Instrumental Variable II

Ye Wang
University of North Carolina at Chapel Hill

Linear Methods in Causal Inference
POLI784

Review

- We discussed how to identify causal effects when there exists non-compliance.
- In this case, treatment assignment Z_i no longer equals treatment status D_i.
- We can always identify the intention-to-treat effect but may care about the local average treatment effect (LATE), or the effect on the compliers.
- It can be identified if 1. Z_i is randomly assigned; 2. it affects Y_i only through D_i; 3. it changes the value of D_i monotonically.
- Then, we can estimate the LATE using the Wald estimator.

History of IV

- We called treatment assignment Z_i under non-compliance an instrumental variable (IV).
- But the idea of using an IV for causal inference was proposed in a very different context.
- It was introduced by economists to study simultaneous structural equations.
- People used to believe that we can describe a social system (e.g., the US economy) with a large number of equations.
- Variables on the left-hand side are referred to as endogenous, while those only appear on the right-hand side are called exogenous.
- Identification means that we can estimate coefficients in these models consistently.
- Instrumental variables were proposed as a solution to the identification problem.

History of IV

- An economist observes the everyday price and quantity of transaction for fish in a market over N days: $(P_i, Q_i)_{i=1}^N$.
- She wants to identify the demand curve of fish: $P_i = d(Q_i)$.
- But we only have the price and quantity at equilibrium, which is also affected by the supply curve: $P_i = s(Q_i)$.
- Suppose both the demand and the supply curves are linear:

$$
\begin{aligned}
P_{di} &= a_d - b_d Q_{di} + \varepsilon_{di}, \\
P_{si} &= a_s + b_s Q_{si} + \varepsilon_{si}.
\end{aligned}
$$

- We know that at equilibrium,

$$
\begin{aligned}
P_i &= \frac{a_s b_d + a_d b_s + b_s(\varepsilon_{di} - \varepsilon_{si})}{b_d + b_s}, \\
Q_i &= \frac{a_d - a_s + \varepsilon_{di} - \varepsilon_{si}}{b_d + b_s}.
\end{aligned}
$$

History of IV

- Even with a large sample, we can only estimate
$$\frac{a_s b_d + a_d b_s}{b_d + b_s} \text{ and } \frac{a_d - a_s}{b_d + b_s}$$
but not (a_d, b_d).
- Suppose there is a shock Z_i on the supply side such as storms at sea:
$$\begin{aligned} P_{di} &= a_d - b_d Q_{di} + \varepsilon_{di}, \\ P_{si} &= a_s + b_s Q_{si} + c_s Z_i + \varepsilon_{si}, \\ Z_i &\perp (\varepsilon_{di}, \varepsilon_{si}). \end{aligned}$$
- Z_i is known as an instrumental variable.

History of IV

- Now, at equilibrium, we have

$$P_i = \frac{a_s b_d + a_d b_s + c_s b_d Z_i + b_s(\varepsilon_{di} - \varepsilon_{si})}{b_d + b_s},$$

$$Q_i = \frac{a_d - a_s - c_s Z_i + \varepsilon_{di} - \varepsilon_{si}}{b_d + b_s}.$$

- From these relationships, we can obtain estimates of the two slopes

$$\frac{c_s b_d}{b_d + b_s} \text{ and } \frac{-c_s}{b_d + b_s}.$$

- Their ratio allows us to identify the parameter b_d.

IV in structural models

- Let's consider an economic model of income and education.
- An individual i maximizes their return by deciding whether to attend college, $D_i \in \{0, 1\}$.
- We assume that the return is decided by

$$m(D_i, \varepsilon_i) - c(D_i, Z_i, \eta_i),$$

  where $m()$ is their expected income, $c()$ refers to the cost of attending college, Z_i represents observable exogenous factors that affect the cost (proximity to hometown), ε_i and η_i are unobservable factors.
- We are interested in the relationship between expected income and college education, $Y_i = m(D_i, \varepsilon_i)$.

IV in structural models

- D_i is an endogenous choice as

$$D_i = \arg\max_d [m(d, \varepsilon_i) - c(d, Z_i, \eta_i)].$$

- We can write $D_i = g(Z_i, \nu_i)$, where $\nu_i = (\varepsilon_i, \eta_i)'$.
- Now, we have a "triangular system:"

$$\begin{aligned} Y_i &= m(D_i, \varepsilon_i), \\ D_i &= g(Z_i, \nu_i), \\ Z_i &\perp \nu_i, \varepsilon_i \not\perp \nu_i. \end{aligned}$$

- Ususally we assume $g()$ is monotonic:

$$\text{If } g(Z_i, \nu_i) > g(Z_i', \nu_i), \text{ then } g(Z_i, \nu_i') > g(Z_i', \nu_i').$$

IV in structural models

- We call Z_i in the triangular system an instrumental variable.
- Like the supply shock in the fishing example, it provides exogenous variations for causal identification in the system.
- The triangular system can incorporate many other scenarios and allows for arbitrary heterogeneity in effects.
- Both ε_i and ν_i might be high-dimensional.
- There is no restriction on whether the variables are continuous or discrete.
- This framework is built upon economic theories and captures complexities in reality.
- Non-compliance is not mentioned in such a system.

IV in linear models

- It is impossible to identify any causal parameter of interest in the triangular system without more structural restrictions.
- Let's consider the simplest scenario: both $m()$ and $g()$ are linear functions with homogeneous effects and no intercept, then

$$
\begin{aligned}
Y_i &= \tau D_i + \varepsilon_i, \\
D_i &= \delta Z_i + \nu_i, \\
Z_i &\perp \nu_i, \varepsilon_i \not\perp \nu_i.
\end{aligned}
$$

- $Cov(D_i, \varepsilon_i) \neq 0$, hence regressing Y_i on D_i leads to bias.
- But we can still estimate τ using a two-step approach.

Two-stage least sqaures

- First, we estimate the second equation with OLS and obtain $\hat{\delta}$.
- This is known as the "first stage" regression.
- Then, note that

$$\begin{aligned} Y_i &= \tau D_i + \varepsilon_i \\ &= \tau(\delta Z_i + \nu_i) + \varepsilon_i \\ &= \xi Z_i + \tilde{\varepsilon}_i, \end{aligned}$$

  where $\xi = \tau\delta$ and $Z_i \perp \tilde{\varepsilon}_i = \tau\nu_i + \varepsilon_i$.
- Hence, regressing Y_i on Z_i leads to a consistent estimate of $\tau\delta$.
- This is known as the "reduced-form" regression.
- The ratio of the two regression estimates is consistent for τ.
- This algorithm is known as the two-stage least squares (2SLS).

Two-stage least sqaures

- There are two other approaches to estimate τ.
- The first stage is always necessary.
- Let's denote the predicted value of D_i as $\hat{D}_i$ and the regression residual as $\hat{\nu}_i$.
- The second approach is to regress Y_i on $\hat{D}_i$ (the second stage).
- The OLS estimate will be consistent for τ.
- Consider the matrix form of the equations:

$$\mathbf{Y} = \mathbf{D}\tau + \varepsilon,$$
$$\mathbf{D} = \mathbf{Z}\delta + \nu.$$

Two-stage least sqaures

- Remember that

$$
\begin{aligned}
\hat{\delta} &= (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{D}), \\
\widehat{\mathbf{D}} &= \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{D}) = \mathbf{P_Z}\mathbf{D}, \\
\hat{\nu} &= \mathbf{D} - \widehat{\mathbf{D}} = (\mathbf{I} - \mathbf{P_Z})\mathbf{D} = \mathbf{Q_Z}\mathbf{D}.
\end{aligned}
$$

- The OLS estimate from regressing $\mathbf{Y}$ on $\widehat{\mathbf{D}}$ equals:

$$
\begin{aligned}
\hat{\tau}_{2SLS} &= \left(\widehat{\mathbf{D}}'\widehat{\mathbf{D}}\right)^{-1}\left(\widehat{\mathbf{D}}'\mathbf{Y}\right) \\
&= \left(\mathbf{D}'\mathbf{P_Z}\mathbf{P_Z}\mathbf{D}\right)^{-1}\left(\mathbf{D}'\mathbf{P_Z}\mathbf{Y}\right) \\
&= \left(\mathbf{D}'\mathbf{P_Z}\mathbf{D}\right)^{-1}\left(\mathbf{D}'\mathbf{P_Z}(\mathbf{D}\tau + \varepsilon)\right) \\
&= \tau + \left(\mathbf{D}'\mathbf{P_Z}\mathbf{D}\right)^{-1}\left(\mathbf{D}'\mathbf{P_Z}\varepsilon\right) \to \tau.
\end{aligned}
$$

Two-stage least sqaures

- It is natural to see that

$$Var[\hat{\tau}_{2SLS}] = \left(\mathbf{D}'\mathbf{P}_{\mathbf{Z}}\mathbf{D}\right)^{-1}\left(\mathbf{D}'\mathbf{P}_{\mathbf{Z}}\varepsilon\varepsilon'\mathbf{P}_{\mathbf{Z}}\mathbf{D}\right)\left(\mathbf{D}'\mathbf{P}_{\mathbf{Z}}\mathbf{D}\right)^{-1}.$$

- It is more complicated than the variance for the OLS estimator but also takes the sandwich form.
- Intuitively, we first project D_i onto the space spanned by Z_i.
- Next, we project Y_i onto the same space, with the projected D_i as the bases.
- We can estimate the variance by replacing ε with the regression residuals in the second stage.
- Incorporating covariates is straightforward by replacing $\mathbf{D}$ with $\tilde{\mathbf{D}} = (\mathbf{D}, \mathbf{X})$ and $\mathbf{Z}$ with $\tilde{\mathbf{Z}} = (\mathbf{Z}, \mathbf{X})$.

Control function

- The third approach is known as "control function" in econometrics.
- In the second stage, we regress Y_i on D_i and $\hat{\nu}_i$.
- We can show that the estimated coefficient for D_i will be consistent for τ.
- Note that ν_i is the endogenous part in D_i and $\hat{\nu}_i$ is unbiased for ν_i.
- If we can control for the endogenous part, the estimation will be unbiased.
- These approaches give you numerically equivalent results when the models are correct.

Control function

- Define the residual-making matrix with regards to $\hat{\nu}$:

$$\mathbf{Q}_{\hat{\nu}} = \mathbf{I} - \hat{\nu}(\hat{\nu}'\hat{\nu})^{-1}\hat{\nu}' = \mathbf{I} - \mathbf{Q_Z D}\left(\mathbf{D}'\mathbf{Q_Z D}\right)^{-1}\mathbf{D}'\mathbf{Q_Z}.$$

- We can see that $\mathbf{Q}_{\hat{\nu}}\mathbf{D} = \mathbf{D} - \mathbf{Q_Z D} = \mathbf{P_Z D}$.
- Multiplying $\mathbf{Q}_{\hat{\nu}}$ on both sides of the main equation, we have

$$\tilde{\mathbf{Y}} = \mathbf{Q}_{\hat{\nu}}\mathbf{Y} = \mathbf{Q}_{\hat{\nu}}\mathbf{D}\tau + \mathbf{Q}_{\hat{\nu}}\varepsilon = \tilde{\mathbf{D}} + \tilde{\varepsilon}.$$

- From the FWL theorem,

$$\begin{aligned}\hat{\tau}_{CF} =& \left(\tilde{\mathbf{D}}'\tilde{\mathbf{D}}\right)^{-1}\left(\tilde{\mathbf{D}}'\tilde{\mathbf{Y}}\right) = \tau + \left(\tilde{\mathbf{D}}'\tilde{\mathbf{D}}\right)^{-1}\left(\tilde{\mathbf{D}}'\tilde{\varepsilon}\right) \\ =& \tau + \left(\mathbf{D}'\mathbf{Q}_{\hat{\nu}}\mathbf{D}\right)^{-1}\left(\mathbf{D}'\mathbf{Q}_{\hat{\nu}}\varepsilon\right) \\ =& \tau + \left(\mathbf{D}'\mathbf{P_Z D}\right)^{-1}\left(\mathbf{D}'\mathbf{P_Z}\varepsilon\right) = \hat{\tau}_{2SLS}.\end{aligned}$$

Control function: application

```
## The OLS estimate is 3.90903
## The 2SLS estimate is 2.905916
## The 2SLS estimate is 2.905916
## The control function estimate is 2.905916
```

From 2SLS to LATE

- We have reviewed the literature on instrumental variables following the econometric tradition.
- We didn't not use the potential outcome notations at all!
- Are the 2SLS estimates causal from the design-based perspective?
- Note that the triangular system satisfies Assumptions 1-4 for identifying the LATE.
- Random assignment results from that $Z_i \perp \nu_i$.
- Exclusion restriction and the first stage hold due to the functional form.
- Monotonicity is assumed for $g()$.

From 2SLS to LATE

- Angrist, Imbens, and Rubin (1996) first showed that when D_i and Z_i are binary, $\hat{\tau}_{2SLS} = \hat{\tau}_{Wald}$.
- The result holds even when the effects are heterogeneous.
- In other words, $\hat{\tau}_{2SLS}$ can be interpreted as an estimate for the average treatment effect on the compliers (LATE).
- Compliers are the agents who are encouraged to select into the treatment by the instrument.
- We start from an economic model and end up with an interpretation rooted in the potential outcomes framework.
- The economic approach and the statistical approach are capturing the same concepts.
- We can justify the potential outcomes framework with social science theory, or analyze social science problems with the idea of counterfactual.
- This finding won Angrist and Imbens a Nobel prize.

From 2SLS to LATE

- In this case, the triangular system boils down to:

$$Y_i = \tau_i D_i + \varepsilon_i,$$
$$D_i = \mathbf{1}\{(\delta_i Z_i + \nu_i) \geq 0\}.$$

- We can see that $D_i(0) = \mathbf{1}\{\nu_i \geq 0\}$ and $D_i(1) = \mathbf{1}\{\delta_i + \nu_i \geq 0\}$.
- δ_i and ν_i determine the response of unit i to Z_i hence i's type.
- For example, compliers are those with $-\delta_i \leq \nu_i < 0$.
- The analysis in Angrist, Imbens, and Rubin (1996) indicates that the 2SLS estimator's result is completely driven by these units.

2SLS: application

```
## The LATE is 1.855

## Estimate from the Wald estimator is 0.107

## The 2SLS estimate is 0.107

## SE of the 2SLS estimate is 0.914
```

From 2SLS to LATE

- These results led to the credibility revolution in the 90s.
- Using IVs in observational studies to identify the LATE became a fad.
- Angrist (1990): draft lottery in the Vietnam war - veteran status - earnings on the labor market.
- Angrist and Krueger (1991): birth season - age when dropping out of high school - earnings on the labor market.
- Acemoglu, Johnson, and Robinson (2001): settlers' mortality - inclusive institutions - economic development.
- But many applications turned out to be not as credible as we thought.

References I

Acemoglu, Daron, Simon Johnson, and James A Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91 (5): 1369–1401.

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *The American Economic Review*, 313–36.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–55.

Angrist, Joshua D, and Alan B Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106 (4): 979–1014.