Canonical Research Designs III: Instrumental Variables I

Paul Goldsmith-Pinkham

March 29, 2022

Today's topic

- In so many examples, we've worried about estimating the effect of some treatment D_i on Y_i, but concerned that this estimate will be biased
- In circumstances where we had strong ignorability (due to perhaps randomization), we felt confident that this was not a concern
- But what about when this isn't the case? We need to create "as-if" random variation in D_i
- The very popular solution: instrumental variables

Today's topic

- There is a huge amount of material to cover regarding IV
- Today, we will be covering the setup and overview for how to think about IV
- Next class, covering a wide range of practical issues and problems that come up for researchers working on these topics

Many faces of Instrumental Variables

- This class is clearly not the first place you've been exposed to IV
  - Now so ubiqutous that likely seen multiple times

- More interestingly, you've likely seen it in many forms

- To start class, we'll discuss what an IV approach can mean or looks like in the three types of tools we discussed way back in Lecture 1
  1. First, we'll try to define what we mean by an instrumental variable, initially using the DAG notation
  2. Then, we'll discuss the classic structural econometrics modeling form, and link it to GMM
     - Moments!
  3. Finally, we'll discuss the potential outcomes or design-based setup

Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments

Joshua D. Angrist and Alan B. Krueger

What is an instrumental variable?

The method of instrumental variables is a signature technique in the econometrics toolkit. The canonical example, and earliest applications, of instrumental variables involved attempts to estimate demand and supply curves.[1] Economists such as P.G. Wright, Henry Schultz, Elmer Working and Ragnar Frisch were interested in estimating the elasticities of demand and supply for products ranging from herring to butter, usually with time series data. If the demand and supply curves shift over time, the observed data on quantities and prices reflect a set of equilibrium points on both curves. Consequently, an ordinary least squares regression of quantities on prices fails to identify—that is, trace out—either the supply or demand relationship.

What is an instrumental variable?

P.G. Wright (1928) confronted this issue in the seminal application of instrumental variables: estimating the elasticities of supply and demand for flaxseed, the source of linseed oil.[2] Wright noted the difficulty of obtaining estimates of the elasticities of supply and demand from the relationship between price and quantity alone. He suggested (p. 312), however, that certain "curve shifters"—what we would now call instrumental variables—can be used to address the problem: "Such additional factors may be factors which (A) affect demand conditions without affecting cost conditions or which (B) affect cost conditions without affecting demand conditions." A variable he used for the demand curve shifter was the price of substitute goods, such as cottonseed, while a variable he used for the supply curve shifter was yield per acre, which can be thought of as primarily determined by the weather.

What is an instrumental variable?

Studying agricultural markets in the 1920s, the father and son research team of Phillip and Sewall Wright were interested in a challenging problem of causal inference: how to estimate the slope of supply and demand curves when observed data on prices and quantities are determined by the intersection of these two curves. In other words, equilibrium prices and quantities—the only ones we get to observe—solve these two stochastic equations at the same time. Upon which curve, therefore, does the observed scatterplot of prices and quantities lie? The fact that population regression coefficients do not capture the slope of any one equation in a set of simultaneous equations had been understood by Phillip Wright for some time. The IV method, first laid out in Wright (1928), solves the statistical simultaneous equations problem by using variables that appear in one equation to shift this equation and trace out the other. The variables that do the shifting came to be known as *instrumental variables* (Reiersol, 1941).

- Variables that do the shifting!

The curve shifting example

- Consider a simple supply and demand curve setup

$$\text{quantity}_d = \alpha_1 + \text{price}\gamma_1 + W_1\tau_1 + u_1 \quad (1)$$

$$\text{quantity}_s = \alpha_2 + \text{price}\gamma_2 + W_2\tau_2 + u_2 \quad (2)$$

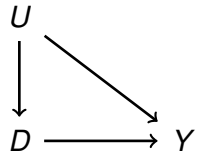
- Equilibrium values of price and quantity are such that these equations equal
  - These can vary due to many various regions – they can be specific to demand or supply, or happen to both

- The *observed* values of price and quantity will give a cloud of points with no real interpretation as either demand or supply curve coefficients (elasticities)

- What we need is shifters of these curves to create variation that traces out either supply or demand
  - Notably, this traces out only a local part of the curve!

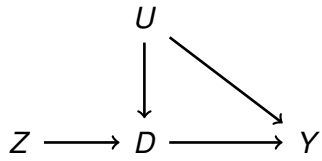
What is an instrumental variable?

- “Curve-shifting variable” is not a particularly extensible concept
- Let’s start with the definition in the context of a DAG
- Consider an effect we are interested in identifying: D on Y
  - In this setting, we know it is not identifiable



What is an instrumental variable?

- Now, we have a variable Z which can identify two effects:
  - Z on D
  - Z on Y
- What is the content of this instrumental variable, Z?
  - It affects Y (**Relevance**)
  - It only affects Y through D (**Exclusion**)
- Without further assumptions, it won't be possible to identify the effect of D on Y using this, but it highlights the features of an IV
  - We'll discuss why shortly



Structural version of instruments - GMM and 2SLS

- The canonical setup with an IV:

$$Y_i = D_i\beta + W_i\gamma_1 + \epsilon_i$$
$$D_i = Z_i\pi + W_i\gamma_2 + u_i$$

  with W_i are a set of exogeneous controls

- A couple notable features about this setup:
  - We've assumed a very parametric model for Y_i
  - In particular, we've assumed a constant effect of D_i on Y_i

- The necessary assumptions to identify D_i in this setting are straightforward:
  - Relevance: $\pi \neq 0$
  - Exclusion: $E(\epsilon_i Z_i | W_i) = 0$

Structural version of instruments - GMM and 2SLS

- The exclusion restriction can be slightly opaque
  - The ϵ_i captures the set of "other" things that can happen
  - But can be harder to map into a counterfactual way of discussing outcomes
- One useful result: let $Z_i^* = Z_i - E(Z_i|W_i)$
  - Then the exclusion restriction can be viewed as saying that $E(\epsilon_i Z_i^*) = 0$
  - That is, the variation in Z_i above and beyond W_i has to be exogeneous for ϵ_i
- Why is this often written as a system of linear equations?
  - Historical precedent is part of it – linear demand systems
  - But, it turns out it is "optimal" in a particular sense

Structural version of instruments - GMM and 2SLS

- In GMM, there's just a more general statement – ignore the first stage for a moment
  - We still maintain the linear second stage – the "structural model"
  - To write this compactly, let $\widetilde{D}_i = (D_i, W_i)$ and $\widetilde{Z}_i = (Z_i, W_i)$

- Recall that the exclusion restriction gives us a set of K *moments*.
  - $E((Y_i - D_i\beta - W_i\gamma)Z_i)$ (excluded instruments)
  - $E((Y_i - D_i\beta - W_i\gamma)W_i)$ (exogeneous instruments)
  - Or compactly, $\mathbf{g}(\beta, \gamma) = E((Y_i - \widetilde{D}_i\widetilde{\beta})\widetilde{Z}_i)$

- Then, recall that given these K moments, for a $K \times K$ positive-definite weight matrix Ω, we can define our linear GMM estimator as the solution to the following problem:

$$(\beta_0, \gamma_0) = \arg\min_{\beta,\gamma} \mathbf{g}(\beta, \gamma)'\Omega\mathbf{g}(\beta, \gamma) \tag{3}$$

Structural version of instruments - GMM and 2SLS

- Fortunately, because **g** is linear, solving for the minimizer is analytically tractable
  - More non-linear second stages are solveable as well, but typically need numerical solutions

- Our general solution is a function of our choice of Ω:

$$\hat{\tilde{\beta}} = \frac{\tilde{D}'\tilde{Z}\Omega\tilde{Z}'\tilde{Y}}{\tilde{D}'\tilde{Z}\Omega\tilde{Z}'\tilde{D}}$$

- Turns out that if the exclusion restriction holds (and relevance, such that the denominator isn't zero), it really doesn't matter what Ω is – your estimator will converge
  - If there's no unobserved heterogeneity in β!

$$\hat{\tilde{\beta}} - \tilde{\beta} = \frac{\tilde{D}'\tilde{Z}\Omega\tilde{Z}'\epsilon}{\tilde{D}'\tilde{Z}\Omega\tilde{Z}'\tilde{D}} \to 0$$

  - All thanks to $E(\tilde{Z}\epsilon) = 0$

Structural version of instruments - GMM and 2SLS

- Where does 2SLS come in? Contrast the formula for 2SLS and this GMM estimator:

$$\hat{\tilde{\beta}}_{2SLS} = \frac{\tilde{D}'\tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'\tilde{Y}}{\tilde{D}'\tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'\tilde{D}}$$

$$\hat{\tilde{\beta}}_{GMM} = \frac{\tilde{D}'\tilde{Z}\Omega\tilde{Z}'\tilde{Y}}{\tilde{D}'\tilde{Z}\Omega\tilde{Z}'\tilde{D}}$$

- 2SLS is a special GMM setting with the weight matrix equal to the inverse of the covariance of the instruments
  - Recall that in GMM, there are ways to get "better" weight matrices to minimize the variance of the estimator (e.g. 2-step GMM, iterated GMM, etc.)
  - However, it turns out 2SLS weight matrix is optimal under homoskedasticity!

Some useful features of 2SLS

$$\hat{\tilde{\beta}}_{2SLS} = \frac{\tilde{D}'\tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'\tilde{Y}}{\tilde{D}'\tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'\tilde{D}}$$

- Recall the projection matrix $P_Z = \tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'$
  - Important property: idempotency – e.g. $P_Z P_Z = P_Z$

$$\hat{\tilde{\beta}}_{2SLS} = \frac{\tilde{D}' P_Z P_Z \tilde{Y}}{\tilde{D}' P_Z P_Z \tilde{D}} = \frac{\hat{\tilde{D}}'\hat{\tilde{Y}}}{\hat{\tilde{D}}'\hat{\tilde{D}}}$$

- Hence it's really the projection of D onto Z, and the projection of Y onto Z
  - So the numerator is the covariance of the predicted pieces, and the denominator is the variance of the predicted endogeneous variable

- This is exactly what our curve shifter had in mind – we need variation in the predicted values

Structural version of instruments - GMM and 2SLS

- This approach is highly focusing on a parameteric specification, but notably the important assumption is that $E(\epsilon_i Z_i) = 0$ and $E(D_i Z_i) \neq 0$. Note that this is much weaker than random assignment!

- But, it is kind of wonky to assume mean independence and not assume full independence

- Why? Well, consider transforming the outcome Y by taking a log (or some other nonlinear transformation). With only mean indpeendence, our instrument is not necessarily valid anymore.
  - That seems like a undesirable property!
  - Unfortunately, comparable to our discussion of difference-in-difference!

The necessary assumptions so far

- So far, we need the following assumptions (and this is what you should always discuss when writing a paper on IV):
  1. relevance $E(D_i Z_i)$
  2. exclusion $E(Z_i \epsilon_i)$

- Tricky part starts now. Two issues with this setup:
  - we have assumed homogeneous effects. E.g. β is the same for all individuals.
    - This is fixable in the model, but question is what estimand do we have?
  - It's not a very coherent "design-based" setup. In other words, it's challenging to think about the shifter Z in terms of the potential outcomes of the outcomes
    - This can make it hard to suss out the validity of the design!

- Large literature in the 1990s (and continuing forward) focused on taking the Neyman-Rubin Casual Model (NRCM) with potential outcomes and mapping it to IV
  - Pushed by Josh Angrist, Guido Imbens, and Don Rubin

Imbens and Angrist (1994)

- Start by focusing on the simplest of cases: binary instrument Z, binary treatment D, and no controls

- Potential outcomes framework needs to be extended to allow an instrument!
  - Define $Y_i(D_i(Z_i), Z_i)$ and $D_i(Z_i)$ as two forms of potential outcomes
  - The *exclusion* restriction here is that $Y_i(D_i(Z_i), Z_i) = Y_i(D_i(Z_i))$, e.g. Z_i only has an effect on Y_i through D_i
  - *Relevance* is that $P(w) = E(D_i|Z_i = w)$ varies across w

- Key point is the $Y_i(1) - Y_i(0)$ can be different for every individual
  - Unlike in the structural models we wrote before, where β was constant

- We'll assume that Z is completely randomly assigned relative to the potential outcomes of Y and D

No ATE is guaranteed - Imbens and Angrist (1994)

- Key point: consider $E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)$ (where $P(1) > P(0)$)

$$\begin{aligned} E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0) &= E(D_i(1)Y_i(1) + (1 - D_i(1))Y_i(0)|Z_i = 1) \\ &\quad - E(D_i(0)Y_i(1) + (1 - D_i(0))Y_i(0)|Z_i = 0) \\ &= E((D_i(1) - D_i(0))(Y_i(1) - Y_i(0))) \\ = Pr(D_i(1) - D_i(0) = 1) &\times E(Y_i(1) - Y_i(0))|D_i(1) - D_i(0) = 1) \\ -Pr(D_i(1) - D_i(0) = -1) &\times E(Y_i(1) - Y_i(0))|D_i(1) - D_i(0) = -1) \end{aligned}$$

- There's a lot to unpack here.

No ATE is guaranteed - Imbens and Angrist (1994)

$$
\begin{aligned}
&E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0) = \\
&= Pr(D_i(1) - D_i(0) = 1) \times E(Y_i(1) - Y_i(0))|D_i(1) - D_i(0) = 1) \\
&- Pr(D_i(1) - D_i(0) = -1) \times E(Y_i(1) - Y_i(0))|D_i(1) - D_i(0) = -1)
\end{aligned}
$$

- First, note that while we assumed that the propensity score was increasing, it does *not* imply that it's increasing for everyone
- Second, we are only identifying the effects of $D\,(Y_i(1) - Y_i(0))$ for those individuals who behavior shifted due to the change in Z
- Third, without restrictions on Y_i, this effect can be zero or even negative, even if the true causal effect is positive!
  - Those shifted into participating by Z could be exactly cancelled by those who shift out

Local Average Treatment Effect (LATE)

- Two potential solutions to this issue:
  1. In a constant effects world, this problem does not exist!
  2. Secondly, if there exists an instrument such that $Pr(D_i(1) - D_i(0) = -1) = 0$, then you're fine as well (e.g. one sided compliance)

- Key innovation: with *monotonicity*, can identify the Local Average Treatment Effect

- **Monotonicity**: $D_i(1) \geq D_i(0))$ for all i (or vice versa)
  - All effects must be monotone in the same direction
  - This is fundamentally untestable! (also suffers from fundamental problem of causal inference)

- Conditional on assuming monotonicity, then the Wald ratio estimates the LATE:

$$\begin{aligned}
\tau_{LATE} &= \frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(D_i|Z_i = 1) - E(D_i|Z_i = 0)} \\
&= \frac{Pr(D_i(1) - D_i(0) = 1) \times E(Y_i(1) - Y_i(0))|D_i(1) - D_i(0) = 1)}{E(D_i|Z_i = 1) - E(D_i|Z_i = 0)} \\
&= E(Y_i(1) - Y_i(0))|D_i(1) - D_i(0) = 1)
\end{aligned}$$

What does this mean?

- Fundamentally, this means that an IV strategy only identifies (non-parametrically) the effect of a treatment for those who respond to the treatment.
  - Monotonicity ensures that the responders all go in one direction

- Language used to describe these groups:
  - Always-takers: $D_i(1) = D_i(0) = 1$
  - Never-takers: $D_i(1) = D_i(0) = 0$
  - Compliers: $D_i(1) - D_i(0) = 1$
  - Defiers: $D_i(1) - D_i(0) = -1$

- Monotonicity ensures that only one of the compliers or defiers exists
  - The compliers can be very different! This is a particular subgroup
  - Important to understand potential differences

Loosening restrictiveness + 2SLS

- This setup is very specific (two binary measures) but it is straightforward to generalize to a multi-valued *instrument*

- Slightly more challenging (notationally) to generalize to a multivalued treatment in a non-parametric way
  - Key concept is an **average causal reponse curve** – effectively a combination of weighted derivatives, depending on where the instrument shifts participation
  - Angrist + Imbens (1995, JASA)

Average Causal Response Curve

- Consider a regression of (log) weekly earnings (Y_i) on years of schooling (S_i) with controls W_i, where we use the quarter-of-birth (Z_i) to instrument for schooling

$$Y_i = W_i\gamma + S_i\tau + \epsilon_i$$

- Recall the potential outcome notation for years of schooling $\in \{0, \ldots, J\}$
  - We can then consider relative comparisons: $\tau_{j,j-1} = E(Y_i(j) - Y_i(j-1))$
  - Note that if this were a linear effect, $\tau_{j,j-1} = 0.5\tau_{j,j-2}$, etc.

- Now consider the potential years of schooling defined by Z_i: $S_i(Z_i)$
  - Can consider this binary (first quarter vs. later)

Average Causal Response Curve

- Under independence assumptions and monotonicity ($S_i(1) - S_i(0) \geq 0$ or vice versa) for each person,

$$\frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(S_i|Z_i = 1) - E(S_i|Z_i = 0)} = \sum_{j=1}^{J} \omega_j E(Y_i(j) - Y_i(j-1)|S_i(1) \geq j > S_i(0))$$

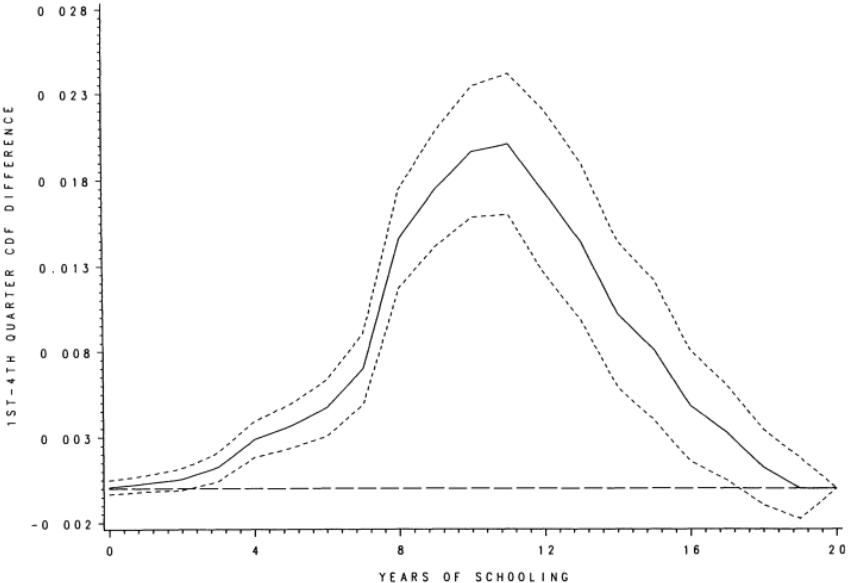
where

$$\omega_j = \frac{Pr(S_i(1) \geq j > S_0)}{\sum_{i=1}^{J} Pr(S_i(1) \geq i > S_0)}.$$

- ω_j can be consistently estimated using the data

- Key takeaway: weighting up non-parametric treatments as a function of where the instrument induces response

Average Causal Response Curve weights



Next class

- Monotonicity is a powerful tool for ensuring the weighted averages
  - but a remarkably strong assumption in some places!
- If it fails, it doesn't mean that there isn't a csaul effect identified, but researchers should be careful
  - Next class, will discuss sensitivities and other issues
- An important note: the IV estimate is just a rescaled reduced form estimate
  - If Z is truly randomly assigned, the reduced form is a valid estimate
  - Do you inherently need the rescaled estimate?