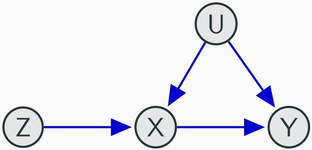
Instrumental Variables: a Short Introduction

Richard Guo

Nov 9, 2020

Department of Statistics, University of Washington, Seattle

Instrumental variable



Instrumental variable (IV) is used in presence of **latent confounder** U.

☞ Z is called an instrumental variable if

1. **Exclusion restriction**: Z has no effect on Y except through X.
   - In terms of potential outcome notation, $Y(x, z) \equiv Y(x)$.
2. **Exogeneity**: Z and U are independent.
   - Weak exogeneity: $Y(x) \perp\!\!\!\perp Z$ for each level of treatment X.
3. **Relevance**: Z and X are not independent (faithfulness).

Validity

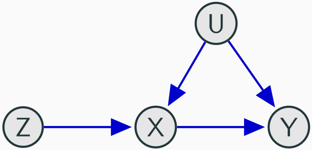
1. **Exclusion restriction**: Z has no effect on Y except through X.
2. **Exogeneity**: Z and U are independent.
3. **Relevance**: Z and X are not independent (faithfulness).

☞ Only **relevance** is **verifiable** (by rejecting the null $Z \perp\!\!\!\perp X$).

☞ Falsification

- Requirements 1 and 2 may imply conditions that can be tested with data (falsification/specification tests).
- But passing these tests does **not** prove that Z is a valid instrument.
- One has to **argue** that Z satisfies these requirements.

Examples



- Encouragement design:
  ▶ X: vaccine, Y: risk of flu, Z: random encouragement from doctor to get vaccine
- Genetic variation (Mendelian randomization):
  ▶ X: alcohol consumption, Y: heart disease, Z: polymorphism related to alcoholic metabolism
- Environmental factor
  ▶ X: economic condition, Y: civil conflict, Z: rainfall (Miguel, Satyanath, and Sergenti, 2004)

Non-identification

That being said, without making **further assumption**, the counterfactual distribution is **not identified** from the observed data.

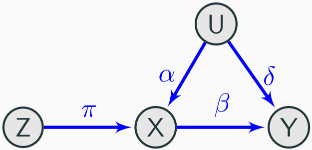
☞ ACE $\mathbb{E}[Y|\text{do}(X=1)] - \mathbb{E}[Y|\text{do}(X=0)]$ is unidentified.

☞ To proceed, we have to make additional assumptions.

► To relate to problems in geometry, I will focus on the **linear structural equation model** (linear SEM).

Linear SEM

Linear SEM is widely adopted in econometrics.



Let Z, X, U, Y be univariate random variables.

Suppose all variables have zero mean and finite variance.

$$Y = \beta X + \delta U + \eta^Y, \quad X = \pi Z + \alpha U + \eta^X.$$

- **Relevance**: $\pi \neq 0$
- **Exogeneity**: $Z \perp\!\!\!\perp U, \eta^X, \eta^Y$.

IV is a division

$$Y = \beta X + \delta U + \eta^Y, \quad X = \pi Z + \alpha U + \eta^X.$$

▶ By substituting the equation on X into the equation on Y, we get the "reduced form"

$$Y = \beta\pi Z + \beta(\alpha U + \eta^X) + \eta^Y,$$
$$X = \pi Z + \alpha U + \eta^X.$$

Now note $\beta(\alpha U + \eta^X) + \eta^Y \perp\!\!\!\perp Z$ and $\alpha U + \eta^X \perp\!\!\!\perp Z$ by exogeneity.

☞ We can consistently estimate $\beta\pi$ and π, and divide:

$$\hat{\beta} = \frac{\widehat{\beta\pi}}{\hat{\pi}} = \frac{\operatorname{cov}(Y, Z)/\operatorname{var} Z}{\operatorname{cov}(X, Z)/\operatorname{var} Z} = \frac{\operatorname{cov}(Y, Z)}{\operatorname{cov}(X, Z)},$$

which can be restated as two-stage least squares (2SLS)

- 1st stage: $\hat{\pi}$ from $X \sim Z$
- 2nd stage: $\hat{\beta}$ from $Y \sim \hat{\pi} Z$, where $\hat{\pi} Z$ is the **fitted value** of X.

IV is a division

☞ Under $\pi \neq 0$ (relevance), it is easy to see that $\hat{\beta}$ is consistent and asymptotically normal.

☞ When Z is binary, this is **Wald's estimator**

$$\hat{\beta} = \frac{\text{cov}(Y,Z)}{\text{cov}(X,Z)} = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[X|Z=1] - \mathbb{E}[X|Z=0]},$$

which follows from

$$cov(Y,Z) = (\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0])p(1-p)$$

for $p = \mathbb{E}\, Z$.

☞ So far we have studied the simplest case $\dim Z = \dim X = 1$.

To gain more insights, let us now assume $\dim Z = \dim X = k$.

☞ Absorbing endogenous errors, the equation on $Y \in \mathbb{R}$ can be written as

$$Y = \beta^\intercal X + \varepsilon^Y, \quad \beta, X \in \mathbb{R}^k,$$

where $\sigma^2 := \mathbb{E}(\epsilon^Y)^2$.

► Given that $Z \in \mathbb{R}^k$ is exogenous, we have $\mathbb{E}[\varepsilon^Y | Z] = \mathbf{0}$, which yields the estimating equation

$$\boldsymbol{Z}^T(\boldsymbol{y} - \boldsymbol{X}\beta) = \mathbf{0},$$

where $\boldsymbol{Z}, \boldsymbol{X} \in \mathbb{R}^{n\times k}, \boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^k$.

This yields the estimator

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\intercal \boldsymbol{X})^{-1} \boldsymbol{Z}^\intercal \boldsymbol{y},$$

which is consistent if (1) $\mathbb{E}[ZX^\intercal]$ is full-rank (2) $\boldsymbol{Z}^\intercal \boldsymbol{\varepsilon}^{\boldsymbol{y}}/n \to_p \mathbf{0}$.

Perspective from estimating equation

☞ CLT

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{d}{\Rightarrow} \mathcal{N}(\mathbf{0}, \Sigma),$$

with a sandwich asymptotic covariance

$$\Sigma = \sigma^2 \, \mathbb{E}[ZX^\intercal]^{-1} \, \mathrm{cov}(Z) \, \mathbb{E}[ZX^\intercal]^{-\intercal} = \sigma^2 \, \mathrm{plim}_n \left(n^{-1} \boldsymbol{X}^\intercal \boldsymbol{P_Z} \boldsymbol{X} \right),$$

where $\boldsymbol{P_Z} = \boldsymbol{Z}(\boldsymbol{Z}^\intercal \boldsymbol{Z})^{-1} \boldsymbol{Z}^\intercal$ is the **projection matrix** into the column space of $\boldsymbol{Z}$.

▶ But there could be other ways of forming the estimating equation, e.g., $f(\boldsymbol{Z})^\intercal(\boldsymbol{y} - \boldsymbol{X}\beta) = \mathbf{0}$ for some function f.

▶ The optimal choice should minimize Σ.

☞ Take $f(Z) = \bar{X} := \mathbb{E}[X|Z]$ is the optimal choice. That is, the "instrumented" treatment X.

To see that $\bar{X} = \mathbb{E}[X|Z]$ is the **asymptotically optimal** exogenous variable in estimating equation, consider the **asymptotic precision**

$$
\begin{aligned}
(\Sigma/\sigma^2)^{-1} &= \text{plim}_n\, n^{-1} \boldsymbol{X}^\intercal \boldsymbol{P_Z} \boldsymbol{X} \\
&= \mathbb{E}[XZ^\intercal] \,\text{cov}(Z)^{-1}\, \mathbb{E}[XZ^\intercal]^{-1} \\
&= \mathbb{E}[\bar{X}Z^\intercal] \,\text{cov}(Z)^{-1}\, \mathbb{E}[\bar{X}Z^\intercal]^{-1} \quad \text{(tower)} \\
&= \text{plim}_n\, n^{-1} \bar{\boldsymbol{X}}^\intercal \boldsymbol{P_Z} \bar{\boldsymbol{X}}.
\end{aligned}
$$

- Specializing to $Z = \bar{\boldsymbol{X}}$, the above becomes $\text{plim}_n\, n^{-1} \bar{\boldsymbol{X}}^\intercal \bar{\boldsymbol{X}}$.

☞ This is optimal because

$$
n^{-1} \bar{\boldsymbol{X}}^\intercal \bar{\boldsymbol{X}} - n^{-1} \bar{\boldsymbol{X}}^\intercal \boldsymbol{P_Z} \bar{\boldsymbol{X}} = n^{-1} \bar{\boldsymbol{X}}^\intercal (\boldsymbol{I} - \boldsymbol{P_Z}) \bar{\boldsymbol{X}} \succeq \boldsymbol{0}.
$$

☞ To estimating k coefficients, one needs at least k estimating equations. But one could have $l \geq k$ instruments.

Over-identification

Now suppose $\dim Z = l \geq k = \dim X$.

- Exactly-identified: $l = k$, Over-identified: $l > k$, Unidentified: $l < k$.

For $\boldsymbol{Z} \in \mathbb{R}^{n \times l}$, $\boldsymbol{J} \in \mathbb{R}^{l \times k}$, suppose we form estimating equation

$$(\boldsymbol{Z}\boldsymbol{J})^\intercal(\boldsymbol{y} - \boldsymbol{X}\beta) = \boldsymbol{0},$$

and look for the optimal $\boldsymbol{J}$ that minimizes the asymptotic covariance.

- Similar to the previous, the asymptotic precision

$$(\Sigma/\sigma^2)^{-1} = \operatorname{plim}_n n^{-1}\bar{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Z}\boldsymbol{J}}\bar{\boldsymbol{X}}.$$

In general, however, we cannot find $\boldsymbol{J}$ such that $\mathbb{E}[X|Z] = Z\boldsymbol{J}$.

☞ Nevertheless, the natural choice is

$$\boldsymbol{Z}\boldsymbol{J} = P_{\boldsymbol{Z}}\bar{\boldsymbol{X}} \quad \Rightarrow \quad \boldsymbol{J} = (\boldsymbol{Z}^\intercal\boldsymbol{Z})^{-1}\boldsymbol{Z}^\intercal\bar{\boldsymbol{X}}.$$

Over-identification

With such choice,

$$\begin{aligned}(\Sigma/\sigma^2)^{-1} &= \text{plim}_n\, n^{-1} \bar{\boldsymbol{X}} \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{Z}\bar{\boldsymbol{X}}}} \bar{\boldsymbol{X}} \\ &= \text{plim}_n\, n^{-1} \bar{\boldsymbol{X}}^\intercal \boldsymbol{P}_{\boldsymbol{Z}\bar{\boldsymbol{X}}} \left(\bar{\boldsymbol{X}}^\intercal \boldsymbol{P}_{\boldsymbol{Z}} \bar{\boldsymbol{X}}^\intercal\right)^{-1} \bar{\boldsymbol{X}}^\intercal \boldsymbol{P}_{\boldsymbol{Z}\bar{\boldsymbol{X}}} \\ &= \text{plim}_n\, n^{-1} \bar{\boldsymbol{X}}^\intercal \boldsymbol{P}_{\boldsymbol{Z}} \bar{\boldsymbol{X}},\end{aligned}$$

where we note $\boldsymbol{P}_{\boldsymbol{Z}}$ is symmetric, idempotent.

☞ This choice of $\boldsymbol{J}$ is optimal because

$$\bar{\boldsymbol{X}}^\intercal \left(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_{\boldsymbol{Z}\boldsymbol{J}}\right) \bar{\boldsymbol{X}} \succeq \mathbf{0}.$$

But $\bar{X} = \mathbb{E}[X|Z]$ is unknown. Nevertheless, $\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}$ is asymptotically equivalent to $\boldsymbol{P}_{\boldsymbol{Z}}\bar{\boldsymbol{X}}$.

Generalized 2SLS

Finally, under $\dim Z = l \geq k = \dim X$, the optimal estimating equation is

$$(\boldsymbol{P}_Z \boldsymbol{X})^\intercal (\boldsymbol{y} - \boldsymbol{X}\beta) = \boldsymbol{0},$$

which yields the **generalized 2SLS** estimator

$$\begin{aligned}\hat{\beta} &= (\boldsymbol{X}^\intercal \boldsymbol{P}_{\boldsymbol{Z}} \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{P}_{\boldsymbol{Z}} \boldsymbol{y} \\ &= \left[(\boldsymbol{P}_Z \boldsymbol{X})^\intercal (\boldsymbol{P}_Z \boldsymbol{X})\right]^{-1} (\boldsymbol{P}_Z \boldsymbol{X})^\intercal \boldsymbol{y},\end{aligned}$$

where $\boldsymbol{P}_Z \boldsymbol{X}$ is the fitted value of $\boldsymbol{X}$ from 1st-stage regression.

☞ CLT

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{d}{\Rightarrow} \mathcal{N}(\boldsymbol{0}, \Sigma), \quad \Sigma = \sigma^2 \, \mathbb{E}[ZX^\intercal]^{-1} \operatorname{cov}(Z) \, \mathbb{E}[ZX^\intercal]^{-T}.$$

☞ This works because $\mathbb{E}[X|Z]$ is **asymptotically independent** of the endogenous error ε^y, although they are dependent in finite samples!

Finite sample behavior

☞ Because 2SLS is a generalized form of division, its finite sample behavior is rather erratic, especially when instrument is **weak**!

Consider again the case of $\dim Z = \dim X = 1$

$$Y = \beta X + \sigma_1 \varepsilon_1, \quad X = \pi Z + \sigma_2 \varepsilon_2,$$

with $\varepsilon_1, \varepsilon_2 \sim \mathcal{N}(0,1)$ with correlation ρ.

We have

$$\hat{\beta} = \frac{\boldsymbol{z}^\intercal \boldsymbol{y}}{\boldsymbol{z}^\intercal \boldsymbol{x}} = \frac{\boldsymbol{z}^\intercal(\beta \boldsymbol{x} + \sigma_1 \boldsymbol{\varepsilon}_1)}{\boldsymbol{z}^\intercal \boldsymbol{x}} = \beta + \sigma_1 \frac{\boldsymbol{z}^\intercal \varepsilon_1}{\boldsymbol{z}^\intercal \boldsymbol{x}}.$$

It follows that

$$\hat{\beta} - \beta = \frac{\sigma_1 \boldsymbol{z}^\intercal \boldsymbol{\varepsilon}_1}{\boldsymbol{z}^\intercal(\pi \boldsymbol{z} + \sigma_2 \boldsymbol{\varepsilon}_2)}.$$

Letting $\boldsymbol{z}^\intercal \boldsymbol{z} = 1$ and writing $\varepsilon_1 = \varepsilon_3 + \rho \varepsilon_2$ for $\varepsilon_3 \sim \mathcal{N}(0,1)$ independent of ε_2

$$\hat{\beta} - \beta = \frac{\sigma_1 \boldsymbol{z}^\intercal(\boldsymbol{\varepsilon}_3 + \rho \boldsymbol{\varepsilon}_2)}{\pi + \sigma_2 \boldsymbol{z}^\intercal \boldsymbol{\varepsilon}_2}$$

Finite sample behavior

Now, taking **conditional expectation** with respect to ε_2 and noting $\boldsymbol{z}^\intercal \boldsymbol{\varepsilon}_2 \sim \mathcal{N}(0,1)$, we get

$$\mathbb{E}[\hat{\beta} - \beta | \boldsymbol{\varepsilon}_2] = \frac{\rho\sigma_1}{\sigma_2} \frac{W}{W + \pi/\sigma_2}, \quad W \sim \mathcal{N}(0,1).$$

- If $\rho = 0$, unbiased and reduced to OLS.
- If $\rho \neq 0$,
  - $\pi = 0$: 2SLS has non-diminishing bias $\rho\sigma_1/\sigma_2$.
  - $\pi \neq 0$: $\mathbb{E}[\hat{\beta} - \beta]$ **does not exist**, even though it is asymptotically unbiased!
    - Generalized $\hat{\beta}$ only has $(l - k)$ moments in finite samples (Kinal, 1980) in the identified/over-identified case ($l \geq k$).
    - Poor asymptotic behavior if $\pi \approx 0$, i.e., weak instrument.

Weak instruments

The asymptotics on $\hat{\beta}$ may be far from reality if instrument is weak. One needs to be **cautious** of this fact when doing inference on IV.

☞ Testing weak instrument

1. Stock and Yogo (2002) based on asymptotic embedding at local asymptotics $\pi = c/\sqrt{n}$.
2. Inference on IV after testing for weak instruments (Bi, Kang, and Taylor, 2020).

☞ Weak-instrument robust test (Anderson and Rubin, 1949) for testing $H_0 : \beta = \beta_0$.

Other topics on IV

☞ Identification strategies without assuming linear SEM

1. Homogeneity
2. Monotonicity and local average treatment effect (LATE).

☞ IV also implies semi-algebraic constraints (e.g., instrument inequalities) that can be used for falsification and partial identification.

References

- Anderson, Theodore W and Herman Rubin (1949). "Estimation of the parameters of a single equation in a complete system of stochastic equations". In: *The Annals of Mathematical Statistics* 20.1, pp. 46–63.
- Bi, Nan, Hyunseung Kang, and Jonathan Taylor (2020). "Inferring Treatment Effects After Testing Instrument Strength in Linear Models". In: *arXiv preprint arXiv:2003.06723*.
- Kinal, Terrence W (1980). "The existence of moments of k-class estimators". In: *Econometrica: Journal of the Econometric Society*, pp. 241–249.

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti (2004). "Economic shocks and civil conflict: An instrumental variables approach". In: *Journal of political Economy* 112.4, pp. 725–753.

Stock, James H and Motohiro Yogo (2002). *Testing for weak instruments in linear IV regression*. Tech. rep. National Bureau of Economic Research.