

## LECTURE 10

## ENDOGENEITY AND INSTRUMENTAL VARIABLES ESTIMATION

## Endogeneity

Consider the partitioned regression model:

$$\begin{aligned} Y_i &= X_i^\top \beta + U_i \\ &= X_{1i}^\top \beta_1 + X_{2i}^\top \beta_2 + U_i, \end{aligned} \tag{1}$$

where  $X_{1i}$  is a  $k_1$ -vector and  $X_{2i}$  is a  $k_2$ -vector of random regressors,  $\beta_1$  and  $\beta_2$  are  $k_1 \times 1$  and  $k_2 \times 1$  vectors of unknown parameters, respectively, with  $k_1 + k_2 = k$ . We assume that  $X_{1i}$  is *endogenous*:

$$E[X_{1i}U_i] \neq 0,$$

as opposed to the (*weakly*) *exogenous* regressors  $X_{2i}$ :

$$E[X_{2i}U_i] = 0.$$

(The assumption  $E[U_i | X_{2i}] = 0$  is called *strong exogeneity*.) Sources of endogeneity:

- **Omitted variables.** Consider the wage equation:

$$\begin{aligned} \log(\text{Wage}_i) &= \alpha + \beta \text{Education}_i + \gamma \text{Gender}_i + \delta \text{Ability}_i + V_i \\ &= \alpha + \beta \text{Education}_i + \gamma \text{Gender}_i + U_i. \end{aligned}$$

Since ability is unobservable, it enters the error term  $U_i = \delta \text{Ability}_i + V_i$ . The gender variable is plausibly exogenous; however, education is correlated with ability and therefore endogenous.

- **Errors in variables.** Suppose that the true model is

$$Y_i = \tilde{X}_{1i}^\top \beta + X_{2i}^\top \beta_2 + V_i;$$

however,  $\tilde{X}_{1i}$  is unobservable. Instead, the econometrician observes  $X_{1i} = \tilde{X}_{1i} + \varepsilon_i$ , where  $\varepsilon_i$  is some noise vector independent of  $\tilde{X}_{1i}$  and  $X_{2i}$ . Substituting  $\tilde{X}_{1i} = X_{1i} - \varepsilon_i$  into the above equation,

$$Y_i = X_{1i}^\top \beta + X_{2i}^\top \beta_2 - \varepsilon_i^\top \beta + V_i.$$

Set  $U_i = -\varepsilon_i^\top \beta + V_i$ . While  $X_{2i}$  is exogenous,  $X_{1i}$  is endogenous, because it is correlated with  $U_i$  through  $\varepsilon_i$ .

- **Simultaneity.** Consider the following equation:

$$\text{Hours}_i = \beta_1 \text{Children}_i + X_{2i}^\top \beta_2 + U_i,$$

where  $\text{Hours}_i$  is the hours of work per week,  $\text{Children}_i$  is the number of children in the family, and  $X_{2i}$  is a vector of exogenous variables. While the number of children affects labor supply, it is reasonable to assume that career decisions affect family size. Accordingly, there is another equation determining the number of children in the family:

$$\text{Children}_i = \gamma_1 \text{Hours}_i + Z_{1i}^\top \gamma_2 + V_i,$$

where  $Z_{1i}$  is another vector of exogenous variables. Substituting the expression for the hours into the equation for the number of children, we obtain (assuming that  $1 - \beta_1 \gamma_1 \neq 0$ )

$$\text{Children}_i = X_{2i}^\top \left( \frac{\beta_2 \gamma_1}{1 - \beta_1 \gamma_1} \right) + Z_{1i}^\top \frac{\gamma_2}{1 - \beta_1 \gamma_1} + \frac{\gamma_1}{1 - \beta_1 \gamma_1} U_i + \frac{1}{1 - \beta_1 \gamma_1} V_i.$$

Assuming that  $X_{2i}$ ,  $Z_{1i}$ , and  $V_i$  are uncorrelated with  $U_i$ ,

$$\begin{aligned} E[U_i \text{Children}_i] &= \frac{\gamma_1}{1 - \beta_1 \gamma_1} E[U_i^2] \\ &\neq 0. \end{aligned}$$

## Properties of the OLS under endogeneity

Consider first the OLS estimator of  $\beta_1$ :

$$\begin{aligned}\widehat{\beta}_{1n} &= (X_1^\top M_2 X_1)^{-1} X_1^\top M_2 Y \\ &= \beta_1 + (X_1^\top M_2 X_1)^{-1} X_1^\top M_2 U,\end{aligned}$$

where  $M_2 = I_n - X_2 (X_2^\top X_2)^{-1} X_2^\top$ . Then

$$\begin{aligned}n^{-1} X_1^\top M_2 X_1 &= n^{-1} \sum_{i=1}^n X_{1i} X_{1i}^\top - n^{-1} \sum_{i=1}^n X_{1i} X_{2i}^\top \left( n^{-1} \sum_{i=1}^n X_{2i} X_{2i}^\top \right)^{-1} n^{-1} \sum_{i=1}^n X_{2i} X_{1i}^\top, \\ n^{-1} X_1^\top M_2 U &= n^{-1} \sum_{i=1}^n X_{1i} U_i - n^{-1} \sum_{i=1}^n X_{1i} X_{2i}^\top \left( n^{-1} \sum_{i=1}^n X_{2i} X_{2i}^\top \right)^{-1} n^{-1} \sum_{i=1}^n X_{2i} U_i.\end{aligned}$$

Assume that:

- $\{(Y_i, X_i) : i \geq 1\}$  are iid.
- $E[X_{i,j}^2] < \infty$  for all  $j = 1, \dots, k$ .
- $E[X_i X_i^\top]$  is positive definite.
- $E[U_i^2] < \infty$ .

By the WLLN, we have

$$\begin{aligned}n^{-1} \sum_{i=1}^n X_{1i} X_{1i}^\top &\rightarrow_p E[X_{1i} X_{1i}^\top], \\ n^{-1} \sum_{i=1}^n X_{1i} X_{2i}^\top &\rightarrow_p E[X_{1i} X_{2i}^\top], \\ n^{-1} \sum_{i=1}^n X_{2i} X_{2i}^\top &\rightarrow_p E[X_{2i} X_{2i}^\top], \\ n^{-1} \sum_{i=1}^n X_{2i} U_i &\rightarrow_p 0, \\ n^{-1} \sum_{i=1}^n X_{1i} U_i &\rightarrow_p E[X_{1i} U_i].\end{aligned}$$

Thus,

$$\begin{aligned}n^{-1} X_1^\top M_2 X_1 &\rightarrow_p E[X_{1i} X_{1i}^\top] - E[X_{1i} X_{2i}^\top] \left( E[X_{2i} X_{2i}^\top] \right)^{-1} E[X_{2i} X_{1i}^\top], \\ n^{-1} X_1^\top M_2 U &\rightarrow_p E[X_{1i} U_i] - E[X_{1i} X_{2i}^\top] \left( E[X_{2i} X_{2i}^\top] \right)^{-1} E[X_{2i} U_i] \\ &= E[X_{1i} U_i] \\ &\neq 0,\end{aligned}$$

and we conclude that  $\widehat{\beta}_{1n}$  is inconsistent:

$$\begin{aligned}\widehat{\beta}_{1n} &\rightarrow_p \beta_1 + \left( E[X_{1i} X_{1i}^\top] - E[X_{1i} X_{2i}^\top] \left( E[X_{2i} X_{2i}^\top] \right)^{-1} E[X_{2i} X_{1i}^\top] \right)^{-1} E[X_{1i} U_i] \\ &\neq \beta_1.\end{aligned}$$

The OLS estimator of  $\beta_2$  is also inconsistent, as the following argument shows. We have

$$\widehat{\beta}_{2n} = \beta_2 + (X_2^\top M_1 X_2)^{-1} X_2^\top M_1 U,$$

where  $M_1 = I_n - X_1 (X_1^\top X_1)^{-1} X_1^\top$ . Then

$$\begin{aligned} \widehat{\beta}_{2n} &\rightarrow_p \beta_2 - \left( E[X_{2i} X_{2i}^\top] - E[X_{2i} X_{1i}^\top] \left( E[X_{1i} X_{1i}^\top] \right)^{-1} E[X_{1i} X_{2i}^\top] \right)^{-1} E[X_{2i} X_{1i}^\top] \left( E[X_{1i} X_{1i}^\top] \right)^{-1} E[X_{1i} U_i] \\ &\neq \beta_2. \end{aligned}$$

## Instrumental Variables Estimation

Let  $Z_{1i}$  be a  $k_1$ -vector of *exogenous* variables:

$$E[Z_{1i} U_i] = 0.$$

It is important that  $Z_{1i}$  is *excluded* from the model (1), that is,  $Z_{1i}$  does not contain any of the elements of  $X_{2i}$ . Define

$$\begin{aligned} X_i &= \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix}, \\ Z_i &= \begin{pmatrix} Z_{1i} \\ X_{2i} \end{pmatrix}. \end{aligned}$$

Here  $X_i$  is the  $k$ -vector of regressors and  $Z_i$  is the  $k$ -vector of *instrumental variables* (IVs). The exogenous regressors appear again in the vector of IVs, and for each endogenous regressor, we include an exogenous variable (IV) that must be excluded from the model  $Y_i = X_i^\top \beta + U_i$ . When all regressors are endogenous ( $k_1 = k$ ),  $X_i$  and  $Z_i$  share no elements.

We assume that the IVs are informative about the regressors. This is expressed as the following *rank condition*:

$$\text{rank}\left(E[Z_i X_i^\top]\right) = k. \tag{2}$$

The rank condition in (2) will fail if, for example,  $E[Z_{1i} X_i^\top] = 0$  ( $Z_{1i}$  is exogenous but merely random noise). The rank condition will also fail if some of the elements of  $Z_{1i}$  are linear combinations of the elements of the included exogenous regressors  $X_{2i}$ .

**Example.** Consider the Hours/Children example. Angrist and Evans (1998) suggested using the sex composition of the first two children as an instrument for the number of children in the family (the sample was restricted to women with at least two children). This is motivated by the observation that if the first two children are of the same sex (boy-boy or girl-girl), the family is more likely to have a third child than in the opposite-sex case (boy-girl or girl-boy), so the dummy variable indicating same-sex siblings is positively correlated with the total number of children (relevance condition). The exclusion restriction is that the sex composition of the first two children affects labor supply only through its effect on fertility; there is no direct channel from sibling sex mix to the mother's labor supply. Since sex composition is determined randomly, the instrument is also uncorrelated with unobserved determinants of labor supply.

Since

$$E[Z_i U_i] = 0,$$

the MM principle suggests an estimator that solves the following system of  $k$  equations:

$$n^{-1} \sum_{i=1}^n Z_i \left( Y_i - X_i^\top \widehat{\beta}_n^{IV} \right) = 0, \text{ or}$$

$$\widehat{\beta}_n^{IV} = \left( \sum_{i=1}^n Z_i X_i^\top \right)^{-1} \sum_{i=1}^n Z_i Y_i$$

$$= (Z^\top X)^{-1} Z^\top Y.$$

The estimator  $\widehat{\beta}_n^{IV}$  is called the IV estimator of  $\beta$ .

Next, we show consistency and asymptotic normality of the IV estimator. We assume:

- $\{(Y_i, X_i, Z_i) : i \geq 1\}$  are iid.
- $E[Z_i U_i] = 0$ .
- $E[X_{i,j}^2] < \infty$  for all  $j = 1, \dots, k$ .
- $E[Z_{i,j}^2] < \infty$  for all  $j = 1, \dots, k$ .
- $E[Z_i X_i^\top]$  is of rank  $k$ .
- $E[U_i^2 Z_i Z_i^\top]$  is positive definite.

Write

$$\widehat{\beta}_n^{IV} = \beta + \left( n^{-1} \sum_{i=1}^n Z_i X_i^\top \right)^{-1} n^{-1} \sum_{i=1}^n Z_i U_i. \quad (3)$$

By the Cauchy–Schwarz inequality,

$$E |Z_{i,r} X_{i,s}| \leq \sqrt{E[Z_{i,r}^2] E[X_{i,s}^2]}$$

$$< \infty \text{ for all } r, s = 1, \dots, k.$$

Therefore, by Slutsky’s Theorem,

$$\widehat{\beta}_n^{IV} \rightarrow_p \beta + \left( E[Z_i X_i^\top] \right)^{-1} E[Z_i U_i]$$

$$= \beta.$$

To show asymptotic normality, we assume in addition that

- $E[Z_{i,j}^4] < \infty$  for all  $j = 1, \dots, k$ .
- $E[U_i^4] < \infty$ .

Write (3) as

$$n^{1/2} \left( \widehat{\beta}_n^{IV} - \beta \right) = \left( n^{-1} \sum_{i=1}^n Z_i X_i^\top \right)^{-1} n^{-1/2} \sum_{i=1}^n Z_i U_i.$$

As in Lecture 7, for all  $r, s = 1, \dots, k$ ,

$$E |U_i^2 Z_{i,r} Z_{i,s}| \leq \left( E[U_i^4] \right)^{1/2} \left( E[Z_{i,r}^4] E[Z_{i,s}^4] \right)^{1/4}$$

$$< \infty.$$

By the CLT and Cramér convergence theorem,

$$\begin{aligned} n^{1/2} \left( \widehat{\beta}_n^{IV} - \beta \right) &\rightarrow_d \left( \mathbb{E}[Z_i X_i^\top] \right)^{-1} N \left( 0, \mathbb{E}[U_i^2 Z_i Z_i^\top] \right) \\ &= N \left( 0, \left( \mathbb{E}[Z_i X_i^\top] \right)^{-1} \mathbb{E}[U_i^2 Z_i Z_i^\top] \left( \mathbb{E}[X_i Z_i^\top] \right)^{-1} \right). \end{aligned}$$

The asymptotic covariance matrix takes the sandwich form and can be estimated consistently by

$$\left( n^{-1} \sum_{i=1}^n Z_i X_i^\top \right)^{-1} n^{-1} \sum_{i=1}^n \widehat{U}_i^2 Z_i Z_i^\top \left( n^{-1} \sum_{i=1}^n X_i Z_i^\top \right)^{-1},$$

where  $\widehat{U}_i = Y_i - X_i^\top \widehat{\beta}_n^{IV}$ .

## Weak Instruments

Consider the following regression model with a single (endogenous) regressor:

$$Y_i = \beta X_i + U_i.$$

We assume that the endogenous regressor is related to a single IV through the following equation:

$$\begin{aligned} X_i &= \pi_n Z_i + V_i, \\ \pi_n &= \frac{c}{\sqrt{n}}. \end{aligned}$$

This parameterization captures a weak but nonzero correlation between  $X_i$  and its instrument  $Z_i$ . We will rely on a large- $n$  approximation for the distribution of the estimators, and, therefore, weakness of the relationship between  $X$  and  $Z$  has to be modeled in terms of the sample size  $n$ . This is because any fixed  $\pi \neq 0$  will be “large” when  $n \rightarrow \infty$ . Therefore, we assume that  $\pi_n \rightarrow 0$  as  $n \rightarrow \infty$ . The rate of convergence is chosen in such a way that small correlations captured by the nonzero constant  $c$  will appear in the limit. In addition, we assume that the IV is uncorrelated with the errors  $U$  and  $V$ , and that the errors are homoskedastic (conditional on  $Z$ ). The homoskedasticity assumption is not essential and is made only for simplicity.

- $\{(Y_i, X_i, Z_i) : i = 1, \dots, n\}$  are iid.
- $\mathbb{E} \left[ Z_i \begin{pmatrix} U_i \\ V_i \end{pmatrix} \right] = 0$ .
- $\mathbb{E} \left[ \begin{pmatrix} U_i \\ V_i \end{pmatrix} \begin{pmatrix} U_i \\ V_i \end{pmatrix}^\top \mid Z_i \right] = \begin{pmatrix} \sigma_U^2 & \sigma_{UV} \\ \sigma_{UV} & \sigma_V^2 \end{pmatrix} = \Sigma$ , a positive definite matrix.
- $\mathbb{E}[Z_i^2] = Q < \infty$ .

Under the above assumptions,

$$\begin{aligned} n^{-1} \sum_{i=1}^n Z_i^2 &\rightarrow_p Q, \\ n^{-1/2} \sum_{i=1}^n Z_i \begin{pmatrix} U_i \\ V_i \end{pmatrix} &\rightarrow_d N \left( 0, Q \begin{pmatrix} \sigma_U^2 & \sigma_{UV} \\ \sigma_{UV} & \sigma_V^2 \end{pmatrix} \right) \\ &\equiv \begin{pmatrix} \Psi_U \\ \Psi_V \end{pmatrix}. \end{aligned}$$

These results follow from the WLLN and the CLT, respectively. In the result above,  $\Psi_U$  and  $\Psi_V$  are jointly normal random variables with zero means, covariance  $\sigma_{UV}Q$ , and variances  $\sigma_U^2Q$  and  $\sigma_V^2Q$ , respectively.

Next, the IV estimator of  $\beta$  is given by

$$\begin{aligned}\widehat{\beta}_n^{IV} - \beta &= \frac{\sum_{i=1}^n Z_i Y_i}{\sum_{i=1}^n Z_i X_i} - \beta \\ &= \frac{\sum_{i=1}^n Z_i U_i}{\sum_{i=1}^n Z_i (\pi_n Z_i + V_i)} \\ &= \frac{\sum_{i=1}^n Z_i U_i}{cn^{-1/2} \sum_{i=1}^n Z_i^2 + \sum_{i=1}^n Z_i V_i} \\ &= \frac{n^{-1/2} \sum_{i=1}^n Z_i U_i}{cn^{-1} \sum_{i=1}^n Z_i^2 + n^{-1/2} \sum_{i=1}^n Z_i V_i} \\ &\rightarrow_d \frac{\Psi_U}{cQ + \Psi_V},\end{aligned}$$

where the last result follows by the CMT and the joint convergence in distribution of  $n^{-1/2} \sum_{i=1}^n Z_i U_i$  and  $n^{-1/2} \sum_{i=1}^n Z_i V_i$ . Hence, due to the weak IV problem, the IV estimator of  $\beta$  is not consistent, and converges instead to a random limit. Furthermore, due to the inconsistency of  $\widehat{\beta}_n^{IV}$ , confidence intervals and hypothesis tests based on  $\widehat{\beta}_n^{IV}$  are invalid: the asymptotic coverage probabilities and size are incorrect. Moreover, the asymptotic distribution of  $\widehat{\beta}_n^{IV}$  depends on unknown parameters  $c$  and  $\Sigma$  that cannot be consistently estimated under these circumstances. Hence, the distribution of  $\widehat{\beta}_n^{IV} - \beta$  cannot be evaluated.

Although consistent point estimation is impossible in this setting, hypotheses about  $\beta$  can still be tested. This can be done with the Anderson-Rubin (AR) statistic. Consider the null hypothesis  $H_0 : \beta = \beta_0$ . The null-restricted residuals are given by

$$U_{0,i} = Y_i - \beta_0 X_i.$$

Provided that the null hypothesis is true, we have the true residuals, and therefore

$$\begin{aligned}n^{-1} \sum_{i=1}^n U_{0,i}^2 &\rightarrow_p \sigma_U^2, \\ n^{-1/2} \sum_{i=1}^n Z_i U_{0,i} &\rightarrow_d \Psi_U \\ &\stackrel{d}{=} N(0, \sigma_U^2 Q).\end{aligned}$$

The AR statistic is given by

$$AR_n(\beta_0) = \frac{n^{-1/2} \sum_{i=1}^n Z_i (Y_i - \beta_0 X_i)}{\sqrt{(n^{-1} \sum_{i=1}^n U_{0,i}^2) (n^{-1} \sum_{i=1}^n Z_i^2)}}.$$

Under the null hypothesis,

$$AR_n(\beta_0) \rightarrow_d N(0, 1).$$

(What we derived above is a simplified scalar version of the AR statistic for the case of a single IV and no exogenous regressors. The general AR statistic accommodates multiple IVs and exogenous regressors; it is constructed as a quadratic form and has a chi-squared asymptotic distribution under the null.) One should reject the null in favor of the alternative, say  $H_1 : \beta \neq \beta_0$ , if  $|AR_n(\beta_0)| > z_{1-\alpha/2}$ , where  $z_\tau$  is the  $\tau$  quantile of the standard normal distribution. Such a test will have asymptotic size  $\alpha$  regardless of the strength of the IV. Furthermore, one can construct confidence intervals for  $\beta$  by collecting all values  $\beta_0$  for which the AR test cannot reject the null:

$$CI_{1-\alpha,n} = \{\beta_0 : |AR_n(\beta_0)| \leq z_{1-\alpha/2}\}.$$

If the null hypothesis is false, then  $U_{0,i}$ 's are not the true residuals, but a function of  $X_i$ 's and the residuals:

$$U_{0,i} = U_i + (\beta - \beta_0) X_i.$$

If the IV  $Z_i$  is related to the regressor  $X_i$ , that is,  $c \neq 0$ , then the asymptotic distribution of the AR statistic will be different from the standard normal, and the test will have power to reject the null. The power of the test depends on the distance from the null  $|\beta - \beta_0|$  and the strength of the IV, which is captured by  $c$ . If  $c = 0$ , that is, the IV is irrelevant, then the AR test has no power to reject false null hypotheses. The test will reject the null with probability  $\alpha$ , equal to the size, regardless of the value of  $\beta$  (the power function is flat and equal to the size,  $\alpha$ , for all values of  $\beta$ ). In this case, the confidence interval described above will have infinite length (it will include all values  $\beta_0 \in \mathbb{R}$ ).

The approach can be extended to the case of IV regression with multiple endogenous and exogenous regressors, as well as heteroskedastic errors. However, in the case of weak IVs, in general, one cannot test hypotheses on subvectors of  $\beta$  using the AR test. Because the coefficient estimators are inconsistent, obtaining true null-restricted residuals requires specifying the values of all coefficients under the null.