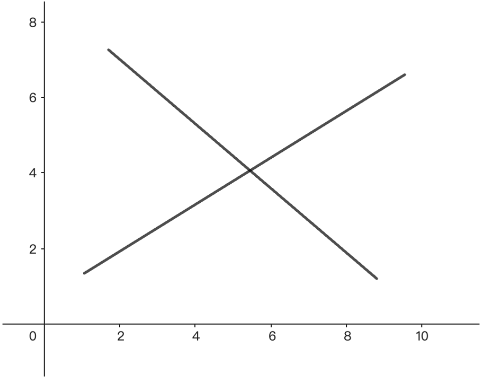
Overview

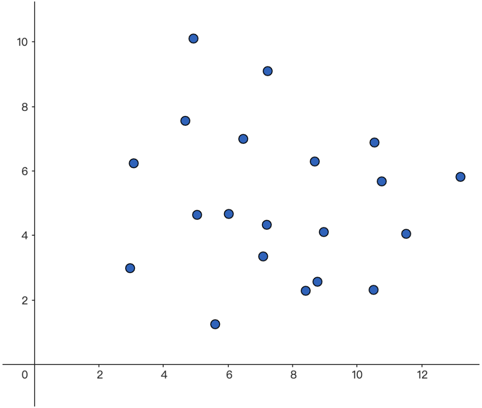
- Instrumental Variable (IV) technique dates back to 1928
- First used to identify demand and supply curve
- Then also used to address measurement error issues
- Again, talk about the restricted case first and then the general case

Demand and Supply Motivation



- The usual demand and supply diagram
- How do we estimate the demand and supply curve?

The Real Data



- Say you got data from a retail firm, the data looks like this
- Why? Because the demand and supply move around at the same time

Setup the Stage

- Recall the control for observables that we talked about previously

$$Y_i = \alpha + \rho D_i + \eta_i$$
$$Y_i = \alpha + \rho D_i + \gamma A_i + \nu_i$$

- Recall that ρ is the (homogeneous) treatment effect of D_i
- e.g., D_i go to school, Y_i wage
- We change X_i to A_i since we assume it's unobserved, e.g., ability
- In general $E[D_i\eta_i] \neq 0$
- Remain the CIA, hence $E[D_i\nu_i] = 0$ (Why?)
- Suppose we observe A_i, then OLS can recover ρ
- But we don't! How to recover ρ?

Instrumental Variable Idea (Mathematically)

- Suppose now we have a variable called Z_i that is
  - Correlated with the treatment assignment D_i, $Cov(D_i, Z_i) \neq 0$, (Relevance)
  - Uncorrelated with any other determinants of Y_i, $Cov(\eta_i, Z_i) = 0$, (Exclusion)
- Then we have

$$\frac{Cov(Y_i, Z_i)/Var(Z_i)}{Cov(D_i, Z_i)/Var(Z_i)} = \frac{Cov(Y_i, Z_i)}{Cov(D_i, Z_i)} = \frac{Cov(\alpha + \rho D_i + \eta_i, Z_i)}{Cov(D_i, Z_i)}$$
$$= \rho \frac{Cov(D_i, Z_i)}{Cov(D_i, Z_i)} + \frac{Cov(\eta_i, Z_i)}{Cov(D_i, Z_i)} = \rho$$

- Note that the LHS is actually the ratio of two regression coefficient
- Coefficient of regressing Y_i on Z_i divided by coefficient of regressing D_i on Z_i
- Meaning that OVB is not an issue if we have such a Z_i
- $\because$ $Cov(\eta_i, Z_i) = 0$, Z_i is sometimes called the exogeneous variation

First Stage and Reduced Form

- Let's write out these two regressions, D_i on Z_i and Y_i on Z_i:

$$D_i = \pi_{10} + \pi_{11} Z_i + \xi_{1i} \text{ (First Stage)}$$
$$Y_i = \pi_{20} + \pi_{21} Z_i + \xi_{2i} \text{ (Reduced Form)}$$

- The above results suggest that $\rho = \frac{\pi_{21}}{\pi_{11}}$
- We can also allow for covariates:

$$D_i = X_i' \pi_{10} + \pi_{11} Z_i + \xi_{1i} \text{ (First Stage)}$$
$$Y_i = X_i' \pi_{20} + \pi_{21} Z_i + \xi_{2i} \text{ (Reduced Form)}$$

What Magic Just Happened?

- Remember that CIA is a very strong assumption
- Potentially there are many variables that you should control for
- Many if not most of them could be unobservable (e.g., ability)
- Essentially you need an experiment to identify the causal effect
- The Z_i solves the problem regardless of the number of omitted variables

Why IV Works (Intuitively)

- We can think of the instrumental variable Z_i as a “clean” source of variation
- Like, for example, a coin flip
- Something that affects whether you get $D_i = 1$ but not relate to the outcome
- The IV strategy uses only the “clean” part of the variation
- Look at the reduced form again

$$Y_i = \pi_{20} + \pi_{21} Z_i + \xi_{2i}$$

  it tells us how much outcome is affected by the “clean variation”
- The first stage is telling us how much to “scale up” from the clean variation

$$D_i = \pi_{10} + \pi_{11} Z_i + \xi_{1i}$$

Example: Causal Effects of Schooling on Earning

- Talk about a famous paper Angrist and Krueger (2001)
- Say we are interested in how schooling affects earning
- Why the regression of schooling on earning won't work?
- What would be the ideal experiment to study this?
- The IV they use: Quarter of Birth (QOB)
- In U.S., students enter school in the calendar year in which they turn 6
- The rule requires students to stay in school until they are 16 years old
- $\therefore$ those born in the 1st quarter can leave school earlier
  - (think about those born in Jan 1st v.s. Dec 31st)
- Assume that QOB not related to earning other than the amount of schooling

Mapping the Notations

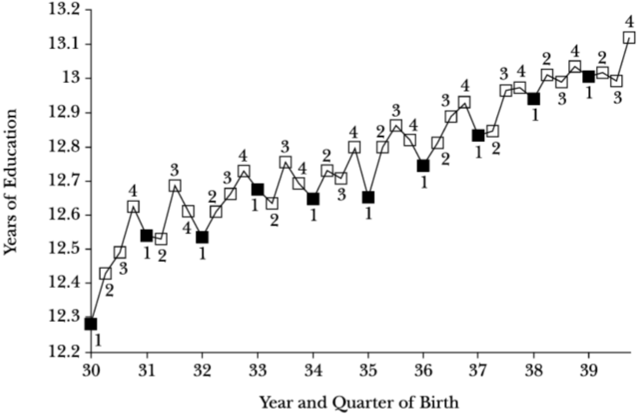
$$D_i = \pi_{10} + \pi_{11} Z_i + \xi_{1i} \text{ (First Stage)}$$
$$Y_i = \pi_{20} + \pi_{21} Z_i + \xi_{2i} \text{ (Reduced Form)}$$

- D_i: years of education
  - or think of whether one goes to college as a binary case
- Y_i: log weekly earning
  - weekly so don't worry about some work more or less
  - why log? percent change interpretation, see HW
- Z_i: whether you are born in the first quarter

Does This Work? First Stage

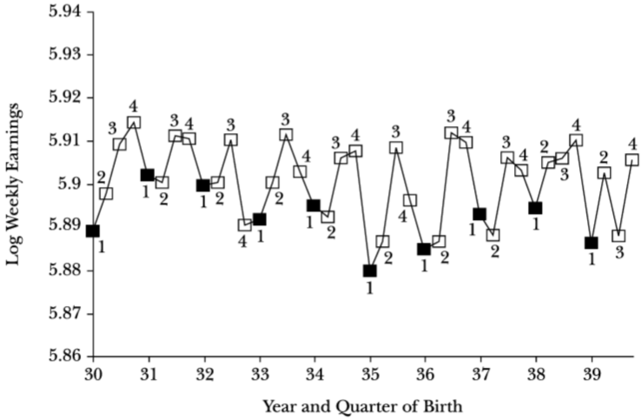
Figure 1

Mean Years of Completed Education, by Quarter of Birth



Does This Work? Reduced Form

Figure 2
Mean Log Weekly Earnings, by Quarter of Birth



Two Stage Least Squared (TSLS or 2SLS)

- Another way to see what happened in math is to substitute the first stage:

$$\begin{aligned} Y_i &= \alpha + \rho D_i + \eta_i \\ &= \alpha + \rho(\pi_{10} + \pi_{11} Z_i + \xi_{1i}) + \eta_i \\ &= \alpha + \rho \underbrace{(\pi_{10} + \pi_{11} Z_i)}_{\equiv \smile_i} + \underbrace{(\rho \xi_{1i} + \eta_i)}_{\equiv \frown_i} \end{aligned}$$

- Note that $E[Z_i \xi_{1i}] = 0$ (By construction, first stage regression)
- Also $E[Z_i \eta_i] = 0$ (IV assumption), so $E[Z_i(\rho \xi_{1i} + \eta_i)] = 0$
- Therefore, $E[\smile_i \frown_i] = 0$
- So if we regress Y_i on $\smile_i$ we get ρ

Two Stage Least Squared (TSLS or 2SLS)

- In practice, we could first get $\hat{\smiley}_i$ by regress D_i on Z_i, then regress Y_i on $\hat{\smiley}_i$

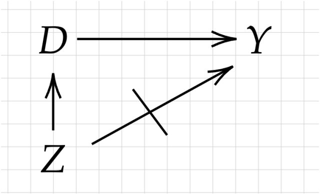
$$
\begin{aligned}
D_i &= \pi_{10} + \pi_{11} Z_i + \xi_{1i} \text{ (First Stage)} \\
\smiley &= \pi_{10} + \pi_{11} Z_i \\
\hat{\smiley}_i &= \hat{\pi_{10}} + \hat{\pi_{11}} Z_i \\
Y_i &= \alpha + \rho \hat{\smiley}_i + (\rho \smiley_i + \frownie_i - \rho \hat{\smiley}_i) \text{ (Second Stage)}
\end{aligned}
$$

- Note that $\hat{\smiley}_i$ is simply the fitted value of the first stage
- So the TSLS is simply get the fitted value first
- and then regress Y_i on the fitted value
- Showing the second stage works is more involved, but essentially $\because$
  - first stage is consistent
  - $E[Z_i(\smiley_i - \hat{\smiley}_i)] = 0$ and $E[Z_i \frownie_i] = 0$

Putting the Rabbit in the Hat

- 2SLS shows how the IV is using the "clean variation" again
- The first stage is a projection of D_i on Z_i
- Can think of taking out the variation in D_i that explains by Z_i
- The fitted value $\hat{☺}_i$, is like when you only use the clean variation to back out D_i
- Since only clean variation is left in $\hat{☺}_i$, regress Y_i on $\hat{☺}_i$ gives the causal effect

In a Diagram



- One common diagram to think about this:
  - Z affects Y only through D

Implementation

- We didn't talk about inference, but a bit similar to OLS
  - Key point: use good packages!
  - The manual 2SLS standard error would be wrong
- It's possible also to include multiple IVs and covariates
  - Make sure you include the same covariates in both stages
  - Or just use the package
- Best practice is to look at first stage and reduced form graphically
- If the relation doesn't show up in reduced form, probably not there
- And then run the regression with the right packages
- Again, don't do it manually

IV Implementations

Table 4.1.1: 2SLS estimates of the economic returns to schooling

	OLS		2SLS					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Years of education	0.075 (0.0004)	0.072 (0.0004)	0.103 (0.024)	0.112 (0.021)	0.106 (0.026)	0.108 (0.019)	0.089 (0.016)	0.061 (0.031)
Covariates:								
Age (in quarters)								✓
Age (in quarters) squared								✓
9 year of birth dummies		✓			✓	✓	✓	✓
50 state of birth dummies		✓			✓	✓	✓	✓
Instruments:			dummy for QOB=1	dummy for QOB=1 or QOB=2	dummy for QOB=1	full set of QOB dummies	full set of QOB dummies int. with year of birth dummies	full set of QOB dummies int. with year of birth dummies

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the the Angrist and Krueger (1991) 1980 Census sample. This sample includes native-born men, born 1930-1939, with positive earnings and non-allocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses.

Some Reflections

- Let's get back to the relevance and exclusion assumption
- $Cov(D_i, Z_i) \neq 0$ saying you need to have "enough" clean variation
  - Can be checked, rule of thumb is F-stat on the IV > 10
  - otherwise we call it a "weak IV"
  - Another view: the first stage is on the denominator, gets trouble if close to 0
- $Cov(\eta_i, Z_i) = 0$ says that the IV cannot affect Y_i other than through D_i
  - Of course IV would affect Y_i, but only through D_i
  - There is no way to test this
  - Most of the time people discuss IV validity on this
  - Can you think of examples that this doesn't hold in the QOB IV?
- The philosophy of IV is not magical, in principle you're finding a quasi-experiment
- And then make the most out of that quasi-experiment part!

IV with Heterogeneous Effect

- So far we assume that the treatment effect is constant across individuals
- What are we estimating if in fact it is heterogeneous?
- Consider the following model

$$Y = \alpha + BD + U$$
$$D = \pi + CZ + V$$

- With proper assumptions, one can show $\beta_{\text{IV}} \equiv \frac{Cov(Y,Z)}{Cov(D,Z)} = E[\frac{C}{E[C]}B]$
- β_{IV} is the weighted average of the causal effect (B)
- But those more strongly impacted by Z (larger C) gets more weight
- Hard to interpret
- In fact, weights could be negative

Recall the Potential Outcome Framework

- Remember our potential outcome framework
- $Y_{i1} = Y_i(1) = Y_i(D_i = 1)$ meaning the potential outcome for i when $D_i = 1$
- Let's think about the binary treatment case and also abstract away from covariates
- We can introduce the IV and define a similar thing
- $D_{i1} = D_i(1) = D_i(Z_i = 1)$ meaning the potential treatment for i when $Z_i = 1$
- Let's the notation ignore i for now
- Y_1, Y_0, D_1, D_0 are the potential outcomes and treatments, respectively
- Note that we have:

$$Y = DY_1 + (1 - D)Y_0$$
$$D = ZD_1 + (1 - Z)D_0$$

Assumptions

1. Relevance: $Cov(D, Z) = E[D|Z = 1] - E[D|Z = 0] \neq 0$
2. Exogeneity: $(Y_1, Y_0, D_1, D_0) \perp Z$
3. Monotonicity (No Defier): $D_1 \geqslant D_0$

Relevance

$$Cov(D, Z) = E[D|Z = 1] - E[D|Z = 0] \neq 0$$

- This is same as before, meaning that the instrument does affect treatment
- The first equation is a result of binary Z and D

Exogeneity

$$(Y_1, Y_0, D_1, D_0) \perp Z$$

- Like the exclusion we talked about
- Z would of course affect D
  - $\because D = ZD_1 + (1 - Z)D_0$
- But the assignment of Z is independent of potential treatments
- A sufficient condition: Z is randomly assigned

Monotonicity

$$D_1 \geqslant D_0$$

- This is new
- Think about 4 cases
  1. $D_0 = 0$ and $D_1 = 0$ (Never Taker)
  2. $D_0 = 1$ and $D_1 = 1$ (Always Taker)
  3. $D_0 = 0$ and $D_1 = 1$ (Complier)
  4. $D_0 = 1$ and $D_1 = 0$ (Defier)
- Compliers are those who only take the treatment if given the instrument
- Monotonicity rules out defiers
- Note that these are potential assignments $\Rightarrow$ not observable

Results

- Result 1:

$$\beta_{IV} \equiv \frac{Cov(Y,Z)}{Cov(D,Z)} = \frac{E[Y|Z=1]-E[Y|Z=0]}{E[D|Z=1]-E[D|Z=0]}$$

- Result 2:

$$\beta_{IV} = E[Y_1 - Y_0|D_1 = 1, D_0 = 0] = \text{LATE}$$

- LATE means local average treatment effect
- The "local" refers to the effects for compliers

Proof: Result 1

Note that for any random variable W and binary Z

$$
\begin{aligned}
Cov(W,Z) &= E[W(Z-E(Z))] \\
&= E[W(Z-E(Z))|Z=1]Pr(Z=1) + E[W(Z-E(Z))|Z=0]Pr(Z=0) \\
&= E(W(1-E(Z))|Z=1)E(Z) - E(WE(Z)|Z=0)(1-E(Z)) \\
&= E(W|Z=1)(1-E(Z))E(Z) - E(W|Z=0)(1-E(Z))E(Z) \\
&= [E(W|Z=1) - E(W|Z=0)]E(Z)(1-E(Z)) \\
\frac{Cov(Y,Z)}{Cov(D,Z)} &= \frac{[E(Y|Z=1) - E(Y|Z=0)]E(Z)(1-E(Z))}{[E(D|Z=1) - E(D|Z=0)]E(Z)(1-E(Z))} \\
&= \frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]}
\end{aligned}
$$

Proof: Result 2

$$\begin{aligned} E[Y|Z=z] &= E[DY_1 + (1-D)Y_0|Z=z] \\ &= E[D_zY_1 + (1-D_z)Y_0|Z=z] \\ &= E[D_zY_1 + (1-D_z)Y_0], \text{(Exogeneity)} \end{aligned}$$

Therefore for any two z and z', we have:

$$\begin{aligned} E[Y|Z=z] - E[Y|Z=z'] &= E[(D_z - D_{z'})Y_1 + (1 - D_z - 1 + D_{z'})Y_0] \\ &= E[(D_z - D_{z'})(Y_1 - Y_0)] \\ &= E[Y_1 - Y_0|D_z = 1, D_{z'} = 0]P(D_z = 1, D_{z'} = 0), \\ &\quad \text{(Law of total probability)} \end{aligned}$$

Similar for D, using the same logic,

$$E[D|Z=z] - E[D|Z=z'] = E[D_z - D_{z'}] = P(D_z = 1, D_{z'} = 0)$$

Taken Together

Now take the ratio, we have:

$$\frac{E[Y|Z=1]-E[Y|Z=0]}{E[D|Z=1]-E[D|Z=0]} = E[Y_1 - Y_0|D_z = 1, D_{z'} = 0]$$

Taken together, we know

$$\beta_{IV} \equiv \frac{Cov(Y,Z)}{Cov(D,Z)} = \frac{E[Y|Z=1]-E[Y|Z=0]}{E[D|Z=1]-E[D|Z=0]} = E[Y_1 - Y_0|D_1 = 1, D_0 = 0] = \text{LATE}$$

Remarks

- What do we learn after all these derivations?
- $\beta_{IV} = \text{LATE} = E[Y_1 - Y_0 | D_1 = 1, D_0 = 0]$
- The IV gives some specific treatment effects
- The treatment effects for the compliers
- Importantly, these treatment effects depend on the IV (Z) you use
- Very often these compliers are not the policy-relevant ones (e.g., earthquake IV)

Covariates

- Things get messy when there are covariates
- For example, instead of assuming $(Y_1, Y_0, D_1, D_0) \perp Z$
- We might want to assume only $(Y_1, Y_0, D_1, D_0) \perp Z|X$
- But in practice you need to saturate the X then
- 2SLS provides an approximation that sometimes works

Who Are the Compliers?

- Compliers are defined with potential treatments, so cannot be identified
- But could still learn something with the LATE assumptions

$$\begin{aligned}
\text{Share of compliers} &= P(D_{1i} > D_{0i}) \\
&= E[D_{1i} - D_{0i}] \\
&= E[D_{1i}] - E[D_{0i}] \\
&= E[D_i | Z_i = 1] - E[D_i | Z_i = 0]
\end{aligned}$$

Who Are the Compliers

- We can also find the characteristic distribution
- Abadie's κ

$$E[X_i|D_{1i} > D_{0i}] = \frac{E[\kappa_i X_i]}{E[\kappa_i]}$$

$$\kappa_i \equiv 1 - \frac{D_i(1 - Z_i)}{1 - P(Z_i = 1|X_i)} - \frac{(1 - D_i)Z_i}{P(Z_i = 1|X_i)}$$

- This can be used to "find the compliers"

Relation to Roy Model

- People find that the Roy model and LATE framework are equivalent
- Model 1: Imbens and Angrist (1994)
  1. $Y = DY_1 + (1-D)Y_0$
  2. $D = ZD_1 + (1-Z)D_0$
  3. (Y_1, Y_0, D_1, D_0)
- Model 2: Non-parametric Roy
  1. $Y = DY_1 + (1-D)Y_0$
  2. $D = \mathbf{1}\{U < \nu(Z)\}$ for ν an unknown function
  3. (Y_1, Y_0, U)

Vytlacil (2002)

- These two are equivalent.
- The key is that $U < \nu(Z)$, it's a separate index.
- For instance, $\nu(U, Z)$ doesn't work.
- This can be used to define MTE:

$$E[Y_1 - Y_0 | U = u]$$

- Those with small u (close to 0) often choose $D = 1$.

Remarks

- We have shown the homogeneous IV and the general IV
- Also on how it relates to the Roy model
- IV is a very powerful tool
- When you have an IV, you have a paper
- But be careful of your IV, are they informative for the problem?
- The fable of finding the keys