LECTURE 10

ENDOGENEITY AND INSTRUMENTAL VARIABLES ESTIMATION

Endogeneity

Consider a partitioned regression model:

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

where X_{1i} is a k_1-vector and X_{2i} is a k_2-vector of random regressors, β_1 is $k_1 \times 1$ and β_2 is $k_2 \times 1$ vectors of unknown parameters, $k_1 + k_2 = k$. We assume that X_{1i} is *endogenous*:

$$E(X_{1i}U_i) \neq 0,$$

as opposed to *(weakly) exogenous* X_{2i}'s:

$$E(X_{2i}U_i) = 0.$$

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

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

$$\begin{aligned} \log Wage_i &= \alpha + \beta Education_i + \gamma Gender_i + \delta Ability_i + V_i \\ &= \alpha + \beta Education_i + \gamma Gender_i + U_i. \end{aligned}$$

  Since ability is unobservable, it "goes" to the residuals $U_i = \delta Ability_i + V_i$. We can assume that the gender variable is exogenous, however, education is correlated with the ability, and, therefore, education is endogenous.

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

$$Y_i = \widetilde{X}_{1i}'\beta + X_{2i}'\beta_2 + V_i,$$

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

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

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

- **Simultaneity.** Consider the following equation

$$Hours_i = \beta_1 Children_i + X_{2i}'\beta_2 + U_i,$$

  where $Hours_i$ is the hours of work per week, and $Children_i$ 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, i.e. there is another equation determining the number of children in the family:

$$Children_i = \gamma_1 Hours_i + Z_{1i}'\gamma_2 + V_i,$$

  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 $1 - \beta_1\gamma_1 \neq 0$)

$$Children_i = X_{2i}'\left(\frac{\beta_2\gamma_1}{1-\beta_1\gamma_1}\right) + Z_{1i}'\frac{\gamma_2}{1-\beta_1\gamma_1} + \frac{\gamma_1}{1-\beta_1\gamma_1}U_i + \frac{1}{1-\beta_1\gamma_1}V_i.$$

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

$$\begin{aligned} E(U_i Children_i) &= \frac{\gamma_1}{1-\beta_1\gamma_1}EU_i^2 \\ &\neq 0. \end{aligned}$$

Properties of the OLS under endogeneity

Consider first the OLS estimator of β_1:

$$\begin{aligned}\widehat{\beta}_{1n} &= (X_1'M_2X_1)^{-1}X_1'M_2Y \\ &= \beta_1 + (X_1'M_2X_1)^{-1}X_1'M_2U,\end{aligned}$$

where $M_2 = I_n - X_2(X_2'X_2)^{-1}X_2'$. We have

$$\begin{aligned}n^{-1}X_1'M_2X_1 &= n^{-1}\sum_{i=1}^n X_{1i}X_{1i}' - n^{-1}\sum_{i=1}^n X_{1i}X_{2i}'\left(n^{-1}\sum_{i=1}^n X_{2i}X_{2i}'\right)^{-1} n^{-1}\sum_{i=1}^n X_{2i}X_{1i}', \\ n^{-1}X_1'M_2U &= n^{-1}\sum_{i=1}^n X_{1i}U_i - n^{-1}\sum_{i=1}^n X_{1i}X_{2i}'\left(n^{-1}\sum_{i=1}^n X_{2i}X_{2i}'\right)^{-1} n^{-1}\sum_{i=1}^n X_{2i}U_i\end{aligned}$$

Assume that:

- $\{(Y_i, X_i) : i \geq 1\}$ are iid
- $EX_{i,j}^2 < \infty$ for all $j = 1, \ldots k$.
- EX_iX_i' positive definite.
- $EU_i^2 < \infty$.

By the WLLN we have

$$\begin{aligned}n^{-1}\sum_{i=1}^n X_{1i}X_{1i}' &\to_p EX_{1i}X_{1i}', \\ n^{-1}\sum_{i=1}^n X_{1i}X_{2i}' &\to_p EX_{1i}X_{2i}', \\ n^{-1}\sum_{i=1}^n X_{2i}X_{2i}' &\to_p EX_{2i}X_{2i}', \\ n^{-1}\sum_{i=1}^n X_{2i}U_i &\to_p 0, \\ n^{-1}\sum_{i=1}^n X_{1i}U_i &\to_p EX_{1i}U_i.\end{aligned}$$

Thus,

$$\begin{aligned}n^{-1}X_1'M_2X_1 &\to_p EX_{1i}X_{1i}' - EX_{1i}X_{2i}'(EX_{2i}X_{2i}')^{-1}EX_{2i}X_{1i}', \\ n^{-1}X_1'M_2U &\to_p EX_{1i}U_i - EX_{1i}X_{2i}'(EX_{2i}X_{2i}')^{-1}EX_{2i}U_i \\ &= EX_{1i}U_i \\ &\neq 0,\end{aligned}$$

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

$$\begin{aligned}\widehat{\beta}_{1n} &\to_p \beta_1 + \left(EX_{1i}X_{1i}' - EX_{1i}X_{2i}'(EX_{2i}X_{2i}')^{-1}EX_{2i}X_{1i}'\right)^{-1}EX_{1i}U_i \\ &\neq \beta_1.\end{aligned}$$

Inconsistency of the OLS estimator of β_2 can be shown similarly. We have

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

where $M_1 = I_n - X_1 (X_1' X_1)^{-1} X_1'$. We have

$$\begin{aligned} \widehat{\beta}_{2n} &\to_p \beta_2 - \left(EX_{2i}X_{2i}' - EX_{2i}X_{1i}' (EX_{1i}X_{1i}')^{-1} EX_{1i}X_{2i}' \right)^{-1} EX_{2i}X_{1i}' (EX_{1i}X_{1i}')^{-1} EX_{1i}U_i \\ &\neq \beta_2. \end{aligned}$$

Instrumental Variables estimation

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

$$EZ_{1i}U_i = 0.$$

It is important that Z_{1i} is *excluded* from the model (1), i.e. 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). Note that the exogenous regressors appear again in the vector of IVs, and for each endogenous regressor we bring an exogenous variable (IV) that must be excluded from the model $Y_i = X_i'\beta + U_i$. When all regressors are endogenous, $k_1 = k$ and we do not have any overlapping elements between X_i and Z_i.

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

$$rank\,(EZ_i X_i') = k. \tag{2}$$

The rank condition in (2) will fail if, for example, $EZ_{1i}X_i' = 0$ (Z_{1i} is exogenous but 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 to use the sex composition of the first two children as an instrument to 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 case (boy-girl or girl-boy). Consequently, the dummy variable for the first two children are of the same sex has to be positively correlated with the total number of children. On the other hand, the instrument is uncorrelated with the errors, because sex composition is determined randomly.

We have that

$$EZ_i U_i = 0.$$

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

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} Z_i \left(Y_i - X_i' \widehat{\beta}_n^{IV} \right) &= 0, \text{ or} \\ \widehat{\beta}_n^{IV} &= \left(\sum_{i=1}^{n} Z_i X_i' \right)^{-1} \sum_{i=1}^{n} Z_i Y_i \\ &= (Z'X)^{-1} Z'Y. \end{aligned}$$

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

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

- $\{(Y_i, X_i, Z_i) : i \geq 1\}$ are iid.
- $EZ_iU_i = 0$.
- $EX_{i,j}^2 < \infty$ for all $j = 1, \ldots k$.
- $EZ_{i,j}^2 < \infty$ for all $j = 1, \ldots k_1$.
- EZ_iX_i' is of rank k.
- $EU_i^2 Z_i Z_i'$ is positive definite.

Write

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

Note that, under the above assumptions, by the Cauchy-Shwartz inequality

$$\begin{aligned} E\left|Z_{i,r}X_{i,s}\right| &\leq \sqrt{EZ_{i,r}^2 EX_{i,s}^2} \\ &< \infty \text{ for all } r, s = 1, \ldots k. \end{aligned}$$

Therefore, by the Slutsky's Theorem,

$$\begin{aligned} \widehat{\beta}_n^{IV} &\to_p \beta + \left(EZ_iX_i'\right)^{-1} EZ_iU_i \\ &= \beta. \end{aligned}$$

In order to show the asymptotic normality, we assume in addition that

- $EZ_{i,j}^4 < \infty$ for all $j = 1, \ldots k$.
- $EU_i^4 < \infty$.

Write (3) as

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

Similarly to the results in Lecture 7, for all $r, s = 1, \ldots, k$,

$$\begin{aligned} E\left|U_i^2 Z_{i,r} Z_{i,s}\right| &\leq \left(EU_i^4\right)^{1/2} \left(EZ_{i,r}^4 EZ_{i,s}^4\right)^{1/4} \\ &< \infty. \end{aligned}$$

Therefore, by the CLT and Cramer Convergence Theorem,

$$\begin{aligned} n^{1/2}\left(\widehat{\beta}_n^{IV} - \beta\right) &\to_d \left(EZ_iX_i'\right)^{-1} N\left(0, \left(EU_i^2 Z_i Z_i'\right)\right) \\ &= N\left(0, \left(EZ_iX_i'\right)^{-1} \left(EU_i^2 Z_i Z_i'\right) \left(EX_iZ_i'\right)^{-1}\right). \end{aligned}$$

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

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

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

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 model defines weak, but different from zero correlation between the endogenous regressor X and its instrument Z. We will rely on large n approximation for the distribution of the estimators, and, therefore, weakness of the relationship between X and Z has to be modelled in terms of the sample size n. This is because any fixed (independent of n) π, will be "large" when $n \to \infty$ as long as it is different from zero. Therefore, we assume that $\pi_n \to 0$ as $n \to \infty$. The rate of convergence is chosen in a such way so that small correlations captured by the non-zero 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 crucial, it is made here only for simplicity.

- $\{(Y_i, X_i, Z_i) : i = 1, \ldots, n\}$ are iid.
- $EZ_i \begin{pmatrix} U_i \\ V_i \end{pmatrix} = 0.$
- $E\left(\begin{pmatrix} U_i \\ V_i \end{pmatrix} \begin{pmatrix} U_i \\ V_i \end{pmatrix}' | Z_i \right) = \begin{pmatrix} \sigma_U^2 & \sigma_{UV} \\ \sigma_{UV} & \sigma_V^2 \end{pmatrix} = \Sigma$, a finite and positive definite matrix.
- $EZ_i^2 = Q < \infty.$

With above assumptions, we have that

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

The results follow by the WLLN and CLT respectively. In the result above result, Ψ_U and Ψ_V denote any bivariate normal random variables with zero means, covariance $\sigma_{UV} Q$, and variances σ_U^2 and σ_V^2 respectively. Note that the above result gives joint convergence in distribution of $n^{-1/2} \sum_{i=1}^n Z_i U_i$ and $n^{-1/2} \sum_{i=1}^n Z_i V_i$.

Next, the IV estimator of β is given by

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

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 β is not consistent, and converges instead to a random limit. Furthermore, due to inconsistency of $\widehat{\beta}_n^{IV}$, confidence intervals and hypotheses testing procedures based on $\widehat{\beta}_n^{IV}$ are invalid (incorrect asymptotic coverage probabilities and size). Note also that the asymptotic distribution of $\widehat{\beta}_n^{IV}$ depends on unknown parameters c and Σ that cannot be consistently estimated under these circumstances . Hence, the distribution of $\widehat{\beta}_n^{IV} - \beta$ cannot be evaluated.

Despite the fact that consistent point estimation in this situation is impossible, it turns out that one can still test hypotheses concerning β. This can be done with, so called, 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

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

The AR statistic is given by

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

We have that under the null hypothesis,

$$AR_n\left(\beta_0\right) \to_d N\left(0,1\right).$$

(The statistic above actually is not exactly the AR statistic, but is based on the same idea. The actual AR statistic is designed for a model with more than one IV and also allows for exogenous regressors and therefore constructed as a quadratic form and has a chi-square asymptotic distribution under the null.) One should reject the null in favor of alternative, say $H_1 : \beta \neq \beta_0$, if $|AR_n\left(\beta_0\right)| > z_{1-\alpha/2}$, where z_τ is the τ quantile of the standard normal distribution. Such a test will have asymptotic size α regardless the strength of the IV. Furthermore, one can construct confidence intervals for β by collecting all values β_0 for which the AR test cannot reject the null:

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

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

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

If the IV Z_i is related to the regressor X_i, i.e. $c \neq 0$, 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. One can show that 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 $c = 0$, i.e. the IV is irrelevant, the AR test will have no power to reject false null hypotheses. The test will always be rejecting the null with probability α equal to the size, regardless of the value of β (the power function is flat and equal to the size, α, for all values of β). In this case, the confidence interval described above will have infinite length (it will include all values $\beta_0 \in R$).

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 sub-vectors of coefficients using the AR test. The reason for this is that coefficient estimators are inconsistent, and therefore, in order to have true null restricted residuals, one has to specify the values of all coefficients under the null.