

Linear Models Lecture 16: IV

Robert Gulotty

University of Chicago

May 14, 2025

2SLS and IV

- `iv_robust(Y ~ D + X|Z + X, data = dat)`

- IV formula:

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'y$$

- Two stage least squares:

- Suppose in the first stage we regress

$$X = Z\gamma + v$$

- In the second stage, we use $\hat{X} = Z\hat{\gamma} = Z(Z'Z)^{-1}Z'X = P_ZX$,

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$$

Equivalence Between 2SLS and IV

- 2SLS is exactly identical to IV when $l = k$

$$\begin{aligned}\hat{\beta}_{2SLS} &= (\hat{X}'\hat{X})^{-1}\hat{X}'y \\ &= (X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'y && ((ABC)^{-1} = C^{-1}B^{-1}A^{-1}) \\ &= (Z'X)^{-1}(Z'Z)(Z'Z)^{-1}Z'y \\ &= (Z'X)^{-1}Z'y = \hat{\beta}_{IV}\end{aligned}$$

Control Function Regression

- Assume that \mathbf{x}_2 is endogenous:

$$Y = \mathbf{x}_1' \beta_1 + \mathbf{x}_2' \beta_2 + e$$

$$\mathbf{x}_2 = \Gamma'_{12} \mathbf{z}_1 + \Gamma'_{22} \mathbf{z}_2 + \mathbf{u}_2$$

- The control function approach directly models the error:

$$e = \mathbf{u}_2' \alpha + v$$

$$\alpha = (E[\mathbf{u}_2 \mathbf{u}_2'])^{-1} E[\mathbf{u}_2 e]$$

$$E[\mathbf{u}_2 v] = 0$$

Control Function Regression

- We then plug this in to the original structural form equation, controlling for the error.

$$Y = \mathbf{x}'_1\beta_1 + \mathbf{x}'_2\beta_2 + e$$

$$Y = \mathbf{x}'_1\beta_1 + \mathbf{x}'_2\beta_2 + u'_2\alpha + v$$

$$E[X_1v] = 0$$

$$E[X_2v] = 0$$

$$E[u_2v] = 0$$

- After we control for u_2 , the error is uncorrelated with \mathbf{x}
- We estimate this new control with the reduced form residual:

$$\hat{u}_{2i} = x_{2i} - \hat{\Gamma}'_{12}\mathbf{z}_1 + \hat{\Gamma}'_{22}\mathbf{z}_2$$

- It is like subtracting off the endogenous part.

$$\mathbf{Y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{U}}_e\hat{\alpha} + \hat{v}$$

Application: Heterogenous Returns to Education

- Consider the canonical returns to education model:

$$\textit{lwage}_i = \mathbf{z}_{i1}\boldsymbol{\delta}_1 + g_{i1}\textit{educ}_i + u_{i1}$$

- The returns to schooling for the population is $\gamma_i = E[g_{i1}]$

$$g_{i1} = \gamma_1 + v_{i1}$$

- Plugging in:

$$\textit{lwage}_i = \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \gamma_1\textit{educ}_i + v_{i1}\textit{educ}_i + u_{i1}$$

$$\textit{educ}_i = \mathbf{z}_i\boldsymbol{\pi}_2 + v_{i2}$$

Control function approach assumes that unobservables are linearly related to v_{i2}

- We then proceed estimation by controlling for \hat{v}_{i2} and the interaction between $\hat{\textit{educ}}_i$ and the estimated \hat{v}_{i2} .

Challenges with IV

- The IV estimator is among the most common tools of econometrics.
- However, it has several weaknesses.
 - Imprecision
 - Small sample Bias
 - Sensitivity to Weak Instruments

Problems with IV estimator: Imprecision

- Suppose Z and X are mean 0, $y = X\beta + e$,

$$Z'X = X'Z = \sum z_i x_i = n * cov(z, x)$$

$$Z'Z = \sum z_i^2 = n * var(z)$$

$$X'X = \sum x_i^2 = n * var(x)$$

$$\hat{\beta}_{IV} = (Z'X)^{-1} Z'y$$

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'y$$

$$Avar(\hat{\beta}_{OLS}) = \sigma_e^2 (X'X)^{-1}$$

$$Avar(\hat{\beta}_{IV}) = \sigma_e^2 (Z'X)^{-1} Z'Z (X'Z)^{-1}$$

Problems with IV estimator: Imprecision

$$Avar(\hat{\beta}_{OLS}) = \sigma_e^2 (X'X)^{-1} = \frac{\sigma_e^2}{n} \frac{1}{\text{var}(x)}$$

$$\begin{aligned} Avar(\hat{\beta}_{IV}) &= \sigma_e^2 (Z'X)^{-1} Z'Z (X'Z)^{-1} = \frac{\sigma_e^2}{n^2} \frac{n * \text{var}(z)}{\text{cov}(x, z)^2} \\ &= \frac{\sigma_e^2}{n} \frac{1}{\text{var}(x)} \frac{\text{var}(x)\text{var}(z)}{\text{cov}(x, z)^2} \\ &= \frac{\sigma_e^2}{n} \frac{1}{\text{var}(x)} \frac{1}{\rho_{xz}^2} \\ &= Avar(\hat{\beta}_{OLS}) \frac{1}{\rho_{xz}^2} \end{aligned}$$

■ As $\rho_{xz}^2 \rightarrow 0$, $Avar(\hat{\beta}_{IV}) \rightarrow \infty$

Problems with IV estimator: Bias

- IV is often is neither biased nor unbiased because it does not even have an expectation.
- Kiviet has shown that the IV estimator has M moments, the number of overidentifying restrictions. If $q = 0$, IV has no expectation.

$$y = X\beta + e$$

$$X = Z\pi + v$$

$$\begin{aligned}\hat{\beta}_{IV} &= (X'P_ZX)^{-1}X'P_Zy \\ &= \beta + (X'P_ZX)^{-1}X'P_Ze \\ &= \beta + (X'P_ZX)^{-1}(\pi'Z' + v')P_Ze \\ &= \beta + (X'P_ZX)^{-1}(\pi'Z' + v')P_Ze \\ &= \beta + (X'P_ZX)^{-1}\pi'Z'P_Ze + (X'P_ZX)^{-1}v'P_Ze \\ &= \beta + (X'P_ZX)^{-1}\pi'Z'e + (X'P_ZX)^{-1}v'P_Ze\end{aligned}$$

Form of small sample bias

$$\begin{aligned} E(\hat{\beta}_{IV}) - \beta &\approx E(X'P_ZX)^{-1}E(\pi'Z'e) + E(X'P_ZX)^{-1}E(v'P_Ze) \\ &= E(X'P_ZX)^{-1}\pi'E(Z'e) + E(X'P_ZX)^{-1}E(v'P_Ze) \\ &= (E(X'P_ZX))^{-1}E(v'P_Ze) \\ &= (E(\pi'Z' + v')P_Z(Z\pi + v)))^{-1}E(v'P_Ze) \\ &= (E(\pi'Z'Z\pi + \pi'Z'v + v'Z\pi + v'P_Zv))^{-1}E(v'P_Ze) \\ &= (E(\pi'Z'Z\pi) + E(v'P_Zv))^{-1}E(v'P_Ze) \quad (\text{b/c } E(Z'e) = E(Z'v) = 0) \\ &= (E(\pi'Z'Z\pi) + \sigma_v^2\rho)^{-1}\sigma_{ev}^2\rho \\ &= \frac{1}{\left(\frac{E(\pi'Z'Z\pi)/\rho}{\sigma_v^2} + 1\right)} \frac{\sigma_{ev}^2}{\sigma_v^2} \end{aligned}$$

F-test

$$E(\hat{\beta}_{IV}) - \beta \approx \frac{1}{\left(\frac{E(\pi'Z'Z\pi)/p}{\sigma_v^2} + 1\right)} \frac{\sigma_{ev}^2}{\sigma_v^2} \approx \frac{1}{(1 + F_{p,n-p})} \frac{\sigma_{ev}^2}{\sigma_v^2}$$

- F is the test where the null is that all instrument coefficients are 0.
- The bias of IV only goes away if $F \rightarrow \infty$
- The bias of IV is the OLS bias as $F \rightarrow 0$.
- Adding useless instruments increases p, which decreases F and increases the bias.

Weak instruments

Suppose we have a single x and a single instrument z . An instrument is weak if ρ_{zx} is small.

$$\begin{aligned} \text{plim} \hat{\beta}_{OLS} &= \text{plim} \frac{\text{cov}(x, y)}{\text{var}(x)} \\ &= \text{plim} \frac{\text{cov}(x, \alpha + \beta + e)}{\text{var}(x)} \\ &= \beta + \text{plim} \frac{\text{cov}(x, e)}{\text{var}(x)} \\ &= \beta + \text{plim} \frac{\text{cov}(x, e)}{\sqrt{\text{var}(x)} \sqrt{\text{var}(e)}} \frac{\sqrt{\text{var}(e)}}{\sqrt{\text{var}(x)}} \\ &= \beta + \rho_{xe} \frac{\sigma_e}{\sigma_x} \end{aligned}$$

Weak instruments

$$\begin{aligned} \text{plim} \hat{\beta}_{IV} &= \beta + \text{plim} \frac{\text{cov}(z, e)}{\text{cov}(z, x)} \\ &= \beta + \text{plim} \frac{\frac{\text{cov}(z, e)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(e)}}}{\frac{\text{cov}(z, x)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(z)}}} \frac{\sqrt{\text{var}(e)}}{\sqrt{\text{var}(x)}} \\ &= \beta + \frac{\rho_{ze} \sigma_e}{\rho_{zx} \sigma_x} \\ &= \beta + \frac{\rho_{ze}}{\rho_{zx} \rho_{xe}} \rho_{xe} \frac{\sigma_e}{\sigma_x} \\ &= \beta + \frac{\rho_{ze}}{\rho_{zx} \rho_{xe}} \text{ABias}(\hat{\beta}_{OLS}) \end{aligned}$$

Weak/Bad instruments are worse than OLS

$$\frac{ABias(\hat{\beta}_{OLS})}{ABias(\hat{\beta}_{IV})} > 1 \rightarrow \frac{\rho_{ze}}{\rho_{zx}\rho_{xe}} > 1$$

If $\frac{\rho_{ze}}{\rho_{zx}\rho_{xe}} \geq 1$, then IV is more biased than OLS.

Suppose $\rho_{xu} = .5$, so X is super endogenous, Z is barely endogenous: $\rho_{zu} = 0.01$.

Small $\rho_{zx} = 0.019$ gives $\frac{ABias(\hat{\beta}_{OLS})}{ABias(\hat{\beta}_{IV})} = 1.052$.

Testing power of instruments

$$\frac{ABias(\hat{\beta}_{OLS})}{ABias(\hat{\beta}_{IV})} \approx \frac{1}{F}$$

F statistic of 100 means IV is 1% as biased as OLS.

Testing endogeneity via Durbin-Hausman-Wu test

- If X is exogenous, then both OLS and IV are consistent, but OLS is BLUE.
- Asymptotically, the difference between OLS and IV should converge to zero.

$$H = (\hat{\beta}_{IV} - \hat{\beta}_{OLS})' [Avar(\hat{\beta}_{IV}) - Avar(\hat{\beta}_{OLS})]^{-1} (\hat{\beta}_{IV} - \hat{\beta}_{OLS}) \sim \chi^2_{dim(\beta)}$$

- Rejecting null says that OLS and IV are not close to one another, so either X is endogenous or Z is an invalid instrument.