Instrumental variables: an epidemiologist’s (or all empirical researchers’) dream

John Rios, Yuanyang Liu
Department of Management Sciences
January 29, 2018

1 Review of causal inference in one slide. Why it matters and why it is difficult.

1.1 A causal definition of ‘Causation’

- $X$ causes $Y (X \rightarrow Y)$ means that: if ($X = x$ implies $Y = y$) and we could intervene and make $X = x'$, then $Y \neq y$. ($X$: treatment, $Y$: outcome.)

- Causation and association: $X, Y$ can be associated because of a common cause $U$ (confounder).
  - E.g. $X =$ education level and $Y =$ income level, confounded by $U =$ one’s innate ability (draw figure).
  - In this case, for a person with high innate ability, she can still be rich with no education.

- Why casual inference is important: we can manipulate outcome (e.g. want to be rich? Get a PHD!)

1.2 Why it is challenging (if not impossible) to infer causation from observational data

- Confounders of the cause and the outcome are unobserved or unknown (e.g. (more education $\rightarrow$ higher income)).
• Because in reality, people choose/select their own level of treatment, which establishes the confounding effect.

• The probability/propensity to be treated is different for different individuals.

• Consider the (education $\rightarrow$ income) example, if people with high innate ability go to college and low innate ability don’t go to college. By comparing income between people with different education level, we are essentially comparing income between people with different innate abilities.

1.3 Why randomized control trial (RCT) is the gold standard for causal inference

• The selection to treatment is completely forbidden. A coin decides who get treatment (imagine a coin decides whether one gets college education or not). So confounding is completely removed.

• Everyone has exactly the same probability/propensity of treatment, by the coin.

• Draw figure for RCT.

1.4 How to mimic RCT with observational study?

• See interlude class notes.
2 Instrumental variables, why it is like a dream

- To infer causal effect from data, we have to face the problem of unmeasured confounder (see Interlude lecture note).

- We learned methods to deal with observed confounders for the treatment assignment. For unobserved confounders, we still have to make the assumption that they don’t exist, or perform sensitive analysis.

- With valid Instrumental Variables (IV), one can estimate causal effect with the presence of unobserved confounders: What I dream everyday.

- However, when I wake up, I find myself still facing the very general ‘no free lunch’ theory of life.

- That is, valid instrumental variables must satisfy some strict conditions, which may only be satisfied in one’s dream.
3 Outline

1. Definition of IV, or conditions of a valid IV.

2. Causal inference with IV.

3. An example: more education causes higher income (Angrist and Keueger, 1991)?

4. An example from Economics.

5. General IV Regression Model.

6. 2SLS in General IV Model

7. Checking instrument validity.

8. Where do valid instruments come from?


10. Main references:

    - [Wooldridge, 2009] (chapter 15) and [Stock and Watson, 2007] (Chapter 12): two classical econometrics textbooks.
    - [Hernán and Robins, 2006] and [Baiocchi et al., 2014]: reference papers on ICON.
    - [Angrist and Keueger, 1991]: one of the most famous IV study, education affects individual income.
4 Causal inference with regression

• A population regression model for the outcome

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]  \hspace{1cm} (1)

– \( Y_i \): outcome of interest (e.g. individual income).
– \( X_i \): treatment variable of interest (e.g. individual year of education).
– \( u_i \): regression error, includes all omitted factors that determine Y (e.g. confounders: family wealth, innate ability).
– \( \beta_1 \): population parameter representing the causal effect of \( X \) to \( Y \) (assuming no reverse causality).

• In a simulation study, we may use Equation (1) to generate data, where \( u_i \) may include other terms, e.g. \( u = \beta_2 X_2 + \beta_3 X_3 + N(0,1) \) and \( \text{cor}(X, X_2) \neq 0 \).

• However, to estimate \( \beta_1 \), we only observe two columns of data \( Y \) and \( X \)

• The problem to get consistent estimate of \( \beta_1 \): correlation between \( X \) and \( u \), from the following sources:

  – Uncontrolled confounders, omitted variable bias.
  – Measurement errors (non-compliance to the treatment).
  – Reverse causality (higher income causes more education).
  – Whatever the source of correlation, the instrumental variables method solves it all.
The idea of the IV method

- The problem: correlation between treatment X and u.

- One idea: find a part of X that is *not correlated* with u, and use this part of X for causal inference (draw a picture to show the part of X uncorrelated to U.).

- Note that in a random experiment, X is completely uncorrelated with U.

- Instrumental variable(s) is used to identify such a part of X not correlated with u.

- Do you like this idea?
6 Definition of an IV / conditions of a valid IV

- Given a population linear model of $Y$:

$$Y_i = \beta_0 + \beta_1 X + u_i$$  \hspace{1cm} (2)

- A valid instrument $Z$ of the treatment variable $X$ satisfies two conditions:

  - **Relevance**: $\text{cor}(Z, X) \neq 0$.
  
  - **Exogeneity**: $\text{cor}(Z, u) = 0$.

  - By exogeneity, the instrument $Z$ has nothing to do with the data generating process of $Y$ that is not controlled by the model.

  - By relevance and exogeneity, the instrument $Z$ is related to $Y$, but only through the treatment $X$.

  - See figure.
7 Random experiment as an example of IV

• Let Z be defined as the outcome of a fair coin flip (Head or Tail)
• For each individual in a random experiment: X=1 if Z=head and X=0 if Z=tail.
• In this case, Z is a valid instrument for the treatment X:
  − Z is relevant: $\text{cor}(Z, X) = 1$
  − Z is exogenous: $\text{cor}(Z, u) = 0$
  − See figure.

• In an observational study, a valid instrument can be weak: $\text{cor}(Z, X)$ is small. Then estimated causal effect by IV can be problematic. Analogous to a random experiment with non-compliance problem.
8 Natural experiment as an example of IV

- Treatment: class size
- Outcome: test score
- Question: small class size causes higher test score?
- Unobserved confounders: e.g. learning opportunity outside of class; quality of teacher.
- Natural event: earthquake. Classes are combined, so class size is large, in area close to the earthquake center.
  - IV: distance between school and earthquake center.
  - See figure.
9 The Two Stage Least Squares (2SLS) Estimator

- $Y$: outcome, $X$: treatment, $Z$: instrument variable. No control variables included here, so the math is simple enough to deliver the main properties of IV estimator clearly.
- Assume for now we already find a valid instrument $Z$.
- A population linear model of $Y$:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$  \hfill (3)

- Recall two assumptions on $Z$
  1. Relavence: $\text{cor}(X, Z) \neq 0$.
  2. Exogenous: $\text{cor}(u, Z) = 0$.
- 2SLS estimator
  - Stage 1 model:

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$  \hfill (4)

  * $X_i$ is decomposed into two parts: $\pi_0 + \pi_1 Z_i$ and $v_i$.
  * By Relavence assumption: $\pi_1 \neq 0$.
  * By Exogenous assumption: $(\pi_0 + \pi_1 Z_i)$ is uncorrelated with $u_i$. It is the part of $X$ used to derive causal relationship between $X$ and $Y$.
  - Stage 2 model:

$$Y_i = \beta_{0}^{2SLS} + \beta_{1}^{2SLS} \hat{X}_i + u_i^{2SLS}$$  \hfill (5)

  * $\hat{X}_i = \pi_0 + \pi_1 Z_i$, from Equation 4
  * $\hat{\beta}_1^{2SLS}$ is the 2 Stage Least Squares estimator of the causal effect of $X$ to $Y$. 
* $\hat{\beta}^{2SLS}_1$ is a consistent estimator of the correct causal effect $\beta_1$ in Equation 3.

- A note: it may be tempting to implement 2SLS estimator by hand, by regressing Equation 4 and Equation 5. However, if we plug Equation 4 into Equation 3

$$u^{2SLS}_i = u_i + \beta_1 v_i.$$  

(6)

So it is better to use a IV software package, which takes this into account and get correct standard errors of $\beta^{2SLS}_1$. 
10 Deriving the formula for 2SLS estimator: a single treatment and a single instrument (can skip).

- Stage 1: \( X_i = \pi_0 + \pi_1 Z_i + v_i \)
- Stage 2: \( Y_i = \beta_{0SLS} + \beta_{1SLS} \hat{X}_i + u_{iSLS} \)
- From Stage 2: \( \hat{\beta}_{1SLS} = \frac{Cov(Y, \hat{X})}{Var(\hat{X})} \)
- By definition of \( \hat{X} \) from Stage 1, \( Cov(Y, \hat{X}) = Cov(Y, (\pi_0 + \pi_1 Z)) = \pi_1 Cov(Y, Z) \)
- Similarly, \( Var(\hat{X}) = Var(\pi_0 + \pi_1 Z) = (\pi_1)^2 Var(Z) \)
- Also from Stage 1, \( \hat{\pi}_1 = \frac{Cov(X, Z)}{Var(Z)} \)
- Finally, put all terms together:

\[
\hat{\beta}_{1SLS} = \frac{Cov(Y, \hat{X})}{Var(\hat{X})} = \frac{\pi_1 Cov(Y, Z)}{(\pi_1)^2 Var(Z)} = \frac{Cov(Y, Z)}{\pi_1 Var(Z)} = \frac{Cov(Y, Z)}{Var(Z)} \times \frac{Var(Z)}{Cov(X, Z)}
\]

i.e. \( \hat{\beta}_{1SLS} = \frac{Cov(Y, Z)}{Cov(X, Z)} \) (Jump to here.)

Note here Cov() represents sample covariance. Also note its similarity with the OLS estimator \( \hat{\beta}_{1OLS} = \frac{Cov(Y, X)}{Cov(X, X)} \).

- A nonformal proof of consistency of \( \hat{\beta}_{1SLS} \):
  
  - Recall the linear model of outcome \( Y_i \):

\[
Y_i = \beta_0 + \beta_1 X_i + u_i
\]
The population covariance between $Y_i$ and $Z_i$:

$$\text{covariance}(Y_i, Z_i) = \text{covariance}(\beta_0 + \beta_1 X_i + u_i, Z_i)$$

$$= \beta_1 \text{covariance}(X_i, Z_i) + \text{covariance}(u_i, Z_i)$$

$$= \beta_1 \text{covariance}(X_i, Z_i) \quad \text{(By exogeneity, } \text{covariance}(u_i, Z_i) = 0.)$$

i.e. $\beta_1 = \frac{\text{covariance}(Y_i, Z_i)}{\text{covariance}(X_i, Z_i)} \quad \text{(By relevance, } \text{covariance}(X_i, Z_i) \neq 0.)$

Note here: This is how we get OLS estimator of $\beta_1$, by replacing $Z_i$ with $X_i$ here, but with the assumption of $\text{covariance}(X, u) = 0$.

As sample size goes to infinity:

$$\hat{\beta}_1^{2SLS} = \frac{\text{sampleCov}(Y, Z)}{\text{sampleCov}(X, Z)} \quad \rightarrow \quad \frac{\text{covariance}(Y_i, Z_i)}{\text{covariance}(X_i, Z_i)} = \beta_1$$

13
11 Large-sample distribution of the 2SLS estimator (can skip)

- Again, let us start from the population regression model for $Y_i$:
  \[ Y_i = \beta_0 + \beta_1 X_i + u_i \]  

- From Equation (8) $Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + (u_i - \bar{u})$. Accordingly, the sample covariance between $Y$ and instrument $Z$, $Cov(Y, Z)$ is

\[
Cov(Y, Z) = \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \bar{Z})(Y_i - \bar{Y})
\]

\[
= \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \bar{Z})[\beta_1 (X_i - \bar{X}) + (u_i - \bar{u})]
\]

\[
= \beta_1 Cov(Z, X) + \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \bar{Z})(u_i - \bar{u})
\]

\[
= \beta_1 Cov(Z, X) + \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \bar{Z})u_i
\]

(since $\bar{u}$ is a constant, and $\sum_{i=1}^{n} (Z_i - \bar{Z}) = 0$)

\[
\therefore \hat{\beta}_{1SLS} = \frac{Cov(Y, Z)}{Cov(X, Z)} = \beta_1 + \frac{\frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \bar{Z})u_i}{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Z_i - \bar{Z})}
\]

- Again, this is very similar to the OLS estimator of $\beta_1$:

\[
\hat{\beta}_{1OLS} = \frac{Cov(Y, X)}{Cov(X, X)} = \beta_1 + \frac{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})u_i}{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})}
\]
• To see the consistency of \( \hat{\beta}_{1}^{2SLS} \) from the above equation, when \( n \) goes to infinity, 
\[
\frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \bar{Z})u_i
\]
converges to population covariance between \( Z \) and \( u \), which equals 0 by exogenous assumption of the instrument \( Z \).

• Also, from the above equation, \( \hat{\beta}_{1}^{2SLS} \approx \beta_1 + \frac{1}{n} \sum_{i=1}^{n} (Z_i - \bar{Z})u_i \)

\[
\text{covariance}(X_i, Z_i)
\]

• So, in large samples, \( \hat{\beta}_{1}^{2SLS} \) is approximately normal

\[
\hat{\beta}_{1}^{2SLS} \sim N(\beta_1, \sigma_{1}^{2\hat{\beta}_{1}^{2SLS}}),
\]

where \( \sigma_{1}^{2\hat{\beta}_{1}^{2SLS}} = \frac{1}{n} \text{variance}[(Z_i - \mu_Z)u_i] \text{covariance}(Z_i, X_i)^2 \). (Discuss this!) (9)

Standard statistical tests and confidence interval can be constructed according to this approximated distribution.
11.1 Must discuss: Problems when the instrument is weak, even it is valid

- From Equation 9, $\hat{\beta}^{2SLS}_1$ can have large standard errors, if $\text{cor}(X, Z)$ is small (weak instrument $Z$).

- If instrument $Z$ is irrelevant, $\text{cor}(Z, X) = 0$. This is equivalent to have a random experiment but everyone comply with probability 0.5. In this case, everything breaks: $\hat{\beta}^{2SLS}_1$ is approximately the ratio of two normal distributions, centered at the limit of $\hat{\beta}^{OLS}_1$.

- Even when $Z$ is weak, the above problem still exist. How strong do we need an valid IV to be? Ask John!

- More seriously, weak and non-exogenous instrument can lead to large asymptotic bias.

\[
\text{plim } \hat{\beta}^{2SLS}_1 = \beta_1 + \frac{\text{cor}(u_i, Z_i)}{\text{cor}(X_i, Z_i)} \frac{\sigma_{u_i}}{\sigma_{X_i}}
\]  

(10)

\[
\text{plim } \hat{\beta}^{OLS}_1 = \beta_1 + \frac{\text{cor}(X_i, u_i)}{\sigma_{X_i}} \frac{\sigma_{u_i}}{\sigma_{X_i}}
\]  

(11)

- (Can skip.) From Equation 10 and Equation 11

  - Inconsistency (asymptotic bias) of $\hat{\beta}^{2SLS}_1$ can be large, if $\text{cor}(X, Z)$ is small and $\text{cor}(u, Z) \neq 0$ (even very small).

  - [Wooldridge, 2009] states that if focus only on consistency, IV estimator is not necessarily better than OLS, by looking at the second term of the RHS of the two equations. For example, if $\text{cor}(X_i, Z_i)$, then even for small $\text{cor}(u_i, Z_i)$, it is possible

\[
\frac{\text{cor}(u_i, Z_i)}{\text{cor}(X_i, Z_i)} > \text{cor}(X_i, u_i) > 0,
\]  

(12)
then IV estimator is biased more upward than OLS.

– Unfortunately, since we may typically have little idea about $\text{cor}(u_i, Z_i)$ and $\text{cor}(X_i, u_i)$, we may never know for sure which estimator has less bias (unless we assume $\text{cor}(u_i, Z_i) = 0$).
12 An example IV study [Angrist and Keueger, 1991]

- So, if we can find a valid and not weak instrument for the treatment variable of interest, we can perform statistical inferences for the treatment effect without considering the unobserved confounders: a dream come true!

- How to find a valid and not weak instrument? Ask John later!

- An example study with IV: [Angrist and Keueger, 1991] (AK-91).

- Treatment: number of years in education (educ).

- Outcome: wage (wage).

- Unobserved confounders: e.g. innate ability

- Possible IV: mother/father’s years in education; number of siblings.

- IV in AK-91: whether or not a man is born in the first quarter of a year (frstqrt).
  - **Exogeneity**: it can be reasonable to assume when you are born is uncorrelated with factors that determines your wage, $\text{Cor}(\text{frstqrt}_i, \text{wage}_i) = 0$.
  - **Relevance**: $\text{Cor}(\text{frstqrt}_i, \text{educ}_i) \neq 0$? Yes, as a result of the school start age policy and compulsory schooling rules in the US.
  - Specifically, US law requires that a child go to school in the calender year when he turns 6 (start age polity).
  - US law also requires the earliest one can drop school is his 16th birthday.
  - For example, assume my birthday is 1/1 and John’s birthday is 12/31, school start date is 6/1. Then, I start school at age 6.5, and John start school at age 5.5. If both of us drop school at age 16, I get 9.5 years of eduction, and John get 10.5 years of education.
Figure 1: Figure 1 in AK-91

- Figure 1 shows the plot of frstqrt and educ. They are correlated, but the correlation is small.

- With sample size of 329,509 US men born in 1930-1939, the OLS estimate of effect of educ to wage is 0.0711 (0.0003); the 2SLS estimate is 0.0891 (0.0161). Note how large the se is for 2SLS estimator, so the CI from 2SLS is much wider than the CI of OLS estimator.

- AK-91 also report estimates of US men born in other decades and with

---

1 See also chapter 4 of ‘Mostly harmless econometrics’ [Angrist and Pischke, 2008].
different model specifications. Consistently, OLS and 2SLS estimates of effect of education to wage is similar. This may be because there is no unobserved confounder in the OLS estimates.

- [Bound et al., 1995] argue for correlation between quarter of birth and unobserved determinants of one’s income.

* quarter of birth is related to: (1) school attendance rate, (2) the likelihood of a student being addressed with behavioral difficulties, (3) the likelihood of a student being referred to mental health services and (4) a student’s academic performance, all of which can affect a student’s future income.

* There is substantial evidence that individuals born early in the year are more likely to suffer from schizophrenia.

* There is also evidence of variation by quarter of birth in the incidence of mental retardation (Knoblock and Pasamanick 1958), autism (Gillberg 1990), dyslexia (Livingston, Adam, and Bracha 1993), multiple sclerosis (Templer et al. 1991), and manic depression (Hare 1975), as well as somewhat mixed evidence regarding differences in IQ among children born at different times of the year (Whorton and Kames 1981).

* The family income of those born early in the year tends to be lower than the family income of those born later in the year.

- Conclusion, AK-91 has weak instrument ($\text{cor}(\text{educ, firstqrt})$ is small), and [Bound et al., 1995] show that instrument in AK-91 is probably not exogenous ($\text{cor}(Z, u) \neq 0$). Therefore, the IV estimator can be problematic.
13 An example IV study in Economics

- Some commodities like cigarettes figure prominently in public policy debates.
- Attempting to reduce illness and deaths from smoking, a tax on cigarettes is being analyzed.
- How big a tax hike is needed to make a dent on cigarette consumption?
- What would the after-tax sales price of cigarette need to be to archive a 20% reduction in cigarette consumption? Elasticity of demand for cigarettes
- Estimate it from data on prices and sales (OLS regression of log quantity on log price). Interaction between supply and demand.

Use 2SLS. Annual data for the 48 continental US states for 1985-1995

- Z: Sales tax. Amount of money associated with the sales taxes in pack of cigarettes.
- Y: Cigarettes consumption. Number of pack of cigarettes sold per capita in the state
- X: Average real price per pack of cigarettes including all taxes.
- High sales tax increases the total sales price. Instrument relevance.
- Sales tax affect the demand for cigarettes only through the price. Instrument exogeneity
- However, there might still be omitted variables that are correlated with the sales tax per pack. Income?
14 General IV Regression Model

The General IV Regression Model has for types of variables:

- **Y**: dependent variable
- **X**: problematic endogenous regressor
- **W**: exogenous regressor
- **Z**: instrumental variable

For IV regression to be possible, there must be at least as many instrumental variables (Zs) as endogenous regressors (Xs). Coefficients may be Underidentified, Exactly Identified, and Overidentified. Only in the last 2 cases the coefficients can be estimated by IV regression.

15 2SLS in General IV Model

The 2SLS estimator in the general IV model with multiple instrumental variables is computed in 2 stages:

- **First-stage regression(s)**: Regress $X_{1i}$ on the instrumental variables $(Z_{1i}, \ldots, Z_{mi})$ and the exogenous variables $(W_{1i}, \ldots, W_{ri})$ using OLS. Compute the predicted value from this regression; call these $\hat{X}_{1i}$. Repeat this for all the endogenous regressors $X_{1i}, \ldots, X_{ki}$, thereby computing the predicted values $\hat{X}_{1i}, \ldots, \hat{X}_{ki}$.

- **Second-stage regression**: Regress $Y_i$ on the predicted value of the endogenous variables $(\hat{X}_{1i}, \ldots, \hat{X}_{ki})$ and the included exogenous variables $(W_{1i}, \ldots, W_{ri})$ using OLS. The 2SLS are the $\hat{\beta}^{2SLS}_0, \ldots, \hat{\beta}^{2SLS}_{(k+r)}$ of the second regression.
16 Checking Instrument Validity

16.1 Assumption 1: Instrument Exogeneity

- Potential violations of this assumption implies unmeasured confounding for the instrument.

According to the diagram:

- Recall:

$$\hat{\beta}_{1 \text{SLS}} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} = \beta_1 + \frac{\frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \bar{Z})u_i}{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Z_i - \bar{Z})}$$

- The 2SLS estimator will converge in probability to something other than population coefficient.
- The 2SLS will be biased and inconsistent
- Can you test statistically the assumption that the instruments are exogenous? Yes (if the coefficients overidentified) and No (if the coefficients are exactly identified)
- If the coefficients are exactly identified, rely on expert opinion
- If the coefficients are overidentified, use the J-statistic.

$$\hat{u}_{i \text{SLS}} = \delta_0 + \delta_1 Z_1 i + ... + \delta_m Z_m i + \delta(m+1)W_1 i + ... + \delta(m+r)W_r i + e_i \quad (13)$$
16.2 Assumption 2: Instrument Relevance

- Instrument relevance plays a role similar to the sample size.
- The more relevant the instruments (variation explained by the instruments) the more information is available for use in IV regression (more accurate estimators).
- 2SLS estimator has a Normal sampling distribution (CLT–large samples)
- The more relevant is the instrument, the better is the normal approximation of the 2SLS estimator
- Verifiable. Statistical analysis to check if instrument and treatment are associated

Weak Instruments

- Instruments that explain little of the variation in X are called weak instrument.
- If instruments are weak, 2SLS is no longer reliable even in large samples.
- Normal distribution provides a poor estimation to the sampling distribution of the TSLS estimator.
- The actual confidence of 2SLS estimator ± 1.96 standard deviation can be far less than 95
- Recall:

\[
\hat{\beta}_{1}^{2SLS} = \frac{\text{sampleCov}(Y, Z)}{\text{sampleCov}(X, Z)} \rightarrow \frac{\text{covariance}(Y_i, Z_i)}{\text{covariance}(X_i, Z_i)} = \beta_1
\]

if the instrument is not just weak but irrelevant, the denominator in the RHS is zero.

- 2SLS estimator can be badly biased. Even more than the unadjusted estimate.
• How relevant must the instrument be for the Normal distribution to provide a good approximation in practice? A rule of thumb.

• Compute the F-statistic testing the hypothesis that all coefficients are zero in the first stage. The more information the instruments contain, the larger is the expected value of the F-statistic. You do not need to worry about weak instruments if the F-statistic exceeds 10.
17 Where do Valid Instruments Come From?

In practice the most difficult aspect of IV estimation is finding instruments that are both relevant and exogenous. Two main approaches:

- Domain knowledge to suggest instruments. For example, Sales tax would affect the total sales price but it is expected to affect the demand of cigarettes only through the total sales price, being independent of unmeasured confounders.

- Look for some exogenous source of variation on X arising from what is in effect a random phenomenon that induces shifts in the endogenous regressor. For example, earthquake damage increased average class size in some school districts, and this variation in class size was unrelated to potential omitted variables that affect student achievements.
18 Summary

- Instrumental variables regression is a way to estimate regression coefficients when one or more regressor is correlated with the error term.

- Endogenous variables are correlated with the error term in the equation of interest; exogenous variables are uncorrelated with this error term.

- For an instrument to be valid, it must (1) be correlated with the included endogenous variable and (2) be exogenous.

- IV regression requires at least as many instruments as included endogenous variables.

- The 2SLS estimator has two stages. First, the included endogenous variables are regressed against the included exogenous variables and the instruments. Second, the dependent variable is regressed against the included exogenous variable and the predicted values of the included endogenous variables from the first-stage regression(s).

- Weak instruments (instruments that are nearly uncorrelated with the included endogenous variables) make the 2SLS estimator biased and 2SLS confidence intervals and hypothesis test unreliable.

- If an instrument is not exogenous, then the 2SLS estimator is inconsistent.
References


