

# Endogeneity and Instrumental Variables

Ping Yu

School of Economics and Finance  
The University of Hong Kong

- 1 Endogeneity
- 2 Instrumental Variables
- 3 Reduced Form
- 4 Identification
- 5 Estimation: Two-Stage Least Squares
- 6 Interpretation of the IV Estimator

# Endogeneity

# Endogeneity

- In the linear regression

$$y_i = \mathbf{x}_i' \beta + u_i, \quad (1)$$

if  $E[\mathbf{x}_i u_i] \neq \mathbf{0}$ , there is *endogeneity*.

- In this case, the LSE will be asymptotically biased.
- The analysis of data with endogenous regressors is arguably the main contribution of econometrics to statistical science.

## Five Sources of Endogeneity

- Simultaneous causality.
  - Example: Does computer usage increase the income? Do Cigarette taxes reduce smoking? Does putting criminals in jail reduce crime?
  - Solution: using instrumental variables (IVs), and designing and implementing a randomizing controlled experiment in which the reverse causality channel is nullified
- Omitted variables.
  - 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.
  - Solution: using IVs, using panel data and using randomizing controlled experiments.

## Continue...

- Errors in variables. This term refers to the phenomenon that an otherwise exogenous regressor becomes endogenous when measured with error.
  - 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.
  - Solution: using IVs (e.g., exogenous determinants of the error ridden explanatory variables, or multiple indicators of the same outcome).
- Sample selection.
  - 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.
  - Solution: Heckman's control function approach.
- Functional form misspecification.  $E[y|\mathbf{x}]$  may not be linear in  $\mathbf{x}$ . Solution: nonparametric methods.

## Simultaneous Causality

- Wright (1928) considered to estimate the elasticity of butter demand, which is critical in the policy decision on the tariff of butter.
- Define  $p_i = \ln P_i$  and  $q_i = \ln Q_i$ , and the demand equation is

$$q_i = \alpha_0 + \alpha_1 p_i + u_i, \quad (2)$$

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):

$$q_i = \beta_0 + \beta_1 p_i + v_i, \quad (3)$$

where  $v_i$  represents the factors that affect supply, such as weather conditions, factor prices, and union status.

- So  $p_i$  and  $q_i$  are determined "within" the model, and they are endogenous. Rigorously, note that

$$p_i = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1},$$

$$q_i = \frac{\alpha_1 \beta_0 - \alpha_0 \beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 v_i - \beta_1 u_i}{\alpha_1 - \beta_1},$$

by solving two simultaneous equations (2) and (3).

continue...

- Suppose  $Cov(u_i, v_i) = 0$ , then

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

which are not zero. If  $\alpha_1 < 0$  and  $\beta_1 > 0$ , then  $Cov(p_i, u_i) > 0$  and  $Cov(p_i, v_i) < 0$ , which is intuitively right (why?).

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

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

$$\stackrel{\text{why?}}{=} \frac{\alpha_1 Var(v_i) + \beta_1 Var(u_i)}{Var(v_i) + Var(u_i)} \in (\alpha_1, \beta_1).$$

- So the LSE is neither  $\alpha_1$  nor  $\beta_1$ , but a weighted average of them. Such a bias is called the *simultaneous equations bias*. The LSE cannot consistently estimate  $\alpha_1$  or  $\beta_1$  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.



continue...

- 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  $\alpha_1$ . Figure 1 illustrates the discussion above intuitively.

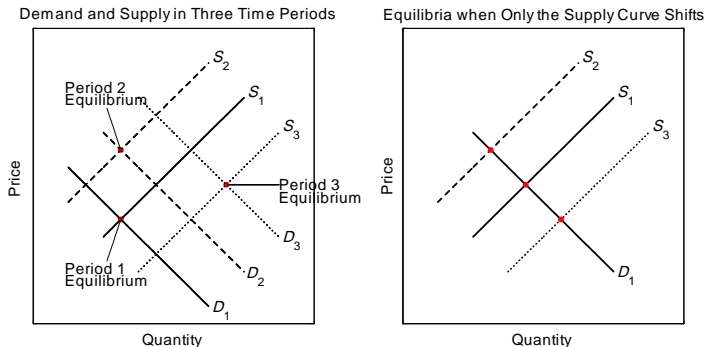


Figure: Endogeneity and Identification of Instrument Variables

continue...

- From above, we can see that  $p_i$  has one part which is correlated with  $u_i$  ( $-\frac{u_i}{\alpha_1 - \beta_1}$ ) and one part is not ( $\frac{v_i}{\alpha_1 - \beta_1}$ ). 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 one 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

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

So

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

and

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

- A natural estimator is

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

which is the *IV estimator*.

continue...

- Another method to estimate  $\alpha_1$  as suggested above is to run regression

$$q_i = \alpha_0 + \alpha_1 \hat{p}_i + \tilde{u}_i,$$

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

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

and  $\tilde{u}_i = \alpha_1 (p_i - \hat{p}_i) + u_i$ .

- It is easy to show that  $Cov(\hat{p}_i, \tilde{u}_i) = 0$ , 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.

## Omitted Variables

- Mundlak (1961) considered the production function estimation, where the error term includes factors that are observable to the economic agent under study but unobservable to the econometrician, and endogeneity arises when regressors are decisions made by the agent on the basis of such factors.
- Suppose that a farmer is producing a product with a Cobb-Douglas technology:

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

where  $Q_i$  is the output on the  $i$ th farm,  $L_i$  is a variable input (labor),  $A_i$  represents an input that is fixed over time (soil quality), and  $v_i$  represents a stochastic input (rainfall), which is not under the farmer's control.

- We shall assume that the farmer knows the product price  $p$  and input price  $w$ , which do not depend on his decisions, and that he knows  $A_i$  but econometricians do not.
- The factor input decision is made before knowing  $v_i$ , and so  $L_i$  is chosen to maximize expected profits. The factor demand equation is

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

so a better farm induces more labors on it.

continue...

- We assume that  $(A_i, v_i)$  is i.i.d. over farms, and  $A_i$  is independent of  $v_i$  for each  $i$ , so  $B = E[\exp(v_i)]$  is the same for all  $i$ , and the level of output the farm expects when it chooses  $L_i$  is  $A_i \cdot (L_i)^{\phi_1} \cdot B$ .
- Take logarithm on both sides of (4), we have a log-linear production function:

$$\log Q_i = \log A_i + \phi_1 \cdot \log(L_i) + v_i.$$

$\log A_i$  is an omitted variable. Equivalently, each farm has a different intercept.

- The LSE of  $\phi_1$  will converge to

$$\frac{\text{Cov}(\log Q_i, \log(L_i))}{\text{Var}(\log(L_i))} = \phi_1 + \frac{\text{Cov}(\log A_i, \log(L_i))}{\text{Var}(\log(L_i))},$$

which is not  $\phi_1$  since there is correlation between  $\log A_i$  and  $\log(L_i)$  as shown in (5).

- Figure 2 shows the effect of  $\log A_i$  on  $\phi_1$  by drawing  $E[\log Q | \log L, \log A]$  for two farms. In Figure 2, the OLS regression line passes through points AB with slope  $\frac{\log Q_1 - \log Q_2}{\log L_1 - \log L_2}$ , but the true  $\phi_1$  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  $\log A_i$ .

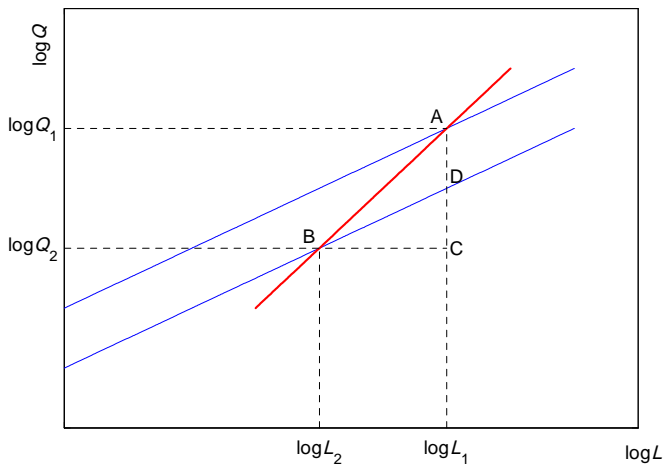


Figure: Effect of Soil Quality on Labor Input

continue...

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

$$\log Q_i = \phi_0 + \phi_1 \cdot \log(L_i) + v_i + u_i, \quad (6)$$

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

where  $\beta_0 = \frac{1}{1 - \phi_1} \left( \phi_0 + \log(B\phi_1) - \log\left(\frac{w}{p}\right) \right)$  is a constant for all farms.

- It is obvious that  $\log L_i$  is correlated with  $(v_i + u_i)$ . Thus, the LSE of  $\phi_1$  in the estimation of log-linear production function confounds the contribution to output of  $u_i$  with the contribution of labor. Actually,

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

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

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

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

## 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:  $C_i = C_i^* + c_i$  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

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

- $E[Y_i u_i] = -kE[y_i^2] < 0$ , so the LSE of  $k$  converges to

$$\frac{E[Y_i C_i]}{E[Y_i^2]} = \frac{kE[(Y_i^*)^2]}{E[(Y_i^*)^2] + E[y_i^2]} < k.$$

- Taking expectation on both sides of (8), we have  $E[C_i] = kE[Y_i] + E[u_i]$ . So  $z = 1$  is a valid IV if  $E[y_i] = E[c_i] = 0$  and  $E[Y_i^*] = E[Y_i] \neq 0$ . The IV estimation using  $z$  as the instrument is  $\frac{\bar{C}_i}{\bar{Y}_i}$ , which is how Friedman estimated  $k$ .



## continue...

- Actually, measurement errors are embodied in regression analysis from the beginning. Galton (1889) analyzed the relationship between the height of sons and the height of fathers. More specifically,

$$S_i = \alpha + \beta F_i + u_i,$$

where  $S_i$  and  $F_i$  are the heights of sons and fathers, respectively.

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

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

- The OLS estimator of  $\beta$  will converge to

$$\frac{\text{Cov}(F_i, S_i)}{\text{Var}(S_i)} = \frac{\text{Var}(F_i)}{\text{Var}(F_i) + \text{Var}(f_i)} < 1,$$

where  $\frac{\text{Var}(F_i)}{\text{Var}(F_i) + \text{Var}(f_i)} \equiv \rho$  is called the *reliability coefficient*. In Galton's analysis, this coefficient is about 2/3. He termed this phenomenon as "regression towards mediocrity".

- The regression line and the true line are shown in Figure 3.

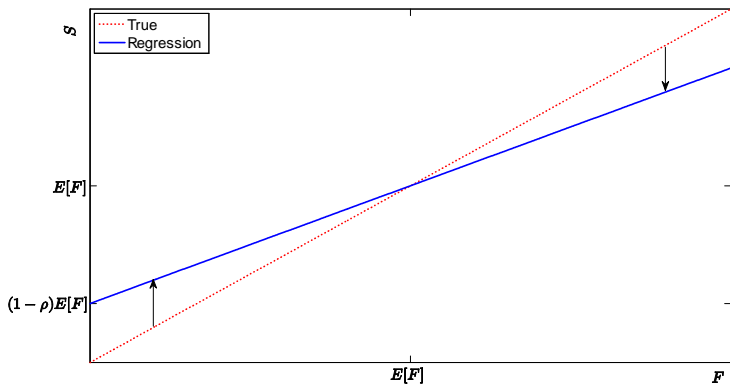


Figure: Relationship Between the Height of Sons and Fathers

# Instrumental Variables

# Instrumental Variables

- $y_i = \mathbf{x}'_i \beta + u_i$  is called the *structural equation* or *primary equation*. In matrix notation, it can be written as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}. \quad (9)$$

- Any solution to the problem of endogeneity requires additional information which we call *instrumental variables* (or simply *instruments*).
- The  $l \times 1$  random vector  $\mathbf{z}_i$  is an instrument for (1) if  $E[\mathbf{z}_i u_i] = \mathbf{0}$ . This condition cannot be tested in practice since  $u_i$  cannot be observed.
- In a typical set-up, some regressors in  $\mathbf{x}_i$  will be uncorrelated with  $u_i$  (for example, at least the intercept). Thus we make the partition

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{x}_{2i} \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix}, \quad (10)$$

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*.

continue...

- By the above definition,  $\mathbf{x}_{1i}$  is an instrumental variable, so should be included in  $\mathbf{z}_i$ , giving the partition

$$\mathbf{z}_i = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{z}_{2i} \end{pmatrix} \begin{matrix} k_1 \\ l_2 \end{matrix}, \quad (11)$$

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.

# Reduced Form

## 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

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

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

- Define

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

as the projection error. Note that  $\mathbf{v}_i$  must be correlated with  $u_i$ . (why?)

- The reduced form linear relationship between  $\mathbf{x}_i$  and  $\mathbf{z}_i$  is the instrumental equation

$$\mathbf{x}_i = \Gamma' \mathbf{z}_i + \mathbf{v}_i. \quad (12)$$

In matrix notation,

$$\mathbf{X} = \mathbf{Z}\Gamma + \mathbf{V}, \quad (13)$$

where  $\mathbf{V}$  is a  $n \times k$  matrix.

- By construction,  $E[\mathbf{z}_i \mathbf{v}_i'] = \mathbf{0}$ , so (12) is a projection and can be estimated by OLS:

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

## continue...

- Substituting (13) into (9), we find

$$\mathbf{y} = (\mathbf{Z}\Gamma + \mathbf{V})\boldsymbol{\beta} + \mathbf{u} = \mathbf{Z}\boldsymbol{\lambda} + \mathbf{e} \quad (14)$$

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

- Observe that

$$E[\mathbf{z}\mathbf{e}] = E[\mathbf{z}\mathbf{v}']\boldsymbol{\beta} + E[\mathbf{z}\mathbf{u}] = \mathbf{0}. \quad (15)$$

Thus (14) is a projection equation and may be estimated by OLS. This is

$$\mathbf{y} = \mathbf{Z}\hat{\boldsymbol{\lambda}} + \hat{\mathbf{e}}, \hat{\boldsymbol{\lambda}} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{y}).$$

- The equation (14) is the reduced form for  $\mathbf{y}$ . (13) and (14) together are the reduced form equations for the system

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

- 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 simultaneous equations* because the second part of equations do not depend on  $\mathbf{y}$ .



# Identification

## Identification

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

$$E[\mathbf{z}_i y_i] = E[\mathbf{z}_i \mathbf{x}'_i] \beta. \quad (16)$$

Multiplying each side by an invertible matrix  $E[\mathbf{z}_i \mathbf{z}'_i]^{-1}$ , we have  $\lambda = \Gamma\beta$ .

- The parameter is identified, meaning that it can be uniquely recovered from the reduced form, if the *rank condition*

$$\text{rank}(\Gamma) = k \quad (17)$$

holds.

- If  $\text{rank}(E[\mathbf{z}_i \mathbf{z}'_i]) = l$  (this is trivial), and  $\text{rank}(E[\mathbf{z}_i \mathbf{x}'_i]) = k$  (this is crucial), this condition is satisfied.
- Assume that (17) holds. If  $l = k$ , then  $\beta = \Gamma^{-1}\lambda$ . If  $l > k$ , then for any  $\mathbf{A} > \mathbf{0}$ ,  $\beta = (\Gamma' \mathbf{A} \Gamma)^{-1} \Gamma' \mathbf{A} \lambda$ .
- If (17) is not satisfied, then  $\beta$  cannot be uniquely recovered from  $(\lambda, \Gamma)$ .
- Note that a necessary (although not sufficient) condition for (17) is the *order condition*  $l \geq k$ .

continue...

- Since  $\mathbf{Z}$  and  $\mathbf{X}$  have the common variables  $\mathbf{X}_1$ , we can rewrite some of the expressions.
- Using (10) and (11) to make the matrix partitions  $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2]$  and  $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ , we can partition  $\Gamma$  as

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

- (13) can be rewritten as

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

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

## What Variable Is Qualified to Be An IV?

- It is often suggested to select an instrumental variable that is

(i) uncorrelated with  $u$ ; (ii) correlated with endogenous variables. (18)

- (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  $\mathbf{Z}$  and  $\mathbf{X}_1$  are not perfectly multicollinear; in other words, there must be "enough" extra variation in  $\hat{\mathbf{x}}_2$  that can not be explained by  $\mathbf{x}_1$ . Such a condition is required in the second stage regression.
- Sometimes (18) is misleading.
- Check the following example with only one endogenous variable:

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

## continue...

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

$$\begin{aligned} E[zu] &= 0, & (19) \\ x_2 &= x_1\gamma_1 + z\gamma_2 + v \text{ with } \gamma_2 \neq 0. \end{aligned}$$

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

- The arguments above indicate that (18) is not sufficient, is it necessary? The answer is still NO!
- For this simple example, can we find some  $z$  such that

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


## continue...

- Observe that  $\text{Cov}(z, x_2) = \text{Cov}(z, x_1\gamma_1 + z\gamma_2 + v) = \text{Cov}(z, x_1)\gamma_1 + \text{Var}(z)\gamma_2$ , so if  $\frac{\text{Cov}(z, x_1)}{\text{Var}(z)} = -\frac{\gamma_2}{\gamma_1}$ , this could happen.
- That is, although  $z$  is not correlated with  $x_2$ ,  $z$  is correlated with  $x_1$ , and  $x_1$  is correlated with  $x_2$ . In mathematical language,  $\text{Cov}(z, x_1) \neq 0$ ,  $\gamma_1 \neq 0$ .
- In such a case,  $z$  is related to  $x_2$  only indirectly through  $x_1$ . If we assume  $\text{Cov}(z, x_1) = 0$ , or  $\gamma_1 = 0$ , then the assumption  $\text{Cov}(z, x_2) \neq 0$  is the right condition for  $z$  to be a valid instrument.
- So the right condition should be that  $z$  is *partially* correlated with  $x_2$  after netting out the effect of  $x_1$ .
- In general, a *necessary* condition for a set of qualified instruments is that at least one instrument appears in each of the first-stage regression.
  - When  $k = \ell$ , each instrument must appear in at least one endogenous regression (why?).

## How to Select Instruments?

- Generally speaking, good instruments are not selected based on mathematics, but based on economic theory.
- Some popular examples are listed:
  - Angrist and Krueger (1991) propose using quarter of birth as an IV for education in the analysis of returns to schooling because of a mechanical interaction between compulsory school attendance laws and age at school entry.
  - Card (1995) uses college proximity<sup>1</sup> 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.
  - 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.

---

<sup>1</sup>Parental education is another popular IV to identify the returns to schooling. 

# Estimation: Two-Stage Least Squares



## IV Estimator

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

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

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

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

## 2SLS Estimator as An IV Estimator

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

$$\mathbf{0} = E[E^*[\mathbf{x}|\mathbf{z}]u] = E[\Gamma'\mathbf{z}u] = E[\Gamma'\mathbf{z}(y - \mathbf{x}'\beta)], \quad (20)$$

where  $E^*[\mathbf{x}|\mathbf{z}]$  is the linear projection of  $\mathbf{x}$  on  $\mathbf{z}$ .

- Replacing population expectations with sample averages in (20) yields

$$\hat{\beta}_{2SLS} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}'\hat{\mathbf{y}},$$

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

- When  $l = k$ , the 2SLS estimator and the IV estimator are numerically equivalent (why?).

## Theil (1953)'s Formulation of 2SLS

- The source of the name "two-stage" is from Theil (1953)'s formulation of 2SLS.
- From (15),

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

i.e.,  $\beta$  is the least squares regression coefficients of the regression of  $y$  on fitted values of  $\Gamma'\mathbf{z}$ , so this method is often called the *fitted-value* method.

- The sample analogue is the following two-step procedure:
  - 1 First, regress  $\mathbf{X}$  on  $\mathbf{Z}$  to get  $\hat{\mathbf{X}}$ .
  - 2 Second, regress  $\mathbf{y}$  on  $\hat{\mathbf{X}}$  to get

$$\hat{\beta}_{2SLS} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}'\mathbf{y} = (\mathbf{X}'\mathbf{P}\mathbf{X})^{-1} (\mathbf{X}'\mathbf{P}\mathbf{y}). \quad (21)$$

## Basmann (1957)'s and Telser (1964)'s version of 2SLS

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

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

so

$$\hat{\beta}_{2SLS} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}'\hat{\mathbf{y}}.$$

- Equivalently,  $\hat{\beta}_{2SLS} = \arg \min_{\beta} (\mathbf{y} - \mathbf{X}\beta)'\mathbf{P}_{\mathbf{Z}}(\mathbf{y} - \mathbf{X}\beta)$ , which is a GLS estimator.
- Telser (1964)'s control function formulation:

$$\begin{pmatrix} \hat{\beta}_{2SLS} \\ \hat{\rho}_{2SLS} \end{pmatrix} = (\hat{\mathbf{W}}'\hat{\mathbf{W}})^{-1} \hat{\mathbf{W}}'\mathbf{y}, \quad (22)$$

where  $\hat{\mathbf{W}} = [\hat{\mathbf{X}}, \hat{\mathbf{V}}]$ .

- This construction exploits another implication of  $E[\mathbf{z}u] = \mathbf{0}$ :

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

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

- Thus, this particular linear combination of the first-stage errors  $\mathbf{v}$  is a function that controls for the endogeneity of the regressors  $\mathbf{x}$ .

Scrutinizing  $\hat{\mathbf{X}}$ 

- Recall that  $\mathbf{Z} = [\mathbf{X}_1, \mathbf{Z}_2]$  and  $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ , so

$$\hat{\mathbf{X}} = [\mathbf{P}\mathbf{X}_1, \mathbf{P}\mathbf{X}_2] = [\mathbf{X}_1, \mathbf{P}\mathbf{X}_2] = [\mathbf{X}_1, \hat{\mathbf{X}}_2],$$

since  $\mathbf{X}_1$  lies in the span of  $\mathbf{X}$ .

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

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

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

## The Wald (1940) Estimator - A Special IV Estimator

- The Wald estimator is a special IV estimator when the single instrument  $z$  is binary.
- Suppose we have the model

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

- The identification conditions are

$$\text{Cov}(z, x) \neq 0, \text{Cov}(z, u) = 0. (\text{Why?}) \quad (23)$$

- It can be shown that the IV estimator is

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

continue...

- 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 it can be shown that  $\hat{\beta}_1$  is equivalent to

$$\hat{\beta}_{Wald} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0} \xrightarrow{p} \frac{E[y|z=1] - E[y|z=0]}{E[x|z=1] - E[x|z=0]},$$

where  $\bar{y}_1$  is mean of  $y$  across the  $n_1$  observations with  $z = 1$ ,  $\bar{y}_0$  is the mean of  $y$  across the  $n_0$  observations with  $z = 0$ , and analogously for  $x$ .

- 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  $z$  divided by the shift in  $q$  by  $z$  is indeed a reasonable slope estimator of the demand curve.

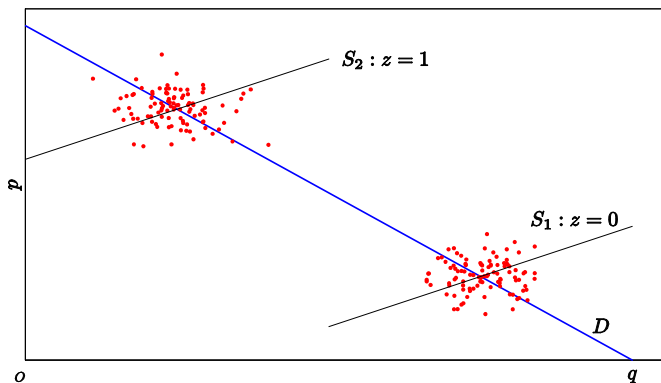


Figure: Intuition for the Wald Estimator in the Linear Demand/Supply System



## Some Popular Examples of the Wald Estimator

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

# Interpretation of the IV Estimator

## The IV Estimation as a Projection

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

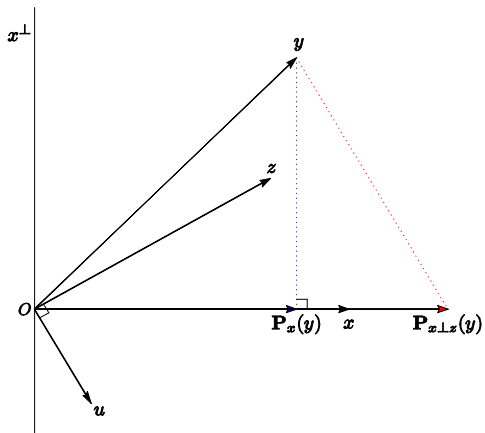


Figure: Projection Interpretation of the IV Estimator

# What is the IV Estimator Estimating?

- The **local average treatment effect estimator** (LATE)
  - the average treatment effect for those individuals whose  $x$  status is affected by  $z$ .
- See Imbens and Angrist (1994).