Lecture 10
Instrumental variables

Lectures in SDPE: *Econometrics I* on February 26, 2024

Markus Jäntti
Swedish Institute For Social Research
Stockholm University

Introduction

- We have endogeneity if

$$Y = \boldsymbol{X}'\beta + e, \text{ and } \mathrm{E}[\boldsymbol{X}e] \neq 0. \tag{1}$$

- This is meaningful only relative to more structure on the problem than that provided by just the linear projection interpretation.
- That is, β needs to have a *structural* interpretation.
- Substantively, we should distinguish between different types of endogeneity, such as:
  - measurement errors in regressors
  - two-way causation
  - simultaneity
  - omitted variables
- Much of this chapter will devoted to the "classical" IV setup in economics involving simultaneous equations.
- Much of the *estimation* and *asymptotic distribution* for IV will be covered under generalized method of moments.

- Let $(Y, \boldsymbol{X}^*)$ be the random variables of interest and the object of study the conditional expectation $\mathrm{E}[Y|\boldsymbol{X}^*] = \boldsymbol{X}^{*\prime}\beta$.
- We observe an error-ridden version $\boldsymbol{X} = \boldsymbol{X}^* + \mathbf{u}$ with the $k \times 1$ vector of measurement errors $\mathbf{u}$, independent of Y and $\boldsymbol{X}^*$.
- Our regression is

$$Y = \boldsymbol{X}^{*\prime}\beta + e = (\boldsymbol{X} - \mathbf{u})'\beta + e = \boldsymbol{X}'\beta + (e - \mathbf{u}'\beta) = \boldsymbol{X}'\beta + v \qquad (2)$$

- Our moment condition is

$$\mathrm{E}[\boldsymbol{X}v] = \mathrm{E}[(\boldsymbol{X}^* + \mathbf{u})(e - \mathbf{u}'\beta)] = -\mathrm{E}[\mathbf{u}\mathbf{u}']\beta \neq \mathbf{0} \qquad (3)$$

as long as $\beta \neq 0$ and $\mathrm{E}[\mathbf{u}\mathbf{u}'] \neq \mathbf{0}$.

Measurement error in regressors

- We then have

$$\beta^* = (\mathrm{E}[\boldsymbol{X}\boldsymbol{X}'])^{-1}\mathrm{E}[\boldsymbol{X}Y] = \beta - (\mathrm{E}[\boldsymbol{X}\boldsymbol{X}'])^{-1}\mathrm{E}[\mathbf{u}\mathbf{u}']\beta \neq \beta. \tag{4}$$

- Our LS estimator

$$\widehat{\beta} = \left(n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_iY_i\right) \tag{5}$$

converges to β^*, not β. This is known as *measurement error bias*.

Supply and demand

- Suppose we are studying supply and demand using

$$\begin{aligned} Q &= -\beta_1 P + e_1, \quad &\text{(demand)} \\ Q &= \quad \beta_2 P + e_2, \quad &\text{(supply)} \end{aligned} \tag{6}$$

- Let $(e_1, e_2) = e$ be iid and $\mathrm{E}[e] = \mathbf{0}$, $\mathrm{E}[ee'] = \boldsymbol{I}_2$ for simplicity and $\beta_1 + \beta_2 = 1$. What happens if we regress Q on P?
- First, express the equation as

$$\begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix} \begin{pmatrix} Q \\ P \end{pmatrix} = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \tag{7}$$

- Solve for (Q, P)

$$\begin{aligned}\begin{pmatrix} Q \\ P \end{pmatrix} &= \begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix}^{-1} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \begin{bmatrix} \beta_2 & \beta_1 \\ 1 & -1 \end{bmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \\ &= \begin{pmatrix} \beta_2 e_1 + \beta_1 e_2 \\ e_1 - e_2 \end{pmatrix}\end{aligned} \tag{8}$$

- Projecting Q onto P yields

$$Q = \beta^* P + u,\ \mathrm{E}[Pu] = 0 \tag{9}$$

with

$$\beta^* = \frac{\mathrm{E}[PQ]}{\mathrm{E}[P^2]} = \frac{\beta_2 - \beta_1}{2} \tag{10}$$

A LS estimate of this projection thus converges to $\frac{\beta_2 - \beta_1}{2}$, not β_2 or β_1. This is known as *simultaneous equations bias*.

- Consider the linear regression, now a *structural equation*

$$Y = \boldsymbol{X}'\beta + e = \boldsymbol{X}_1'\beta_1 + \boldsymbol{X}_2'\beta_2 + e,\ \mathrm{E}[\boldsymbol{X}e] \neq 0 \tag{11}$$

  where we have endogeneity.

- Now, $\boldsymbol{X}_1, \beta_1$ are $k_1 \times 1$ and $\boldsymbol{X}_2, \beta_2$ are $k_2 \times 1$, and are partitioned such that

$$\mathrm{E}[\boldsymbol{X}_1 e] = \mathbf{0} \quad \text{but} \quad \mathrm{E}[\boldsymbol{X}_2 e] \neq \mathbf{0} \tag{12}$$

  I.e., $\boldsymbol{X}_2$ has non-zero covariance with e and is *endogenous* while $\boldsymbol{X}_1$ is *exogenous*.

- To get an unbiased/consistent estimate, we need to solve β from the sample analog of

$$\mathrm{E}[\boldsymbol{Z}e] = \mathrm{E}\left[\begin{pmatrix}\boldsymbol{X}_1 \\ \boldsymbol{Z}_2\end{pmatrix} e\right] = \mathbf{0} \tag{13}$$

  where $\boldsymbol{X}_1 = \boldsymbol{Z}_1$ are the $k_1 \times 1$ *included exogenous regressors* and $\boldsymbol{Z}_2$ are the $l_2 \times 1$ *excluded exogenous regressors*.

- If $l = k_1 + l_2 = k$ the model is *just-identified* (or exactly identified), if $l > k$ it is over-identified
- That you *need* $l_2 \geq k_2$ instruments to consistently estimate β does *not* mean they *exist*!

- Recall Slide 3, equation 3.
- Suppose there exist a set of l instruments ($l \geq k$), $\boldsymbol{Z}$, such that
  - it is uncorrelated with the regression error $\mathrm{E}[\boldsymbol{Z}e] = 0$
  - it is uncorrelated with the measurement error $\mathrm{E}[\mathbf{Zu}] = 0$
  - it is correlated with the true regressors $\mathrm{E}[\boldsymbol{Z}\boldsymbol{X}^{*\prime}] = \boldsymbol{Q}$ (where $\boldsymbol{Q} > 0$)
- An unbiased estimator of β can be constructed using the sample analog of the condition that

$$\mathrm{E}[\boldsymbol{Z}(Y - \boldsymbol{X}'\beta)] = 0. \tag{14}$$

- The key issue is whether or not such a $\boldsymbol{Z}$ exists.

Reduced Form

- In contrast to the structural form, we approach estimation of the parameters in the *reduced form* equations using linear projection.
- Consider the two sets of linear projection coefficients:

$$\begin{aligned}\boldsymbol{\Gamma} &= \mathrm{E}[\boldsymbol{Z}\boldsymbol{Z}']^{-1}\mathrm{E}[\boldsymbol{Z}\boldsymbol{X}'] \\ \boldsymbol{\lambda} &= \mathrm{E}[\boldsymbol{Z}\boldsymbol{Z}']^{-1}\mathrm{E}[\boldsymbol{Z}Y]\end{aligned} \tag{15}$$

  ($\boldsymbol{\Gamma}$ is $l \times k$, $\boldsymbol{\lambda}$ is $l \times 1$.)
- From the first, we have the $k \times 1$ projection error:

$$\mathbf{u} = \boldsymbol{X} - \boldsymbol{\Gamma}'\boldsymbol{Z} \Rightarrow \boldsymbol{X} = \boldsymbol{\Gamma}'\boldsymbol{Z} + \mathbf{u},\ \mathrm{E}[\mathbf{Z}\mathbf{u}'] = \mathbf{0}. \tag{16}$$

- Substitute 16 into 11 to get the *reduced form* for Y:

$$Y = \beta'(\boldsymbol{\Gamma}'\boldsymbol{Z} + \mathbf{u}) + e \Rightarrow Y = \boldsymbol{\lambda}'\boldsymbol{Z} + v. \tag{17}$$

- The reduced form and structural parameters are related:

$$\boldsymbol{\lambda} = \boldsymbol{\Gamma}\beta, \tag{18}$$

  as are the errors:

$$v = \beta'\mathbf{u} + e. \tag{19}$$

Reduced Form

- By construction, we have

$$E[\mathbf{Z}v] = E[\mathbf{Z}\mathbf{u}']\beta + E[\mathbf{Z}e] = 0. \tag{20}$$

- Using sample data, both sets of linear projection coefficients can be estimated by LS:

$$\begin{aligned}\widehat{\boldsymbol{\Gamma}} &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X} \\ \widehat{\boldsymbol{\lambda}} &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}\end{aligned} \tag{21}$$

Identification

- *Identification* is about this: can the structural parameter β be recovered from the reduced form coefficients?
- Solve β from eq. 18:

$$\beta = \mathbf{\Gamma}^{-1}\boldsymbol{\lambda},\ l = k; \quad \beta = (\mathbf{\Gamma}'\boldsymbol{W}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}'\boldsymbol{W}\boldsymbol{\lambda},\ l > k, \boldsymbol{W} > 0. \tag{22}$$

  For this to be possible, the *rank condition* must hold:

$$\text{rank}(\mathbf{\Gamma}) = k \tag{23}$$

- Note that $\boldsymbol{Z}' = (\boldsymbol{X}_1', \boldsymbol{Z}_2')$ and $\boldsymbol{X}' = (\boldsymbol{X}_1', \boldsymbol{X}_2')$ so

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{\Gamma}_{11} & \mathbf{\Gamma}_{12} \\ \mathbf{\Gamma}_{21} & \mathbf{\Gamma}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \mathbf{\Gamma}_{12} \\ \mathbf{0} & \mathbf{\Gamma}_{22} \end{bmatrix} \tag{24}$$

  and the regression of $\boldsymbol{X}$ on $\boldsymbol{Z}$ can be re-written as

$$\begin{aligned} \boldsymbol{X}_1 =& \mathbf{Z}_1 \quad (\text{Note:}\ \mathbf{Z}_1 \equiv \boldsymbol{X}_1) \\ \boldsymbol{X}_2 =& \mathbf{\Gamma}_{12}'\mathbf{Z}_1 + \mathbf{\Gamma}_{22}'\mathbf{Z}_2 + \mathbf{u} \end{aligned} \tag{25}$$

- The key to identification is the rank of $\mathbf{\Gamma}_{22}$ which must be k_2 to ensure $\text{rank}(\mathbf{\Gamma}) = k$.

Estimation

- We treat IV estimation and the distribution of the estimators in greater detail under GMM (chapter 12; lecture 13) but now cover the main estimators.

IV estimator

- With a just-identified model $l = k$, we have the moment condition

$$\mathrm{E}[\boldsymbol{Z}e] = \mathbf{0}. \tag{26}$$

- With $e = Y - \boldsymbol{X}'\beta$, we have

$$\begin{aligned}&\mathrm{E}[\boldsymbol{Z}(Y - \boldsymbol{X}'\beta)] = \mathbf{0} \Leftarrow \\ &\mathrm{E}[\boldsymbol{Z}Y] - \mathrm{E}[\boldsymbol{Z}\boldsymbol{X}']\beta = \mathbf{0},\end{aligned} \tag{27}$$

so

$$\beta = (\mathrm{E}[\boldsymbol{Z}\boldsymbol{X}'])^{-1}\mathrm{E}[\boldsymbol{Z}Y]. \tag{28}$$

- The Instrumental Variables (IV) estimator $\widehat{\boldsymbol{\beta}}_{IV}$ is the "plug-in" estimator

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{IV} =&\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_iY_i\right) \\ =&(\boldsymbol{Z}'\boldsymbol{X})^{-1}(\boldsymbol{Z}'\boldsymbol{Y}).\end{aligned} \tag{29}$$

- To generalize a bit, for any instruments $\boldsymbol{W}$, an IV estimator is

$$\widehat{\boldsymbol{\beta}}_{IV} = (\boldsymbol{W}'\boldsymbol{X})^{-1}(\boldsymbol{W}'\boldsymbol{Y}). \tag{30}$$

Two-stage least squares

- The IV estimator is simple to solve from the moment conditions as $l = k$ (just identified).
- With $l \geq k$, this will not work.
- The reduced form can be written as

$$Y = \boldsymbol{Z}'\boldsymbol{\Gamma}\beta + v, \quad \mathrm{E}[\boldsymbol{Z}v] = \boldsymbol{0}. \tag{31}$$

- Now let $\boldsymbol{w} = \boldsymbol{\Gamma}'\boldsymbol{Z}$ so

$$Y = \boldsymbol{w}'\beta + v, \quad \mathrm{E}[\boldsymbol{w}v] = \boldsymbol{0} \tag{32}$$

- If $\boldsymbol{\Gamma}$ was known, the "natural" estimator would be

$$\begin{aligned}\widehat{\beta} &= (\boldsymbol{W}'\boldsymbol{W})^{-1}(\boldsymbol{W}'\boldsymbol{Y}) \\ &= (\boldsymbol{\Gamma}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Gamma})^{-1}(\boldsymbol{\Gamma}'\boldsymbol{Z}'\boldsymbol{Y}).\end{aligned} \tag{33}$$

- Since this is not feasible, we must use $\widehat{\boldsymbol{\Gamma}}$ in its place:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{2SLS} &= (\widehat{\boldsymbol{\Gamma}}'\boldsymbol{Z}'\boldsymbol{Z}\widehat{\boldsymbol{\Gamma}})^{-1}(\widehat{\boldsymbol{\Gamma}}'\boldsymbol{Z}'\boldsymbol{Y}) \\ &= (\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \\ &= (\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}.\end{aligned} \tag{34}$$

- The parameters are estimated by least squares in two stages, hence the name.
- The properties of the 2SLS estimator will be examined in the next lecture.

Endogeneity tests

- The 2SLS estimator is needed as the regressors $\boldsymbol{X}_2$ in the structural equation are thought to be endogenous ($\mathrm{E}[\boldsymbol{X}_2 e] \neq 0$).
- This is a testable restriction, with hypotheses

$$\mathbb{H}_0 : \mathrm{E}[\boldsymbol{X}_2 e] = \mathbf{0} \quad \text{against} \quad \mathbb{H}_0 : \mathrm{E}[\boldsymbol{X}_2 e] \neq \mathbf{0} \tag{35}$$

- This can be approached using *control functions*.

Control functions

- We can write the structural and first-stage regressions as

$$
\begin{aligned}
Y &= \boldsymbol{X}_1'\beta_1 + \boldsymbol{X}_2'\beta_2 + e \\
\boldsymbol{X}_2 &= \boldsymbol{\Gamma}_{12}'\mathbf{Z}_1 + \boldsymbol{\Gamma}_{22}'\mathbf{Z}_2 + \mathbf{u}_2.
\end{aligned}
\tag{36}
$$

- The endogeneity of $\boldsymbol{X}_2$ means that $\mathbf{u}_2$ and e are correlated
- Consider then the linear projection of e onto $\mathbf{u}_2$:

$$
e = \mathbf{u}_2'\alpha + \epsilon; \quad \alpha = (\mathrm{E}[\mathbf{u}_2\mathbf{u}_2'])^{-1}\mathrm{E}[\mathbf{u}_2 e]; \quad \mathrm{E}[\mathbf{u}_2\epsilon] = \mathbf{0}. \tag{37}
$$

- Substituting this back into the structural equation we have

$$
\begin{aligned}
Y &= \boldsymbol{X}_1'\beta_1 + \boldsymbol{X}_2'\beta_2 + \mathbf{u}_2'\alpha + \epsilon \\
&\mathrm{E}[\boldsymbol{X}_1\epsilon] = 0; \quad \mathrm{E}[\boldsymbol{X}_2\epsilon] = 0; \quad \mathrm{E}[\mathbf{u}_2\epsilon] = 0
\end{aligned}
\tag{38}
$$

- Note that $\boldsymbol{X}_2$ is uncorrelated with ϵ. It's correlation with e is through $\mathbf{u}_2$ and ϵ is the error after e has been projected orthogonally onto $\mathbf{u}_2$.

- While we do not observe $\mathbf{u}_2$, it can be estimated by the first-stage residual

$$\hat{\mathbf{u}}_2 = \boldsymbol{X}_2 - \widehat{\boldsymbol{\Gamma}}'_{12}\boldsymbol{Z}_1 + \widehat{\boldsymbol{\Gamma}}'_{22}\boldsymbol{Z}_2 \tag{39}$$

- The coefficients $(\beta_1, \beta_2, \alpha)$ can be estimated by LS on

$$Y = \boldsymbol{X}'\beta + \hat{\mathbf{u}}_2\alpha + \hat{\epsilon}. \tag{40}$$

Endogeneity tests

- Since $\mathrm{E}[\boldsymbol{X}_2 e] = 0$ if and only if $\mathrm{E}[\mathbf{u}_2 e] = 0$, the above $\mathbb{H}_0$ can be restated as
$$\mathbb{H}_0 : \alpha = 0 \quad \text{against} \quad \mathbb{H}_1 : \alpha \neq 0.$$
- This can be tested using a Wald statistic; the problem is slightly more involved as the testing relies on the "generated regressor" $\widehat{\mathbf{u}_2}$ rather than the true values (see Hansen 2021, chs 12.26–12.27)
- An alternative approach is to use a so-called Hausman test (see Hansen 2021, ch 9.15)

Identification failure

- Recall from eq. 25

$$\boldsymbol{X}_2 = \boldsymbol{\Gamma}'_{12}\boldsymbol{Z}_1 + \boldsymbol{\Gamma}'_{22}\boldsymbol{Z}_2 + u \tag{41}$$

  that if $\boldsymbol{\Gamma}_{22}$ is not of rank k_2, identification fails. (The closely related problem of weak instruments – $\boldsymbol{\Gamma}_{22} \neq 0$ but is very small – will be studied in Econometrics II.)

- Suppose we study three scalar variables (Y, X, Z):

$$\begin{aligned} Y &= X\beta + e \\ X &= Z\gamma + u \end{aligned} \tag{42}$$

- Suppose $\gamma = 0$, so $\mathrm{E}[ZX] = 0$. What happens to the IV and LS estimators?
- Assume homoscedastic, unit variance and correlated errors (so X is endogenous),

$$\mathrm{Var}\left(\begin{pmatrix} e \\ u \end{pmatrix} \middle| Z\right) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad \mathrm{E}[Z^2] = 1. \tag{43}$$

Identification failure

- Study $(Ze, Zu)'$, for which we know by the CLT that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}Z_i e_i\\ Z_i u_i\end{pmatrix}\xrightarrow{d}\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix}\sim N\left(\mathbf{0},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right). \tag{44}$$

- Define $\xi_0 = \xi_1 - \rho\xi_2$, which is normal and $\perp \xi_2$.
- For the (O)LS estimator of β

$$\widehat{\beta}-\beta = \frac{n^{-1}\sum_{i=1}^{n}u_i e_i}{n^{-1}\sum_{i=1}^{n}u_i^2}\xrightarrow{p}\rho \neq 0. \tag{45}$$

  (I.e., the endogeneity of X renders LS inconsistent.)
- With identification failure, $\gamma = 0$, the asymptotic distribution of the IV estimator is

$$\widehat{\boldsymbol{\beta}}_{IV}-\beta = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i e_i}{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i X_i}\xrightarrow{d}\frac{\xi_1}{\xi_2} = \rho + \frac{\xi_0}{\xi_2}. \tag{46}$$

Identification failure

- $\widehat{\boldsymbol{\beta}}_{IV}$ is inconsistent: it converges to a random variable, not the correct constant (or indeed any constant)
- The ratio ξ_0/ξ_2 is symmetrically distributed around zero so $\widehat{\boldsymbol{\beta}}_{IV}$ has median at $\beta + \rho$
- ξ_0/ξ_2 is a ratio of two independent normal variables so follows a Cauchy distribution; it therefore does not have a finite expectation

Identification failure

- For the t-statistic, we need

$$
\begin{aligned}
\widehat{\sigma}^2 =& \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i\widehat{\boldsymbol{\beta}}_{IV}\right)^2 \\
=& \frac{1}{n}\sum_{i=1}^{n} e_i^2 - \frac{2}{n}\sum_{i=1}^{n} e_i X_i\left(\widehat{\boldsymbol{\beta}}_{IV} - \beta\right) + \frac{1}{n}\sum_{i=1}^{n} X_i^2\left(\widehat{\boldsymbol{\beta}}_{IV} - \beta\right)^2 \\
\xrightarrow{d}& 1 - 2\rho\frac{\xi_1}{\xi_2} + \left(\frac{\xi_1}{\xi_2}\right)^2 .
\end{aligned}
\tag{47}
$$

- Then we have the asymptotic distribution of the t-statistic,

$$
T = \frac{\widehat{\boldsymbol{\beta}}_{IV} - \beta}{\sqrt{\widehat{\sigma^2}\sum_{i=1}^{n} Z_i^2 / \sum_{i=1}^{n}|X_i Z_i|}} \xrightarrow{d} \frac{\xi_1}{\sqrt{1 - 2\rho\frac{\xi_1}{\xi_2} + \left(\frac{\xi_1}{\xi_2}\right)^2}}
\tag{48}
$$

- This is non-normal. E.g., if $\rho \to 1$, $\xi_1/\xi_2 \xrightarrow{p} 1$ and $\widehat{\sigma}^2 \xrightarrow{p} 0$. As a concsequence, the standard error of $\widehat{\boldsymbol{\beta}}_{IV}$ converges to zero and the t-statistic converges to ∞! so $T \to \infty$ as $\xi_1/\xi_2 \to 1$.

References

Hansen, Bruce E (2021). *Econometrics*. Madison, WI: University of Wisconsin.