

## 1 Omitted Variables and Lack of Orthogonality

Suppose we have an outcome of interest  $y_i$ , and two sets of regressor variables  $x_i$  and  $v_i$ , and we assume:

$$E[y_i|x_i, v_i] = x_i'\gamma + v_i'\delta.$$

Can write this as:

$$y_i = x_i'\gamma + v_i'\delta + \varepsilon_i,$$

where  $\varepsilon_i$  is a conditional mean residual.

Suppose want to estimate  $\gamma$ , but we **do not** observe  $v_i$ . Write

$$y_i = x_i'\gamma + u_i, \quad \text{where } u_i = v_i'\delta + \varepsilon_i.$$

Note that  $E[u_i|x_i] \neq 0$ , in general, and this in turn implies

$$E[x_i u_i] = E[x_i E[u_i|x_i]] \neq 0.$$

Consider a least squares regression of  $y_i$  on  $x_i$ :

$$\begin{aligned} \hat{\gamma}_{ls} &= (X'X)^{-1}X'y \\ &= \left( \sum_{i=1}^n x_i x_i' \right)^{-1} \sum_{i=1}^n x_i y_i \\ &= \left( \sum_{i=1}^n x_i x_i' \right)^{-1} \sum_{i=1}^n x_i (x_i \gamma + u_i) \\ &= \gamma + \left( \sum_{i=1}^n x_i x_i' \right)^{-1} \sum_{i=1}^n x_i u_i. \end{aligned}$$

However, as an estimator for  $\gamma$ , the OLS estimator is biased:

$$\begin{aligned} E[\hat{\gamma}_{ls}|X] &= \gamma + \left( \sum_{i=1}^n x_i x_i' \right)^{-1} \sum_{i=1}^n x_i E[u_i|x_i] \\ &\neq \gamma. \end{aligned}$$

It is also inconsistent:

$$\begin{aligned} \text{plim } \hat{\gamma}_{ls} &= \text{plim} \left[ \gamma + \left( \frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^n x_i u_i \right] \\ &= \gamma + E[x_i x_i']^{-1} E[x_i u_i] \\ &\neq \gamma. \end{aligned}$$

This highlights that  $E[x_i u_i] \neq 0$  is a key source of difficulty.

## 2 Example: Wages and Schooling

Let's fix ideas by returning to the wage-schooling relationship considered in the previous lecture note.

Let  $y_i$  be the log of hourly wage, and let  $x_i$  be a vector containing a constant, the college indicator, age, gender, and race.

What other unobserved factors would we (ideally) like to control for? The vector  $v_i$  could include measures of ability, motivation, and other unobserved components of human capital that might be relevant for wages.

A comment on exposition: it is common to see empirical papers write down equations like

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

and assert that this is the model. Logically, this is devoid of content, without some explanation of the meaning of  $u_i$  or  $\beta$ . If we interpret  $u_i$  as  $y_i - E[y_i|x_i]$ , then the equation above implies that the conditional mean of  $y_i$  given  $x_i$  is  $x_i' \beta$ , and OLS would be unbiased and consistent for it.

Often, people have in mind that there are other unobserved factors which should be controlled for, that appear in  $u_i$ . Then OLS will not be a good estimator for  $\beta$ .

## 3 Linear IV: Just-Identified Case

Continue with the wage-schooling example.

Suppose we have another vector of random variables  $z_i$ , where

$$m := \dim(v_i) \geq \dim(x_i) =: k.$$

The vector  $z_i$  should satisfy:

$$E[z_i u_i] = 0$$

(so the elements of  $z_i$  should be uncorrelated with ability and motivation). This is called an "orthogonality condition."

In addition, we need that

$$\text{rank} \{E[z_i x_i']\} = k.$$

A vector  $z_i$  that satisfies these conditions is called an **instrumental variable**.

For example, we might have

$$z_i = (1, \text{age}_i, \text{gender}_i, \text{race}_i, \text{dist}_i)'$$

where  $\text{dist}_i$  is the distance of the individual's home (during high school) to the nearest college or university.

Notice that  $z_i$  has some components in common with  $x_i$ . However,  $z_i$  does **not** contain the college indicator, because we assume the college variable is correlated with unobserved ability and motivation.

Also,  $z_i$  contains a variable not in  $x_i$ : distance to nearest college. For distance to be a reasonable variable to include in  $z_i$ , we need:

- Distance is uncorrelated with ability or motivation. This might be reasonable if we assume that individual's locations are selected randomly or at least in some way that is not related to the unobserved variables.
- We also need to assume that it **is** somehow related to schooling choice, because otherwise the rank condition might not be satisfied.
- Also, we are implicitly assuming that distance does not have a direct role in determining wages; otherwise, it should be included as part of the regressors  $x_i$ .

Since  $E[z_i u_i] = 0$ , we might try to work off this orthogonality condition to derive an estimator for  $\gamma$ .

Since  $u_i = y_i - x_i' \gamma$ , we can write

$$E[z_i(y_i - x_i' \gamma)] = 0.$$

We might try to apply the method of moments (or “sample analog”) principle: replace population expectations by sample averages, and define our estimator  $\hat{\gamma}$  as the solution to:

$$\frac{1}{n} \sum_{i=1}^n z_i(y_i - x_i' \hat{\gamma}) = 0.$$

Note that  $z_i$  is an  $m \times 1$  vector, and  $y_i - x_i' \hat{\gamma}$  is a scalar. So the previous expression is really a system of  $m$  equations:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n z_{i1}(y_i - x_i' \hat{\gamma}) &= 0 \\ \frac{1}{n} \sum_{i=1}^n z_{i2}(y_i - x_i' \hat{\gamma}) &= 0 \\ &\vdots \\ \frac{1}{n} \sum_{i=1}^n z_{im}(y_i - x_i' \hat{\gamma}) &= 0 \end{aligned}$$

Since  $\gamma$  is  $k \times 1$ , this is a system of  $m$  equations in  $k$  unknowns.

In our example, we had  $m = k$ . So there are an equal number of unknowns as equations to solve. We call this the “just-identified case.” Typically, there will exist a unique solution for  $\hat{\gamma}$ . Let's focus on this case for the time being.

Assuming there is a unique solution, we must have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n z_i (y_i - x_i' \hat{\gamma}) &= 0 \\ \Rightarrow \sum_{i=1}^n z_i y_i - \left[ \sum_{i=1}^n z_i x_i' \right] \hat{\gamma} &= 0 \\ \Rightarrow \hat{\gamma} &= \left[ \sum_{i=1}^n z_i x_i' \right]^{-1} \sum_{i=1}^n z_i y_i. \end{aligned}$$

We call this the **linear instrumental variables** (IV) estimator. Sometimes, to highlight the distinction from OLS, we will write the estimator as  $\hat{\gamma}_{iv}$ .

Note the similarity with the OLS estimator. In fact, if there are no omitted variables, so that  $E[x_i u_i] = 0$ , we could set  $z_i = x_i$ . Then  $\hat{\gamma}$  would be the OLS estimator.

It's useful to write the IV estimator in vector-matrix form:

$$\hat{\gamma}_{iv} = (Z'X)^{-1}Z'y,$$

where

$$Z = \begin{pmatrix} z_1' \\ \vdots \\ z_n' \end{pmatrix}.$$

What are the properties of the IV estimator?

- $\hat{\gamma}_{iv}$  is biased:

Using  $y_i = x_i \gamma + u_i$ , we can write

$$\hat{\gamma}_{iv} = \left[ \sum_{i=1}^n z_i x_i' \right]^{-1} \sum_{i=1}^n z_i (x_i \gamma + u_i) = \gamma + \left[ \sum_{i=1}^n z_i x_i' \right]^{-1} \sum_{i=1}^n z_i u_i.$$

We might look at this and think that  $E[z_i u_i] = 0$  will imply unbiasedness, but this is not correct, because  $x_i$  is random and correlated with  $u_i$ . So if we focus on  $E[\hat{\gamma}_{iv} | Z]$ , the expression

$$E \left[ \left( \sum_{i=1}^n z_i x_i' \right)^{-1} \sum_{i=1}^n z_i u_i \mid Z \right] \neq E \left[ \left( \sum_{i=1}^n z_i x_i' \right)^{-1} \mid Z \right] E \left[ \sum_{i=1}^n z_i u_i \mid Z \right].$$

- $\hat{\gamma}_{iv}$  is consistent:

We can write

$$\begin{aligned} \hat{\gamma}_{iv} &= \gamma + \left[ \frac{1}{n} \sum_{i=1}^n z_i x_i' \right]^{-1} \frac{1}{n} \sum_{i=1}^n z_i u_i \\ &\xrightarrow{P} \gamma + E[z_i x_i']^{-1} E[z_i u_i] \end{aligned}$$

provided that we can assume the relevant moments exist and that the LLN applies.

- $\hat{\gamma}_{iv}$  is asymptotically normally distributed:

Write:

$$\sqrt{n}(\hat{\gamma}_{iv} - \gamma) = \left[ \frac{1}{n} \sum_{i=1}^n z_i x_i' \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n z_i u_i.$$

Suppose that  $(x_i, z_i, u_i)$  are IID from a joint distribution, and that its fourth moments are finite. Then the vector  $(z_i u_i)$  is IID, with  $E[z_i u_i] = 0$ , and

$$V[z_i u_i] = E[z_i u_i (z_i u_i)'] = E[u_i^2 z_i z_i'].$$

(Recall that  $u_i$  is scalar.) By the central limit theorem,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n z_i u_i \xrightarrow{d} N(0, E[u_i^2 z_i z_i']).$$

So

$$\sqrt{n}(\hat{\gamma}_{iv} - \gamma) \xrightarrow{d} N(0, E[z_i x_i']^{-1} E[u_i^2 z_i z_i'] E[x_i z_i']^{-1}).$$

To estimate the variance of  $\hat{\gamma}_{iv}$ , we can form

$$\hat{u}_i = y_i - x_i' \hat{\gamma}_{iv}.$$

Then we can replace the expectations with sample analogs:

$$\hat{V} = \left[ \frac{1}{n} \sum_{i=1}^n z_i x_i' \right]^{-1} \left( \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 z_i z_i' \right) \left[ \frac{1}{n} \sum_{i=1}^n x_i z_i' \right]^{-1}.$$

Note: if we assume that  $E[u_i^2 | z_i] = \sigma_u^2$ , a type of homoskedasticity, then

$$E[u_i^2 z_i z_i'] = E[E[u_i^2 | z_i] z_i z_i'] = \sigma_u^2 E[z_i z_i'],$$

and we can estimate the middle term of the variance by

$$\hat{\sigma}_u^2 = \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2,$$

$$\hat{E}[u_i^2 z_i z_i'] = \hat{\sigma}_u^2 \cdot \frac{1}{n} \sum_{i=1}^n z_i z_i'.$$