DSE - Data-Driven Economic Analysis
Econometrics Module
Lecture 7 - Instrumental Variables

Michele De Nadai
michele.denadai@unimi.it

Trimester II, 2023

UNIVERSITÀ
DEGLI STUDI
DI MILANO

Endogeneity

- Suppose we are looking to estimate this linear model:

$$y = \beta_0 + \beta_1 d + \beta_3 x + u, \quad E[u|d, x] = 0$$

- Under Conditional Indipendence Assumption β_1 has a **causal interpretation**.
- If we do not have access to x we could aim to estimate a **simpler** model:

$$y = \beta_0 + \beta_1 d + u^*,$$

- Note that in this model the error term is **not mean independent** of d:

$$E[u^*|d] = E[\beta_3 x + u|d] = \beta_3 E[x|d] \neq 0$$

- When this happens we say that d is **endogenous**.

Endogeneity

- Note that we could potentially write down a linear model where the error term is mean independent of d.
- Assume $E[x|d] = a + bd$ then

$$\begin{aligned} E[y|d] &= \beta_0 + \beta_1 d + \beta_2(a + bd) + E[u|d] \\ &= \underbrace{(\beta_0 + a)}_{\gamma_0} + \underbrace{(\beta_1 + b)}_{\gamma_1} d \end{aligned}$$

- We could then write

$$y = \gamma_0 + \gamma_1 d + e, \quad E[e|d] = 0,$$

- OLS regression of y on d would **only consistently estimate** γ_0 and γ_1.

Motivation: Omitted Variables

- The problem arises when:
  - a relevant explanatory variable is omitted from the regression specification, **and**
  - this variable is correlated with the explanatory variables included in the specification.
- That is, the regression we estimate is **not** the regression we **would like** to estimate.
- Let us see it with an example

Motivation: Omitted Variables

- Returns to schooling:

$$\log \text{wage}_i = \alpha + \beta \text{educ}_i + \gamma x_i + e_i,$$

  where x_i is other observable characteristics.

- Ability or motivation are unobserved but possibly affect wages.
- Female labor supply:

$$\text{labor supply}_i = \alpha + \beta \text{family size}_i + \gamma x_i + e_i$$

- Mothers with weak labor force attachment or low earnings potential may be more likely to have children than mothers with strong labor force attachment or high earnings potential.

Omitted Variables: why it matters

- Let us consider $x_i = \text{ability}_i$
- Estimating β ignoring ability_i by OLS yields:

$$\begin{aligned}
\beta_{\text{OLS}} &= \frac{Cov(y, \text{educ})}{Var(\text{educ})}, \\
&= \frac{Cov(\alpha + \beta \cdot \text{educ} + \gamma \cdot \text{ability} + \epsilon, \text{educ})}{Var(\text{educ})}, \\
&= \beta \frac{Var(\text{educ})}{Var(\text{educ})} + \frac{Cov(\gamma \cdot \text{ability} + \epsilon, \text{educ})}{Var(\text{educ})}, \\
&= \beta + \gamma \frac{Cov(\text{ability}, \text{educ})}{Var(\text{educ})},
\end{aligned}$$

where $\frac{Cov(\text{ability},\text{educ})}{Var(\text{educ})}$ is the slope coefficient of a regression of ability on educ.

Example

- β_{OLS} does **not** identify the coefficient of interest β unless:
  - $\gamma = 0$; i.e. ability is not a relevant regressor in the structural equation,
  - $Cov(\text{ability}, \text{educ}) = 0$; i.e. ability and years of schooling are uncorrelated.
- Neither of the two is likely, as we might expect that persons with higher ability have higher wages, but are also more likely to invest in more years of schooling.

Measurement error

- There are two cases:
  - measurement error in the **dependent variable**,
  - measurement error in one (or more than one) **explanatory variable**.
- Only the latter gives rise to **endogeneity**
- Let us consider both cases separately.

Measurement error: Dependent Variable

- Structural model:

$$y_i^* = \mathbf{x}_i^{'}\boldsymbol{\beta} + e_i, \;\; \mathrm{E}[e_i|\mathbf{x}_i] = 0$$

where $\mathbf{y}_i^*$ is unobservable, $\mathbf{y}_i = \mathbf{y}_i^* + \mathbf{u}_i$ is observed, and $\mathbf{u}_i$ is a measurement error independent of $\mathbf{y}_i^*$ and x_i.

- Rewrite the model

$$\begin{aligned} y_i - \mathbf{u}_i &= \mathbf{x}_i^{'}\boldsymbol{\beta} + e_i \\ y_i &= \mathbf{x}_i'\boldsymbol{\beta} + e_i + \mathbf{u}_i \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \end{aligned}$$

where $\varepsilon_i = e_i + \mathbf{u}_i$ satisfies $\mathrm{E}[\mathbf{x}_i \varepsilon_i] = 0$.

- $\widehat{\beta}$ is still unbiased and consistent.

Measurement Error - Explanatory Variable

- Structural model:

$$y_i = \mathbf{x}_i^{*'}\boldsymbol{\beta} + e_i, \;\; \mathrm{E}[e_i|\mathbf{x}_i^*] = 0$$

  where $\mathbf{x}_i^*$ is unobservable, $\mathbf{x}_i = \mathbf{x}_i^* + \mathbf{u}_i$ is observed, and $\mathbf{u}_i$ is a measurement error independent of $\mathbf{x}_i^*$ and y_i.

- Rewrite the model

$$\begin{aligned} y_i &= \mathbf{x}_i^{*'}\boldsymbol{\beta} + e_i \\ &= (\mathbf{x}_i - \mathbf{u}_i)'\boldsymbol{\beta} + e_i \\ &= \mathbf{x}_i'\boldsymbol{\beta} + e_i - \mathbf{u}_i'\boldsymbol{\beta} \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \end{aligned}$$

  where $\varepsilon_i = e_i - \mathbf{u}_i'\boldsymbol{\beta}$ with $\mathrm{E}[\mathbf{x}_i\varepsilon_i] \neq 0$.

- $\widehat{\beta}$ is biased and inconsistent.

Measurement error: Explanatory Variable

- Indeed it is:

$$
\begin{aligned}
\mathrm{E}[\mathbf{x}_i \varepsilon_i] &= \mathrm{E}[\mathbf{x}_i \{e_i - (\mathbf{x}_i - \mathbf{x}_i^*)' \boldsymbol{\beta}\}] \\
&= \mathrm{E}[\mathbf{x}_i e_i] - \mathrm{E}[\mathbf{x}_i (\mathbf{x}_i - \mathbf{x}_i^*)' \boldsymbol{\beta}], \\
&= \left\{ -\mathrm{E}[\mathbf{x}_i \mathbf{x}_i'] + \mathrm{E}[\mathbf{x}_i \mathbf{x}_i^{*'}] \right\} \boldsymbol{\beta} \\
&= \mathrm{E}[\mathbf{x}_i \mathbf{x}_i'] \left\{ \mathrm{E}[\mathbf{x}_i \mathbf{x}_i']^{-1} \mathrm{E}[\mathbf{x}_i \mathbf{x}_i^{*'}] - 1 \right\} \boldsymbol{\beta}
\end{aligned}
$$

- This is only zero when:
  - $\mathrm{E}[\mathbf{x}_i \mathbf{x}_i']^{-1} \mathrm{E}[\mathbf{x}_i \mathbf{x}_i^{*'}] = 1$, i.e. there is no measurement error. Why?
  - $\boldsymbol{\beta} = 0$, i.e. $\mathbf{x}_i^*$ is not a relevant regressor.

Measurement error: Explanatory Variable

- More generally the OLS estimator for $\boldsymbol{\beta}$ consistently estimates:

$$
\begin{aligned}
\boldsymbol{\beta}_{OLS} &= \mathsf{E}[\mathbf{x}_i\mathbf{x}_i']^{-1}\mathsf{E}[\mathbf{x}_i y_i], \\
&= \mathsf{E}[\mathbf{x}_i\mathbf{x}_i']^{-1}\mathsf{E}[\mathbf{x}_i(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i)], \\
&= \mathsf{E}[\mathbf{x}_i\mathbf{x}_i']^{-1}\mathsf{E}[\mathbf{x}_i\mathbf{x}_i']\boldsymbol{\beta} + \mathsf{E}[\mathbf{x}_i\mathbf{x}_i']^{-1}\mathsf{E}[\mathbf{x}_i\varepsilon_i], \\
&= \boldsymbol{\beta} + \mathsf{E}[\mathbf{x}_i\mathbf{x}_i']^{-1}\mathsf{E}[\mathbf{x}_i\mathbf{x}_i']\left\{\mathsf{E}[\mathbf{x}_i\mathbf{x}_i']^{-1}\mathsf{E}[\mathbf{x}_i\mathbf{x}_i^{*'}] - 1\right\}\boldsymbol{\beta} \\
&= \boldsymbol{\beta} + (\mathsf{E}[\mathbf{x}_i\mathbf{x}_i']^{-1}\mathsf{E}[\mathbf{x}_i\mathbf{x}_i^{*'}] - 1)\boldsymbol{\beta} \\
&= \mathsf{E}[\mathbf{x}_i\mathbf{x}_i']^{-1}\mathsf{E}[\mathbf{x}_i\mathbf{x}_i^{*'}]\boldsymbol{\beta}
\end{aligned}
$$

- This is the well known **attenuation bias** due to mismeasured regressors, since elements of $\mathsf{E}[\mathbf{x}_i\mathbf{x}_i']^{-1}\mathsf{E}[\mathbf{x}_i\mathbf{x}_i^{*'}]$ are values between 0 and 1.
- The larger the variability of measurement error $(Var(\mathbf{u}_i))$ the larger the bias.

Definition

- A linear model with endogeneity:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i, \quad \mathrm{E}[\mathbf{x}_i e_i] \neq 0$$

- $\mathbf{x}_i$ is called *endogeneous*.
- OLS inconsistent.
- $\mathrm{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}[e|X]$.

Instrumental Variables

- Consider the simple model:

$$y_i = x_i\beta + e_i, \quad \mathrm{E}[x_i e_i] \neq 0$$

- Assume availability of an **instrumental variable** (z_i) such that:
  - *Relevance* - $\mathrm{E}[x_i z_i] \neq 0$,
  - *Validity* - $\mathrm{E}[z_i e_i] = 0$.
- We could then write the linear projection of x_i on z_i as

$$x_i = \delta_0 + \delta_1 z_i + v_i, \quad \mathrm{E}[v_i z_i] = 0$$

Instrumental Variables (cont.)

- Which implies:

$$\begin{aligned} y_i &= \beta(\delta_0 + \delta_1 z_i + v_i) + e_i, \\ &= \beta\delta_0 + \beta\delta_1 z_i + \underbrace{\beta v_i + e_i}_{e_i^*} \end{aligned}$$

- Note that $E[z_i e_i^*] = \beta E[z_i v_i] + E[z_i e_i] = 0$.
- Now:
  - Slope of OLS regression of y on z, $\widehat{\beta}_{y,z} \xrightarrow{p} \beta\delta_1$,
  - Slope of OLS regression of x on z, $\widehat{\beta}_{x,z} \xrightarrow{p} \delta_1$,
  - The ratio $\frac{\widehat{\beta}_{y,z}}{\widehat{\beta}_{x,z}} \xrightarrow{p} \beta$.

Instrumental Variables - Examples

- Returns to schooling: Use college proximity as an instrument
  - Students living close to a college are more likely to enrol into college *ceteris paribus*.
- Female labor supply: Use twins at second birth and sex composition in the first two children as instruments
  - Occurence of twins is (assumed to be) random.
  - American parents with two boys or two girls are more likely to have a third child (sibling sex composition is random).