathbf{x}_1 = \mathbf{0}$ (otherwise, the multicollinearity problem would happen). $\mathbf{x}_2 = \boldsymbol{\Gamma}_{12}'\mathbf{z}_1 + \boldsymbol{\Gamma}_{22}'\mathbf{z}_2 + \mathbf{v}_2$ implies $\mathbf{v}_2 = \mathbf{x}_2 - \boldsymbol{\Gamma}_{12}'\mathbf{z}_1 - \boldsymbol{\Gamma}_{22}'\mathbf{z}_2$, so the rank condition that rank$(\boldsymbol{\Gamma}_{22}) = k_2$ guarantees that there is separate variation in $\mathbf{v}_2$ from $\mathbf{x} = (\mathbf{z}_1', \mathbf{x}_2')'$ in the regression of y on $\mathbf{x}$ and $\mathbf{v}_2$.

Exercise 15 *In the structural model*

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

(i) show that $\widehat{\boldsymbol{\beta}}_{2,2SLS} = \left(\mathbf{X}_2'\mathbf{P}_{\mathbf{Z}_2^{\perp}}\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{P}_{\mathbf{Z}_2^{\perp}}\mathbf{y}$*, where* $\mathbf{Z}_2^{\perp} = \mathbf{M}_{\mathbf{X}_1}\mathbf{Z}_2$*; (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* $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{2SLS}$*; (iii) Does* $\widehat{\boldsymbol{\beta}}_{1,2SLS}$ *equal* $\widehat{\boldsymbol{\beta}}_{1,OLS}$*?*

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

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

The identification conditions are

$$
Cov(z,x) \neq 0, Cov(z,u) = 0.^{20} \tag{27}
$$

From Exercise 15, the IV estimator is

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

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

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

[20] 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 so that $\gamma_1 = Cov(z,x)/Var(z)$; as a result, $Cov(z,x) \neq 0$ is not only necessary but also sufficient for identification.

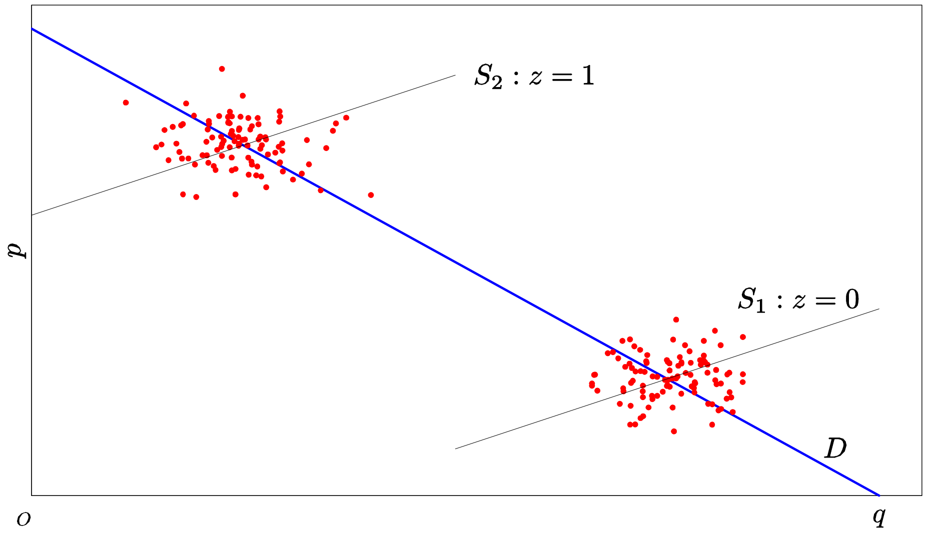

Figure 4: Intuition for the Wald Estimator in the Linear Demand/Supply System

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*

$$\frac{E\left[y|z=1\right]-E\left[y|z=0\right]}{E\left[x|z=1\right]-E\left[x|z=0\right]}. \tag{28}$$

Note that the numerator and denominator of (28) are exactly the slope coefficients in the reduced form equations:

$$\begin{aligned} y &= \lambda_0+\lambda_1 z+e, \\ x &= \gamma_0+\gamma_1 z+v, \end{aligned}$$

so the form of $\widehat{\beta}_{Wald}$ *is a direct application of ILS. 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, Giles et al. (2003) 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.[21] 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.[22]

Exercise 16 *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?

Exercise 17 (*) *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.)

[21] See Heckman (1997) for a critique on the validity of this instrument. Suppose $z \neq x$ is because $x = 0$ although $z = 1$, i.e., draft evaders ($x = 1$ while $z = 0$, the volunteers, seem fine with exclusion although they may anticipate high earnings gains from military service). If this is for medical reasons, or more generally reasons that make these candidates ineligible to serve, then the exclusion assumption seems plausible. If, on the other hand these are individuals fit but unwilling to serve, they may have had to take actions to stay out of the military that could have affected their subsequent civilian labor market careers. Such actions may include extending their educational career, or temporarily leaving the country. Note that these issues are not addressed by the random assignment of the instrument.

[22] All the samples bought lottery. x is the magnitude of the prize, while z is the indicator for winning the prize, so $E\left[x|z=0\right] = 0$. y is the behavior (e.g., consumption) change before and after winning the prize. Imbens et al. essentially assume the exogeneity of x, so z is not really required.

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 in a nonseparable model?

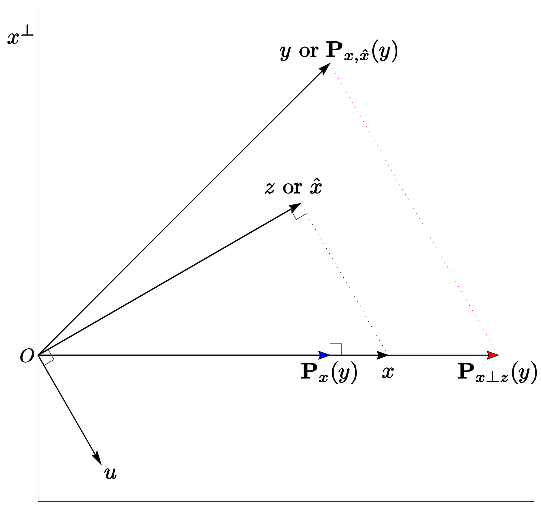
$x^{\perp}$
y or $\mathbf{P}_{x,\hat{x}}(y)$
z or $\hat{x}$
O
$\mathbf{P}_x(y)$
x
$\mathbf{P}_{x\perp z}(y)$
u

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 $\mathbf{P}_{\mathbf{X}\perp\mathbf{Z}}(\mathbf{y})$ in Section 4.1 of Chapter 2). Since $z \perp u$, this is also the projection of y onto $span(x)$ along u if $\dim\left(span^{\perp}(z)\right) = 1$ as in the figure. 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 stay 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? (*)

In the linear model, the IV estimator is estimating $\boldsymbol{\beta}$, the *constant* effect of $\mathbf{x}$ on y. In a generally nonseparable model (e.g., the equations (18), or $\mathbf{x}$ and y are both binary),

$$\begin{aligned} y &= m(\mathbf{x}, \mathbf{u}), \\ \mathbf{x} &= h(\mathbf{z}, \mathbf{v}), \end{aligned}$$

the effect of $\mathbf{x}$ on y is heterogenous. 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 (28)

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

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,

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

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

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

In the table, we exclude the possibility of defiers. This is the *monotonicity* assumption in Imbens and Angrist (1994).[23] The denominator of (28)

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

so we have

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

In summary, if we interpret z as a random assignment and x as the realized treatment status due to imperfect compliance, then $\text{plim}\left(\widehat{\beta}_1\right)$ is the **intention-to-treat** (ITT) effect $E\left[y|z = 1\right] - E\left[y|z = 0\right]$ scaled by the proportion of individuals that are induced to change their treatment status through the intented assignment.

$\text{plim}\left(\widehat{\beta}_1\right) = E\left[(y_1 - y_0) | S_1 - S_0 = 1\right]$ is called the **local average treatment effect** (LATE) 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**. 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 $J \geq 2$ levels, Angrist and Imbens (1995) show that

$$\widehat{\beta}_1 \xrightarrow{p} \sum_{j=1}^{J} \omega_j \cdot E\left[y_j - y_{j-1} | S_1 \geq j > S_0\right] \equiv \beta \tag{29}$$

where $\omega_j = \frac{P(S_1 \geq j > S_0)}{\sum_{j=1}^{J} P(S_1 \geq j > S_0)} = \frac{P(S<j|z=0) - P(S<j|z=1)}{E[S|z=1] - E[S|z=0]}$ is the impact of z on the cdf of S at the level j. That is, β_1 is a weighted average of per-unit treatment effect. For the case with continuous S,

[23] 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.

see Angrist et al. (2000).

Exercise 18 *Show (29) in the case of $J = 2$.*

Exercise 19 *$S_1 - S_0$ is called the* compliance intensity *which need not be 1, so in the expression of β in (29), the sets of compliers $\{S_1 \geq j > S_0\}_{j=1}^{J}$ are overlapping subpopulations. Denote the complier subpopulation with $S_0 = k$ and $S_1 = l$ as $\mathcal{C}_{kl}$, where $k < l$, $k, l = 0, 1, \cdots, J$. Show that β can be re-expressed in terms of nonoverlapping subpopulations as*

$$\beta = \sum_{k=0}^{J-1} \sum_{l>k} w_{kl} \cdot E\left[y_l - y_k | \mathcal{C}_{kl}\right],$$

where $w_{kl} = \frac{P(\mathcal{C}_{kl})}{\sum_{k=0}^{J-1}\sum_{l>k}(l-k)P(\mathcal{C}_{kl})}$.

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),[24] 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 Sections 2, 3 and 4. The notations there can be applied in this case. Define

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

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

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

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

[24] Theodore (Ted) Anderson (1918-2016) was a American statistician and econometrician, who made fundamental contributions to multivariate statistical theory. Important contributions include the Anderson-Darling distribution test, the Anderson-Rubin statistic, the method of reduced rank regression, and his most famous econometrics contribution – the LIML estimator. He continued working throughout his long life, even publishing theoretical work at the age of 97!

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

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

where

$$\kappa\left(\boldsymbol{\beta}_2\right) = \frac{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_1\mathbf{Y}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\boldsymbol{\gamma}} = \frac{\left(\mathbf{y}^{\perp} - \mathbf{X}_2^{\perp}\boldsymbol{\beta}_2\right)'\left(\mathbf{y}^{\perp} - \mathbf{X}_2^{\perp}\boldsymbol{\beta}_2\right)}{\left(\mathbf{y}^{\perp} - \mathbf{X}_2^{\perp}\boldsymbol{\beta}_2\right)'\mathbf{M}_{\mathbf{Z}_2^{\perp}}\left(\mathbf{y}^{\perp} - \mathbf{X}_2^{\perp}\boldsymbol{\beta}_2\right)} \tag{30}$$

with $\boldsymbol{\gamma} = \left(1, -\boldsymbol{\beta}_2'\right)'$, $\mathbf{Y} = [\mathbf{y}, \mathbf{X}_2]$, $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'$ and for any random matrix $\mathbf{A}$, $\mathbf{A}^{\perp} = \mathbf{M}_1\mathbf{A}$. Maximizing $\ell_n\left(\boldsymbol{\beta}_2\right)$ is equivalent to minimizing $\kappa\left(\boldsymbol{\beta}_2\right)$. 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 $\kappa\left(\boldsymbol{\beta}_2\right)$ 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 (30) 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 Appendix A):

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LIML}} = \left(\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2'\right)' = \left[\mathbf{X}'\left(\mathbf{I}_n - \widehat{\kappa}\mathbf{M}_{\mathbf{Z}}\right)\mathbf{X}\right]^{-1}\mathbf{X}'\left(\mathbf{I}_n - \widehat{\kappa}\mathbf{M}_{\mathbf{Z}}\right)\mathbf{y}, \tag{31}$$

where $\widehat{\kappa}$ is the smallest characteristic root of $\mathbf{W}_1\mathbf{W}^{-1}$ or $\mathbf{W}^{-1/2}\mathbf{W}_1\mathbf{W}^{-1/2}$ with

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

or the smallest root of the determinantal equation $|\mathbf{W}_1 - \kappa\mathbf{W}| = 0$. For inference, it is useful to observe that (31) shows that $\widehat{\boldsymbol{\beta}}_{\mathrm{LIML}}$ can be written as an IV estimator

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LIML}} = \left(\widetilde{\mathbf{X}}'\mathbf{X}\right)^{-1}\widetilde{\mathbf{X}}'\mathbf{y} \tag{32}$$

using the instrument

$$\widetilde{\mathbf{X}} = \left(\mathbf{I}_n - \widehat{\kappa}\mathbf{M}_{\mathbf{Z}}\right)\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 - \widehat{\kappa}\widehat{\mathbf{V}}_2 \end{pmatrix},$$

where $\widehat{\mathbf{V}}_2 = \mathbf{M}_{\mathbf{Z}}\mathbf{X}_2$ is the reduced-form residuals from regressing $\mathbf{x}_{2i}$ on $\mathbf{z}_i$. Expressing LIML using this IV formula is useful for variance estimation.

In Appendix B, we show that the LIML and 2SLS have the same asymptotic distribution, which holds under the same assumptions as for 2SLS, and in particular does not require normality of the errors. Consequently, one method to obtain an asymptotically valid covariance estimate for LIML is to use the same formula as for 2SLS. However, this is not the best choice. Rather, using the IV representation for LIML (32), we can estimate the asymptotic covariance matrix by

$$\widehat{\mathbf{V}} = \left(\frac{1}{n}\widetilde{\mathbf{X}}'\mathbf{X}\right)^{-1} \widehat{\mathbf{\Omega}} \left(\frac{1}{n}\mathbf{X}'\widetilde{\mathbf{X}}\right)^{-1}, \tag{33}$$

where

$$\widehat{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i' \widehat{u}_i^2$$

with $\widetilde{\mathbf{x}}_i$ being the ith row of $\widetilde{\mathbf{X}}$ and $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{\text{LIML}}$. This simplifies to the 2SLS formula when $\widehat{\kappa} = 1$ but otherwise differs. The estimator (33) is a better choice than the 2SLS formula for covariancematrix estimation as it takes advantage of the LIML estimator structure. When the model is homoskedastic, we can replace $\widehat{\mathbf{V}}$ by

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

where $\widehat{\sigma}^2 = n^{-1}\left(\mathbf{y} - \mathbf{X}'\widehat{\boldsymbol{\beta}}_{\text{LIML}}\right)'\left(\mathbf{y} - \mathbf{X}'\widehat{\boldsymbol{\beta}}_{\text{LIML}}\right)$. 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).

Minimizing (30) with respect to $\boldsymbol{\gamma}$ is invariant to the scale of $\boldsymbol{\gamma}$, so we use a normalization $\overline{\boldsymbol{\gamma}} = (1, -\boldsymbol{\beta}_2')'$ above. An alternative normalization is to set $\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\boldsymbol{\gamma} = 1$. Using the second normalization,

$$\widehat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \frac{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_1\mathbf{Y}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\boldsymbol{\gamma}}$$

is the generalized eigenvector of $\mathbf{Y}'\mathbf{M}_1\mathbf{Y}$ with respect to $\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}$ associated with the smalled generalized eigenvalue.[25] Computationally this is straightforward. For example, in MATLAB, the generalized eigenvalues and eigenvectors of the matrix $\mathbf{A}$ with respect to $\mathbf{B}$ is found by the command $\mathsf{eig}(\mathbf{A}, \mathbf{B})$. Once this $\widehat{\boldsymbol{\gamma}}$ is found, any other normalization can be obtained by rescaling. For example, to obtain the MLE for $\boldsymbol{\beta}_2$ make the partition $\widehat{\boldsymbol{\gamma}}' = \left[\widehat{\gamma}_1, \widehat{\boldsymbol{\gamma}}_2'\right]$ and set $\widehat{\boldsymbol{\beta}}_2 = -\widehat{\boldsymbol{\gamma}}_2/\widehat{\gamma}_1$. To obtain the MLE for $\boldsymbol{\beta}_1$, recall the structural equation $y_i = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + u_i$. Replacing $\boldsymbol{\beta}_2$ with the MLE $\widehat{\boldsymbol{\beta}}_2$ and then apply regression. Thus

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

These solutions are the same as those in (31). Both kinds of solutions involve similar computations,

[25] Let $\mathbf{A}$ and $\mathbf{B}$ be $k \times k$ matrices. The generalized characteristic equation is $|\mathbf{A} - \mu\mathbf{B}| = 0$. The solutions μ are known as **generalized eigenvalues** of $\mathbf{A}$ with respect to $\mathbf{B}$. Associated with each generalized eigenvalue μ is a **generalized eigenvector** $\mathbf{v}$ which satisfies $\mathbf{Av} = \mathbf{Bv}\mu$. They are typically normalized so that $\mathbf{v}'\mathbf{Bv} = 1$ and thus $\mu = \mathbf{v}'\mathbf{Av}$.

e.g., calculate $\widehat{\kappa}$, but (31) is important for the distribution theory and to reveal the algebraic connection between LIML, leasts quares, and 2SLS.

If we want to test $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$, then notice that

$$\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_{20} = \mathbf{X}_1\boldsymbol{\lambda}_1^* + \mathbf{Z}_2\boldsymbol{\lambda}_2^* + \mathbf{e}^*,$$

where $\boldsymbol{\lambda}_1^* = \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}_{12}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_{20})$, $\boldsymbol{\lambda}_2^* = \boldsymbol{\Gamma}_{22}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_{20})$, and $\mathbf{e}^* = \mathbf{u} + \mathbf{V}_2(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_{20})$. Here, the exogenous variables excluded from the structural equation are added directly to the equation instead of being used to replace the endogenous explanatory varaibles by fitted values as in 2SLS. Now, testing $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$ is equivalent to test $\boldsymbol{\lambda}_2^* = \mathbf{0}$, and the resulting F-statistic

$$F(\boldsymbol{\beta}_{20}) = \frac{(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_{20})[\mathbf{M}_1 - \mathbf{M}_\mathbf{Z}](\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_{20})/l_2}{(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_{20})\mathbf{M}_\mathbf{Z}(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_{20})/(n-l)} = \frac{\left(\mathbf{y}^\perp - \mathbf{X}_2^\perp\boldsymbol{\beta}_{20}\right)'\mathbf{P}_{\mathbf{Z}_2^\perp}\left(\mathbf{y}^\perp - \mathbf{X}_2^\perp\boldsymbol{\beta}_{20}\right)/l_2}{\left(\mathbf{y}^\perp - \mathbf{X}_2^\perp\boldsymbol{\beta}_{20}\right)'\mathbf{M}_{\mathbf{Z}_2^\perp}\left(\mathbf{y}^\perp - \mathbf{X}_2^\perp\boldsymbol{\beta}_{20}\right)/(n-l)}$$

follows $F_{l_2,n-l}$ under the null.[26] This test statistic was first proposed by Anderson and Rubin (1949). The advantage of the F test over the Wald test based on the 2SLS estimation is that it is valid even if the identification fails (i.e., rank$(\boldsymbol{\Gamma}_{22}) < k_2$); see Dufour (1997). The $1-\alpha$ confidence set of $\boldsymbol{\beta}_2$ by inverting the F test, say, $\mathbf{C}_{\beta_2}(\alpha)$, is not generally an ellipsoid.[27] Also, the AR-test is designed to test the complete vector $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$, and is not suitable to build confidence sets for individual components of $\boldsymbol{\beta}_2$ or some tranformation $\mathbf{r}(\boldsymbol{\beta}_2) \in \mathbb{R}^q$, where $q < k_2$. A generic solution is the **projection-based confidence set**; see, e.g., Dufour (1990, 1997), Wang and Zivot (1998), Dufour and Jasiak (2001) and Dufour and Taamouti (2005, 2007). Such a confidence set takes the image set $\mathbf{r}\left(\mathbf{C}_{\boldsymbol{\beta}_2}(\alpha)\right) = \left\{\mathbf{r}(\boldsymbol{\beta}_2) | \boldsymbol{\beta}_2 \in \mathbf{C}_{\boldsymbol{\beta}_2}(\alpha)\right\}$ as the confidence set. Because $\boldsymbol{\beta}_2 \in \mathbf{C}_{\boldsymbol{\beta}_2}(\alpha)$ implies $\mathbf{r}(\boldsymbol{\beta}_2) \in \mathbf{r}\left(\mathbf{C}_{\boldsymbol{\beta}_2}(\alpha)\right)$, we have

$$P\left(\mathbf{r}(\boldsymbol{\beta}_2) \in \mathbf{r}\left(\mathbf{C}_{\boldsymbol{\beta}_2}(\alpha)\right)\right) \geq P\left(\boldsymbol{\beta}_2 \in \mathbf{C}_{\boldsymbol{\beta}_2}(\alpha)\right) \geq 1-\alpha,$$

i.e., such a confidence set is valid. When $\mathbf{r}(\boldsymbol{\beta}_2) = \beta_j$, an individual element of $\boldsymbol{\beta}_2$, $\mathbf{r}\left(\mathbf{C}_{\boldsymbol{\beta}_2}(\alpha)\right)$ can be interpreted as the projection of $\mathbf{C}_{\boldsymbol{\beta}_2}(\alpha)$ on the β_j-axis; $\mathbf{r}\left(\mathbf{C}_{\boldsymbol{\beta}_2}(\alpha)\right)$ need not be an interval, but we can take its convex hull, which is an interval, as the confidence interval. The projection-based confidence set is typically nonsimilar and conservative.

[26] This null distribution relies on the normality of u_i but not that of $\mathbf{v}_{2i}$, which implies that the reduced-form equation of $\mathbf{x}_{2i}$ may suffer from the omitted instruments.

[27] The confidence set can be even empty when $\widehat{\lambda}$ in Exercise 22 exceeds some constant. Empty confidence set is an indication that the model is misspecified, e.g., there does not exist $\boldsymbol{\beta}_2$ such that $\boldsymbol{\lambda}_2 = \boldsymbol{\Gamma}_{22}\boldsymbol{\beta}_2$.

8 k-class Estimators (*)

The LIML estimator (31) is a special k-**class estimator**[28] (see Theil (1961) and Nagar (1959)),

$$\widehat{\boldsymbol{\beta}}(k)=\left[\mathbf{X}'\left(\mathbf{I}_n-k\mathbf{M}_{\mathbf{Z}}\right)\mathbf{X}\right]^{-1}\mathbf{X}'\left(\mathbf{I}_n-k\mathbf{M}_{\mathbf{Z}}\right)\mathbf{y}.$$

Inspection of the 2SLS formula (24) shows that the 2SLS estimator is a k-class estimator with $k=1$ (which implies that under just-identification the IV estimator is MLE under normality), and the OLS estimator is a k-class estimator with $k=0$. 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 because its distribution has fat tails, so it generally has large disperson especially with weak instruments; 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 median of LIML is typically much closer to $\boldsymbol{\beta}$ than is the mean or median of TSLS. 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 $\mathbf{Z}_2$.[29] In the fixed-instrument, normal-error model (i.e., $\mathbf{Z}$ is fixed and $(u_i,\mathbf{v}_i')$ are iid jointly normal) with $l_1=k_1=0$ and $k=1$, Rothenberg (1984) shows that the bias of 2SLS (to three terms) is

$$\frac{(l-2)\,\rho}{\mu^2}\frac{\sigma_u}{\sigma_v},$$

which increases with l, but the bias of LIML,

$$-\frac{\rho}{\mu^2}\frac{\sigma_u}{\sigma_v},$$

does not, where ρ is the correlation between u_i and v_i, and $\mu^2=\boldsymbol{\Gamma}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\Gamma}/\boldsymbol{\sigma}_v^2$ is often called the **concentration parameter**; for the general asymptotic expansion of the 2SLS estimator, see Sargan and Mikhail (1971) and Anderson and Sawa (1973). When the instruments are fixed and the errors are symmetrically distributed, Rothenberg (1983) shows that LIML is the best median-unbiased k-class estimator to second order. 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

[28] The "k" in k-class estimators and the "k" of $\dim(\mathbf{x})$ can be differentiated from the context. Also, $\widehat{\kappa}$ is usually written as k in the literature, which is the origin of the name of k-class estimators.

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

direction. Bekker (1994) presents small-sample results for LIML and a generalization of LIML, and also shows that unlike 2SLS, LIML is consistent under many-instrument asymptotics. 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 $k = \widehat{\kappa}_{\text{LIML}} - \alpha/(n-l)$ 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. Rothenberg (1984) shows that with fixed instruments and normal errors, the Fuller-k estimator with $\alpha = 1$ is best unbiased to second order among estimators with $k = 1 + a\left(\widehat{\kappa}_{\text{LIML}} - 1\right) - c/\left(n-l\right)$ for some constants a and c. In Monte Carlo simulations, Hahn et al. (2004) reported substantial reductions in bias and MSE using Fuller-k estimators, relative to 2SLS and LIML, when instruments are weak.

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 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.

9 Split-Sample IV and JIVE (*)

We briefly describe the **jackknife instrumental variables estimator** (JIVE) of Angrist, Imbens and Krueger (1999, AIK hereafter) here. To motivate the JIVE, first note that the ideal instrument for estimation of $\boldsymbol{\beta}$ is $\mathbf{w} = \boldsymbol{\Gamma}\mathbf{z}$. We can write the ideal IV estimator as

$$\widehat{\boldsymbol{\beta}}_{\text{ideal}} = \left(\sum_{i=1}^{n} \mathbf{w}_i \mathbf{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{w}_i y_i\right) = \left(\mathbf{W}'\mathbf{X}\right)^{-1}\left(\mathbf{W}'\mathbf{y}\right).$$

This estimator is not feasible since $\boldsymbol{\Gamma}$ is unknown. The 2SLS estimator replaces $\boldsymbol{\Gamma}$ with the multivariate least squares estimator $\widehat{\boldsymbol{\Gamma}}$ and $\mathbf{w}_i$ by $\widehat{\mathbf{x}}_i = \widehat{\boldsymbol{\Gamma}}\mathbf{z}_i$ leading to the following representation for

2SLS

$$\widehat{\boldsymbol{\beta}}_{\text{2SLS}} = \left(\sum_{i=1}^{n} \widehat{\mathbf{x}}_i \mathbf{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \widehat{\mathbf{x}}_i y_i\right) = \left(\widehat{\mathbf{X}}'\mathbf{X}\right)^{-1} \widehat{\mathbf{X}}'\mathbf{y}.$$

Since $\widehat{\boldsymbol{\Gamma}}$ is estimated on the full sample including observation i it is a function of the reduced form error $\mathbf{v}_i$ which is correlated with the structural error u_i. It follows that $\widehat{\mathbf{x}}_i$ and u_i are correlated, which means that $\widehat{\boldsymbol{\beta}}_{\text{2SLS}}$ is biased for $\boldsymbol{\beta}$. More specifically,

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

where $\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\boldsymbol{\Gamma}} = \mathbf{P}_{\mathbf{Z}}\mathbf{X}$ with $\widehat{\boldsymbol{\Gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})$. Note that $\widehat{\mathbf{X}} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{P}_{\mathbf{Z}}\mathbf{V}$, so although $\mathbf{Z}\boldsymbol{\Gamma}$ is not correlated with $\mathbf{u}$, $\mathbf{P}_{\mathbf{Z}}\mathbf{V}$ is. Specifically,

$$\begin{aligned} E\left[u_i \widehat{\mathbf{x}}_i'\right] &= E\left[\mathbf{z}_i'\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{z}_j\mathbf{v}_j'\right)u_i\right] = E\left[\mathbf{z}_i'\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{z}_i\mathbf{v}_i'u_i\right] \\ &= E\left[\mathbf{z}_i'\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{z}_i E\left[\mathbf{v}_i'u_i|\mathbf{Z}\right]\right] = E\left[\mathbf{z}_i'\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{z}_i\sigma_{\mathbf{v}u}\right] = \frac{l}{n}\sigma_{\mathbf{v}u}, \end{aligned}$$

where we denote $E\left[\mathbf{v}_i'u_i|\mathbf{Z}\right]$ as $\sigma_{\mathbf{v}u}$ which is not zero.[30] For fixed l, this bias will vanish in large samples, but in finite samples, this bias is not neglectable.

A possible solution to this problem is to replace $\widehat{\mathbf{x}}_i$ with a predicted value which is uncorrelated with the error u_i. One method is the **split-sample IV** (**SSIV**) estimator of Angrist and Krueger (1995). Divide the sample randomly into two independent halves A and B. Use A to estimate the reduced form and B to estimate the structural coefficient. Specifically, use sample A to construct $\widehat{\boldsymbol{\Gamma}}_A = \left(\mathbf{Z}_A'\mathbf{Z}_A\right)^{-1}\left(\mathbf{Z}_A'\mathbf{X}_A\right)$. Combine this with sample B to create the predicted values $\widehat{\mathbf{X}}_B = \mathbf{Z}_B\widehat{\boldsymbol{\Gamma}}_A$. The SSIV estimator is $\widehat{\boldsymbol{\beta}}_{\text{SSIV}} = \left(\widehat{\mathbf{X}}_B'\mathbf{X}_B\right)^{-1}\widehat{\mathbf{X}}_B'\mathbf{y}_B$. This has lower bias than $\widehat{\boldsymbol{\beta}}_{\text{2SLS}}$. A limitation of SSIV is that the results will be sensitive to the sample spliting. One split will produce one estimator; another split will produce a different estimator. Any specific split is arbitrary, so the estimator depends on the specific random sorting of the observations into the samples A and B. A second limitation of SSIV is that it is unlikely to work well when the sample size n is small.

A much better solution is obtained by a leave-one-out estimator for $\boldsymbol{\Gamma}$. Specifically, let

$$\widehat{\boldsymbol{\Gamma}}_{(-i)} = \left(\mathbf{Z}'\mathbf{Z} - \mathbf{z}_i\mathbf{z}_i'\right)^{-1}\left(\mathbf{Z}'\mathbf{X} - \mathbf{z}_i\mathbf{x}_i'\right) = \left(\mathbf{Z}_{(-i)}'\mathbf{Z}_{(-i)}\right)^{-1}\left(\mathbf{Z}_{(-i)}'\mathbf{X}_{(-i)}\right)$$

be the estimator of $\boldsymbol{\Gamma}$ computed using all but the ith observation in the first stage, and let $\mathbf{x}_i^* = \widehat{\boldsymbol{\Gamma}}_{(-i)}'\mathbf{z}_i$ be the reduced form predicted values. Using $\mathbf{x}_i^*$ as an instrument we obtain the estimator

$$\widehat{\boldsymbol{\beta}}_{\text{JIVE1}} = \left(\sum_{i=1}^{n}\mathbf{x}_i^*\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i^* y_i\right) = \left(\sum_{i=1}^{n}\widehat{\boldsymbol{\Gamma}}_{(-i)}'\mathbf{z}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\widehat{\boldsymbol{\Gamma}}_{(-i)}'\mathbf{z}_i y_i\right) =: \left(\mathbf{X}^{*\prime}\mathbf{X}\right)^{-1}\mathbf{X}^{*\prime}\mathbf{y}.$$

[30] Note that $\sum_{i=1}^{n}\mathbf{z}_i'\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{z}_i = \text{tr}(\mathbf{P}_{\mathbf{Z}}) = l$. Since for a random sample, the expectation of each summand is the same, we have $E\left[\mathbf{z}_i'\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{z}_i\right] = l/n$.

This is the first JIVE of AIK. It first appeared in Phillips and Hale (1977).[31] Now,

$$E\left[u_i\mathbf{x}_i^{*\prime}\right]=E\left[\mathbf{z}_i'\left(\mathbf{Z}_{(-i)}'\mathbf{Z}_{(-i)}\right)^{-1}\mathbf{Z}_{(-i)}'E\left[\mathbf{X}_{(-i)}u_i|\mathbf{Z}\right]\right]=\mathbf{0},$$

i.e., $\widehat{\boldsymbol{\beta}}_{\text{JIVE1}}$ is unbiased.

AIK point out that a somewhat simpler adjustment also removes the correlation and bias. Define the estimator and predicted value

$$\begin{aligned}\widetilde{\boldsymbol{\Gamma}}_{(-i)} &= \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\mathbf{Z}'\mathbf{X}-\mathbf{z}_i\mathbf{x}_i'\right)=\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\mathbf{Z}_{-i}'\mathbf{X}_{-i}\right),\\ \mathbf{x}_i^{**} &= \widetilde{\boldsymbol{\Gamma}}_{(-i)}'\mathbf{z}_i,\end{aligned}$$

which only adjusts the $\mathbf{Z}'\mathbf{X}$ component. Their second JIVE of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\text{JIVE2}}=\left(\sum_{i=1}^n\mathbf{x}_i^{**}\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^n\mathbf{x}_i^{**}y_i\right)=\left(\sum_{i=1}^n\widetilde{\boldsymbol{\Gamma}}_{(-i)}'\mathbf{z}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^n\widetilde{\boldsymbol{\Gamma}}_{(-i)}'\mathbf{z}_iy_i\right)=\left(\mathbf{X}^{**\prime}\mathbf{X}\right)^{-1}\mathbf{X}^{**\prime}\mathbf{y}.$$

The unbiasedness of $\widehat{\boldsymbol{\beta}}_{\text{JIVE2}}$ can be similarly shown.

Using the formula for leave-one-out estimators in Section 6 of Chapter 3, $\widehat{\boldsymbol{\beta}}_{\text{JIVE1}}$ and $\widehat{\boldsymbol{\beta}}_{\text{JIVE2}}$ use two linear operations: the first to create the predicted values $\mathbf{x}_i^*$ or $\mathbf{x}_i^{**}$, and the second to calculate the IV estimator. Thus the estimators do not require significantly more computation than 2SLS.

AIK showed that their JIVEs and 2SLS are asymptotically equivalent under conventional fixed-model asymptotics. The literature usually refers to $\widehat{\boldsymbol{\beta}}_{\text{JIVE1}}$ as the JIVE, and we follow this convention. Calculations drawing on work of Chao and Swanson (2002) reveal that under weak-instrument asymptotics, the JIVE is asymptotically equivalent to a k-class estimator with $\widehat{\kappa}=1+l/(n-l)$. Theoretical calculations by Chao and Swanson (2002) and Monte Carlo simulations by AIK indicate that the JIVE is similar to LIML and improves on 2SLS in bias when there are many weak instruments.[32] Davison and MacKinnon (2006) criticize the JIVE to have low efficiency relative to LIML under homoskedasticity, but as suggested in Hausman et al. (2012), this criticism can be overcomed by combining the JIVE idea with LIML. For asymptotic distribution theory for $\widehat{\boldsymbol{\beta}}_{\text{JIVE1}}$ and $\widehat{\boldsymbol{\beta}}_{\text{JIVE2}}$ in the presence of heteroskedasticity and many instruments, see Chao, et al. (2012).

Exercise 20 *Define P_{ij} as the ijth element of $\mathbf{P}_{\mathbf{Z}}$. Show that (i) $\mathbf{z}_i'\widehat{\boldsymbol{\Gamma}}_{(-i)}=\frac{\mathbf{z}_i'\widehat{\boldsymbol{\Gamma}}-P_{ii}\mathbf{x}_i'}{1-P_{ii}}$; (ii) $\widehat{\boldsymbol{\beta}}_{JIVE2}$ can be rewritten as $\left(\sum_{i\neq j}\mathbf{x}_i'P_{ij}\mathbf{x}_j\right)^{-1}\left(\sum_{i\neq j}\mathbf{x}_i'P_{ij}y_j\right)$.*

Exercise 21 (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*

[31] $\widehat{\boldsymbol{\beta}}_{\text{JIVE1}}$ is the UJIVE of Blomquist and Dahlberg (1999), who propose their JIVE as $\left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*\prime}\mathbf{y}$.

[32] Monte Carlo simulations in Blomquist and Dahlberg (1999) show that the variance of JIVE and LIML is larger than 2SLS.

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).*

Appendix A: 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

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

so

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

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

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

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 $\widehat{\boldsymbol{\Pi}}$ is indeed the minimizer. Since there are no restrictions on the rows of $\boldsymbol{\Pi}$ 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 $\left|(\mathbf{Y}\mathbf{B}-\mathbf{Z}\boldsymbol{\Gamma})'(\mathbf{Y}\mathbf{B}-\mathbf{Z}\boldsymbol{\Gamma})\right|$, which equals that of $\left|\left(\mathbf{Y}-\mathbf{Z}\boldsymbol{\Gamma}\mathbf{B}^{-1}\right)'\left(\mathbf{Y}-\mathbf{Z}\boldsymbol{\Gamma}\mathbf{B}^{-1}\right)\right|$, becomes

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

which can be rewritten as

$$\left|\begin{array}{cc} \left(\mathbf{Y}^{\perp}\boldsymbol{\gamma}\right)'\left(\mathbf{Y}^{\perp}\boldsymbol{\gamma}\right) & \left(\mathbf{Y}^{\perp}\boldsymbol{\gamma}\right)'\left(\mathbf{X}_2^{\perp}-\mathbf{Z}_2^{\perp}\boldsymbol{\Gamma}_{22}\right) \\ \left(\mathbf{X}_2^{\perp}-\mathbf{Z}_2^{\perp}\boldsymbol{\Gamma}_{22}\right)'\left(\mathbf{Y}^{\perp}\boldsymbol{\gamma}\right) & \left(\mathbf{X}_2^{\perp}-\mathbf{Z}_2^{\perp}\boldsymbol{\Gamma}_{22}\right)'\left(\mathbf{X}_2^{\perp}-\mathbf{Z}_2^{\perp}\boldsymbol{\Gamma}_{22}\right) \end{array}\right| \tag{34}$$

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 $\mathbf{B}$,

$$\left|\begin{array}{cc} \mathbf{A}'\mathbf{A} & \mathbf{A}'\mathbf{B} \\ \mathbf{B}'\mathbf{A} & \mathbf{B}'\mathbf{B} \end{array}\right| = \left|\mathbf{A}'\mathbf{A}\right|\left|\mathbf{B}'\mathbf{M}_{\mathbf{A}}\mathbf{B}\right|, \tag{35}$$

we have our target (34) equal to

$$\left(\mathbf{Y}^{\perp}\boldsymbol{\gamma}\right)'\left(\mathbf{Y}^{\perp}\boldsymbol{\gamma}\right)\left|\left(\mathbf{X}_2^{\perp}-\mathbf{Z}_2^{\perp}\boldsymbol{\Gamma}_{22}\right)'\mathbf{M}_{\mathbf{v}}\left(\mathbf{X}_2^{\perp}-\mathbf{Z}_2^{\perp}\boldsymbol{\Gamma}_{22}\right)\right|, \tag{36}$$

where $\mathbf{v}=\mathbf{Y}^{\perp}\boldsymbol{\gamma}$, and note that $\left(\mathbf{Y}^{\perp}\boldsymbol{\gamma}\right)'\left(\mathbf{Y}^{\perp}\boldsymbol{\gamma}\right)$ is scalar so its determinant is itself. The parameters $\boldsymbol{\Gamma}_{22}$ appears only in the second factor of (36). This factor is the determinant of the matrix of sums of squares and cross-products of the residuals from regressions of $\mathbf{M}_{\mathbf{v}}\mathbf{X}_2^{\perp}$ on $\mathbf{M}_{\mathbf{v}}\mathbf{Z}_2^{\perp}$. From the discussion above, the minimizer (corresponding to $\boldsymbol{\Gamma}_{22}$) should be the OLS estimate, and the matrix of residuals is $\mathbf{M}_{\mathbf{M}_{\mathbf{v}}\mathbf{Z}_2^{\perp}}\mathbf{M}_{\mathbf{v}}\mathbf{X}_2^{\perp}=\mathbf{M}_{\mathbf{v},\mathbf{Z}_2^{\perp}}\mathbf{X}_2^{\perp}$. Consequently, the second factor of (36), minimized with respect to $\mathbf{\Gamma}_{22}$, is

$$\left|\mathbf{X}_2^{\perp\prime}\mathbf{M}_{\mathbf{v},\mathbf{Z}_2^{\perp}}\mathbf{X}_2^{\perp}\right|. \tag{37}$$

The fact that $\mathbf{v}$ and $\mathbf{Z}_2^{\perp}$ appear in a symmetrical fashion in (37) can be exploited in order to make (37) depend on $\boldsymbol{\gamma}$ only through a scalar factor. Consider the determinant

$$\left|\begin{array}{cc} \mathbf{v}'\mathbf{M}_{\mathbf{Z}_2^{\perp}}\mathbf{v} & \mathbf{v}'\mathbf{M}_{\mathbf{Z}_2^{\perp}}\mathbf{X}_2^{\perp} \\ \mathbf{X}_2^{\perp\prime}\mathbf{M}_{\mathbf{Z}_2^{\perp}}\mathbf{v} & \mathbf{X}_2^{\perp\prime}\mathbf{M}_{\mathbf{Z}_2^{\perp}}\mathbf{X}_2^{\perp} \end{array}\right|. \tag{38}$$

By use of (35), this determinant can be factorized just as (34) was. We obtain

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

Using the facts that $\mathbf{M}_1\mathbf{M}_{\mathbf{Z}_2^{\perp}} = \mathbf{M}_{\mathbf{Z}}$ and that $\mathbf{v} = \mathbf{M}_1\mathbf{Y}\boldsymbol{\gamma}$, (38) can be rewritten as

$$\left|\begin{array}{cc} \boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\boldsymbol{\gamma} & \boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\boldsymbol{\gamma} & \mathbf{X}_2'\mathbf{M}_{\mathbf{Z}}\mathbf{X}_2 \end{array}\right| = \left|\mathbf{B}'\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\mathbf{B}\right| = \left|\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\right|, \tag{39}$$

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 (39) does not depend on $\mathbf{B}$ at all.

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

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

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

$$\mathbf{Y}'\mathbf{M}_1\mathbf{Y}\boldsymbol{\gamma}{-}\kappa\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\boldsymbol{\gamma} = \mathbf{0}, \tag{40}$$

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 (40) 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 (31).

Exercise 22 *Define* $\widehat{\lambda}=\min_{\boldsymbol{\beta}}\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\mathbf{P}_{\mathbf{Z}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}$ *and* $\widetilde{\lambda}=\frac{\widehat{\lambda}-\frac{\alpha}{n-l}\left(1-\widehat{\lambda}\right)}{1-\frac{\alpha}{n-l}\left(1-\widehat{\lambda}\right)}$. *Show that the LIML estimator can be re-expressed as*

$$\widehat{\boldsymbol{\beta}}_{LIML}=\left(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X}-\widehat{\lambda}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{y}-\widehat{\lambda}\mathbf{X}'\mathbf{y}\right),$$

and Fuller's estimator can be re-expressed as

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

Appendix B: LIML Asymptotic Distribution

For the distribution theory, it is useful to rewrite (31) as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LIML}}=\left(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X}-\widehat{\mu}\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{y}-\widehat{\mu}\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{y}\right),$$

where

$$\widehat{\mu}=\widehat{\kappa}-1=\min_{\boldsymbol{\gamma}}\frac{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_1\mathbf{Z}_2\left(\mathbf{Z}_2'\mathbf{M}_1\mathbf{Z}_2\right)^{-1}\mathbf{Z}_2'\mathbf{M}_1\mathbf{Y}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\boldsymbol{\gamma}}.$$

This second equality holds since the span of $\mathbf{Z}=[\mathbf{Z}_1,\mathbf{Z}_2]$ equals the span of $[\mathbf{Z}_1,\mathbf{M}_1\mathbf{Z}_2]$. This implies

$$\mathbf{P}_{\mathbf{Z}}=\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'=\mathbf{Z}_1\left(\mathbf{Z}_1'\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1'+\mathbf{M}_1\mathbf{Z}_2\left(\mathbf{Z}_2'\mathbf{M}_1\mathbf{Z}_2\right)^{-1}\mathbf{Z}_2'\mathbf{M}_1.$$

We now show that $n\widehat{\mu}=O_p(1)$. The reduced form of y_i implies

$$\mathbf{Y}=\mathbf{Z}_1\boldsymbol{\Pi}_1+\mathbf{Z}_2\boldsymbol{\Pi}_2+\mathbf{F},$$

where $\boldsymbol{\Pi}_1=[\boldsymbol{\lambda}_1,\boldsymbol{\Gamma}_{12}]$, $\boldsymbol{\Pi}_2=[\boldsymbol{\lambda}_2,\boldsymbol{\Gamma}_{22}]$ and $\mathbf{F}$ stacks $\mathbf{f}_i'$ with $\mathbf{f}_i'=[e_i,\mathbf{v}_{2i}']$. Since $\boldsymbol{\Pi}_2=[\boldsymbol{\Gamma}_{22}\boldsymbol{\beta}_2,\boldsymbol{\Gamma}_{22}]$, it follows that $\boldsymbol{\Pi}_2\overline{\boldsymbol{\gamma}}=\mathbf{0}$ for $\overline{\boldsymbol{\gamma}}=\left(1,-\boldsymbol{\beta}_2'\right)'$. Note that $\mathbf{F}\overline{\boldsymbol{\gamma}}=\mathbf{u}$, so $\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\overline{\boldsymbol{\gamma}}=\mathbf{M}_{\mathbf{Z}}\mathbf{u}$ and $\mathbf{M}_1\mathbf{Y}\overline{\boldsymbol{\gamma}}=\mathbf{M}_1\mathbf{u}$.

Hence

$$
\begin{aligned}
n\widehat{\mu} &= \min_{\boldsymbol{\gamma}} \frac{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_1\mathbf{Z}_2\left(\mathbf{Z}_2'\mathbf{M}_1\mathbf{Z}_2\right)^{-1}\mathbf{Z}_2'\mathbf{M}_1\mathbf{Y}\boldsymbol{\gamma}}{\frac{1}{n}\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\boldsymbol{\gamma}} \\
&\leq \frac{\left(\frac{1}{\sqrt{n}}\mathbf{u}'\mathbf{M}_1\mathbf{Z}_2\right)\left(\frac{1}{n}\mathbf{Z}_2'\mathbf{M}_1\mathbf{Z}_2\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{Z}_2'\mathbf{M}_1\mathbf{u}\right)}{\frac{1}{n}\mathbf{u}'\mathbf{M}_{\mathbf{Z}}\mathbf{u}} \\
&= O_p\left(1\right).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{LIML}} - \boldsymbol{\beta}\right) &= \left(\frac{1}{n}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X} - \widehat{\mu}\frac{1}{n}\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{u} - \sqrt{n}\widehat{\mu}\frac{1}{n}\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{u}\right) \\
&= \left(\frac{1}{n}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X} - o_p\left(1\right)\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{u} - o_p\left(1\right)\right) \\
&= \sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{2SLS}} - \boldsymbol{\beta}\right) + o_p\left(1\right).
\end{aligned}
$$

Obviously, as long as $\sqrt{n}\widehat{\mu} = o_p\left(1\right)$, all k-estimators have the same asymptotic distribution as the 2SLS estimator.