DSE - Data-Driven Economic Analysis
Econometrics Module
Lecture 8 - Instrumental Variables II

Michele De Nadai
michele.denadai@unimi.it

Trimester II, 2023

UNIVERSITÀ
DEGLI STUDI
DI MILANO

More than one Endogenous Variable

- General model

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

- Let k be the number of endogenous regressors ($\mathbf{x}_i$).
- Instrumental variables: An $l \times 1$ vector $\mathbf{z}_i$ that satisfies
  1. Validity: $\mathrm{E}[\mathbf{z}_i e_i] = 0$
  2. Relevance: $\mathrm{E}[\mathbf{z}_i \mathbf{x}_i'] \neq 0$
- Then we should have $l \geqslant k$ to **identify** $\boldsymbol{\beta}$.

Instrumental Variables (cont.)

- In words, there should be at least **one IV for each endogenous regressor**.
- In general we could have a model with k_1 exogenous and k_2 endogenous regressors ($k_1 + k_2 = k$). When l_2 instruments are available, then we can use the $l = k_1 + l_2$ instruments to estimate $\boldsymbol{\beta}$ ($k \times 1$).
- $l = k$ or $l_2 = k_2$: just-identified
- $l > k$ or $l_2 > k_2$: overidentified

First Stage

- So called "first stage" regression describes the relationship between **each endogenous** $\mathbf{x_i}$ and the set of instruments $\mathbf{z_i}$:

$$\mathbf{x_i} = \mathbf{\Gamma}'\mathbf{z_i} + \mathbf{u_i}, \;\; \mathrm{E}[\mathbf{z_i}\mathbf{u_i}'] = 0. \tag{1}$$

- Recall: $\mathbf{z_i}$ includes exogenous regressors.
- First stage in matrix form:

$$\mathbf{X} = \mathbf{Z}\mathbf{\Gamma} + \mathbf{U}$$

- OLS **estimate** of Γ:

$$\widehat{\mathbf{\Gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \xrightarrow{p} \mathbf{\Gamma}$$

Reduced Form

- Relationship between the **outcome** and the set of instruments, obtained as:

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = (\mathbf{Z}\boldsymbol{\Gamma} + \mathbf{U})\boldsymbol{\beta} + \mathbf{e} \\ &= \mathbf{Z}\boldsymbol{\lambda} + \mathbf{v}, \end{aligned} \tag{2}$$

  where $\boldsymbol{\lambda} = \boldsymbol{\Gamma}\boldsymbol{\beta}$ and $\mathbf{v} = \mathbf{U}\boldsymbol{\beta} + \mathbf{e}$.

- This model satisfies also $E[\mathbf{z}_i v_i] = 0$ where v_i is the i-th element of $\mathbf{v}$.

- OLS **estimate** of $\boldsymbol{\lambda}$ is:

$$\widehat{\boldsymbol{\lambda}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \xrightarrow{p} \boldsymbol{\lambda} = \boldsymbol{\Gamma}\boldsymbol{\beta}$$

- Equations (2) is called the **reduced form** equation.

Identification

- Recall $\boldsymbol{\lambda} = \boldsymbol{\Gamma\beta}$. For $\boldsymbol{\beta}$ to be identified (to be recovered from $\boldsymbol{\Gamma}$ and $\boldsymbol{\lambda}$), a necessary condition is

$$\mathrm{rank}(\boldsymbol{\Gamma}) = k.$$

- When $l = k$, $\boldsymbol{\beta} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\lambda}$.
- When $l > k$, for any $l \times l$ matrix $\mathbf{W} > 0$, $\boldsymbol{\beta} = (\boldsymbol{\Gamma}'\mathbf{W}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\mathbf{W}\boldsymbol{\lambda}$.
- This is the least square estimate of the regression of $\boldsymbol{\lambda}$ on $\boldsymbol{\Gamma}$ with no error.

Estimation

- Assume that $\boldsymbol{\beta}$ is identified.
- When $l = k$ (just-identified), the instrumental variables (IV) estimator is

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{IV} &= \widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\lambda}} \\
&= \left((\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \\
&= \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{y} \\
&\xrightarrow{p} \boldsymbol{\Gamma}^{-1}\boldsymbol{\lambda} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}\boldsymbol{\beta} = \boldsymbol{\beta}
\end{aligned}$$

Estimation (cont.)

- When $l > k$ (overidentified), the two-stage least squares (2SLS) estimator is

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

- In general we have:

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\widehat{\boldsymbol{\Gamma}}'\boldsymbol{W}\widehat{\boldsymbol{\Gamma}})^{-1}\widehat{\boldsymbol{\Gamma}}'\boldsymbol{W}\widehat{\boldsymbol{\lambda}} \\ &= \left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{W}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{W}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}\end{aligned}$$

- When $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$ we have $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{2SLS}$
- Where does it come from?

Why 2SLS ?

- First stage: Get fitted values of $\mathbf{X}$

$$\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\boldsymbol{\Gamma}}, \ \ \widehat{\boldsymbol{\Gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

- Second stage: Regress $\mathbf{y}$ on $\widehat{\mathbf{X}}$ (same as projecting $\mathbf{y}$ on $\widehat{\mathbf{X}}$)

$$\widehat{\boldsymbol{\beta}} = (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'\mathbf{y} =$$

Control Function Approach

- The structural equation and reduced form:

$$
\begin{aligned}
y_i &= \mathbf{x}_i'\boldsymbol{\beta} + e_i, \\
\mathbf{x}_i &= \boldsymbol{\Gamma}'\mathbf{z}_i + \mathbf{u}_i
\end{aligned}
$$

- IV assumption: $\mathrm{E}[\mathbf{z}_i e_i] = 0$
- Implication: $\mathbf{x}_i$ is endogenous iff $\mathbf{u}_i$ and e_i are correlated.
- Linear projection of e_i on $\mathbf{u}_i$:

$$
e_i = \mathbf{u}_i'\boldsymbol{\gamma} + \varepsilon_i, \quad \mathrm{E}[\mathbf{u}_i \varepsilon_i] = 0
$$

- Substitute this into the structural equation.

Control Function Approach (cont.)

- We have

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{u}_i'\boldsymbol{\gamma} + \varepsilon_i, \quad (3) \\ E[\mathbf{x}_i \varepsilon_i] &= 0, \\ E[\mathbf{u}_i \varepsilon_i] &= 0. \end{aligned}$$

- $\mathbf{x}_i$ is uncorrelated with ε_i. Why?
- Since $\mathbf{u}_i$ is not observable, we use $\widehat{\mathbf{u}}_i = \mathbf{x}_i - \widehat{\boldsymbol{\Gamma}}'\mathbf{z}_i$.
- Estimate $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in (3) by least-squares of y_i on $(\mathbf{x}_i, \widehat{\mathbf{u}}_i)$.
- The resulting estimator $\widehat{\boldsymbol{\beta}}$ is equivalent to $\widehat{\boldsymbol{\beta}}_{2sls}$.
- When the structural model is non-linear, the control function estimator would be different from the 2SLS.

Asymptotic Results

- Consistency

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{2SLS} &= \left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{X}\boldsymbol{\beta}+\mathbf{e}) \\
&= \boldsymbol{\beta} + \left(\frac{1}{n}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{e} \\
&\xrightarrow{p} \boldsymbol{\beta} + (Q_{zx}'Q_{zz}^{-1}Q_{zx})^{-1}Q_{zx}'Q_{zz}^{-1}\mathrm{E}[\mathbf{z}_i e_i] \\
&= \boldsymbol{\beta},
\end{aligned}$$

  where $Q_{zx} = \mathrm{E}[\mathbf{z}_i\mathbf{x}_i']$, $Q_{zz} = \mathrm{E}[\mathbf{z}_i\mathbf{z}_i']$.

- Unfortunately in general $\mathrm{E}[y_i|\mathbf{X},\mathbf{Z}] \neq \mathbf{X}\boldsymbol{\beta}$, which implies that $\mathrm{E}[\widehat{\boldsymbol{\beta}}_{2SLS}] \neq \boldsymbol{\beta}$.

Asymptotic Results

- Asymptotic normality:

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}_{2SLS}-\boldsymbol{\beta}) &= \left(\frac{1}{n}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\cancel{\frac{1}{n}}\left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\right)\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{e} \\ &\xrightarrow{d} (Q_{zx}'Q_{zz}^{-1}Q_{zx})^{-1}Q_{zx}'Q_{zz}^{-1}N\left(0, E[\mathbf{z}_i\mathbf{z}_i'e_i^2]\right) \\ &= N(0,\Sigma),\end{aligned}$$

  where
  $\Sigma = (Q_{zx}'Q_{zz}^{-1}Q_{zx})^{-1}Q_{zx}'Q_{zz}^{-1}\Omega Q_{zz}^{-1}Q_{zx}(Q_{zx}'Q_{zz}^{-1}Q_{zx})^{-1}$
  and $\Omega = E[\mathbf{z}_i\mathbf{z}_i'e_i^2]$.

- When errors are homosckedastic we have $E[\mathbf{z}_i\mathbf{z}_i'e_i] = \sigma^2 Q_{zz}$, which implies $\Sigma = \sigma^2(Q_{zx}'Q_{zz}^{-1}Q_{zx})^{-1}$.

- Note: It is incorrect to calculate the variance (or standard error) of the second stage OLS estimator.

References

- "Econometrics", B. Hansen (2022) **Chapter 12.1-12.12 and 12-15-12.16**