UCLA Anderson
SCHOOL of MANAGEMENT

PMDA

Lecture 7: Instrumental Variables (IV)

A few updates and reminders

Classes:

- **Final Exam:** Thursday, March 21st 8:30 to 11:30 am, classroom:
  - Section 1: C301
  - Section 2: B313

Problem sets:

- **Problem Set 3**: Panel Data & Diff-in-diff **due Feb 25th**
- Problem Set 4: Lasso Regression due March 16th

Last Class: Diff-in-diffs estimator

- A group of people (treated group) was assigned a treatment (saw the ad) at a time t^Treat, but the treatment was **not** randomly assigned
- You have panel data for the outcome of interest (average revenue, CTR) for the treated group before and after t^Treat, let's call them $Y^{treated}_{before}$ and $Y^{treated}_{after}$
- If
  - You also have panel data for a group of people who were not exposed to the treatment (control group), $Y^{control}_{before}$ and $Y^{control}_{after}$
  - The outcome in the treated and control groups would have the same time variation if it weren't for the treatment ("parallel trends" assumption)
- Then you can estimate the treatment effect on observational data using the diff-in-diff estimator:

$$(Y^{treated}_{after} - Y^{control}_{after}) - (Y^{treated}_{before} - Y^{control}_{before})$$

Same as

$$(Y^{treated}_{after} - Y^{treated}_{before}) - (Y^{control}_{after} - Y^{control}_{before})$$

Last Class: Checking for common trends

The key identifying assumption in DD models is that the treatment states have similar trends to the control states in the absence of treatment.

- With only one treatment and control group, graph your results, and look at trends in periods before the treatment
- More generally,
    - We can use “leads” and “lags”
    - Run a regression with more interaction terms of treatment and time, for times before and after the treatment happened
    - The leads are the coefficients for the interaction terms before
    - The lags are the coefficients for the interaction terms after
    - Under the assumption of common trends the lead coefficients should be statistically insignificant

$$Y_{it} = \delta_i + \lambda_t + \beta_{-2} D_{it,-2} + \beta_{-1} D_{it,-1} + \beta_1 D_{it,1} + \beta_2 D_{it,2} + \beta_3 D_{it,3} + e_{it}$$

- Under the assumption of common trends beta_{-2} and beta_{-1} should be statistically insignificant (Autor 2003)
- This also tells you whether the effects stays constant, fades or increases over time

Agenda today

- **Another method for causal inference in observational data:**
  - Instrumental Variables (IV) regression
  - Workshop

Instrumental Variables (IV) regression

- IV is a general method to eliminate endogeneity, i.e. when you have a regression model

$$Y = \beta_0 + \beta_1 X + e$$

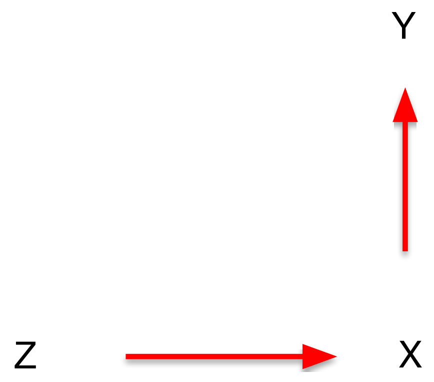
  with cov(X,e)≠0

- The key requirement is that you must find (at least) another variable, an instrument Z, that satisfies two conditions:

1. (Relevance) it is correlated with X: cov(Z,X)≠0

2. (Exogeneity) it is not correlated with e: cov(Z,e)=0

IV in a causal graph



- Visualization of the two conditions for instrument validity:
  - Relevance = there is a link between Z and X
  - Exogeneity = there is no direct link between Z and Y

- **Q:** How is this different from the causal graphs with omitted variables we have considered before?
- **A:** An omitted variable by definition is part of the error e, so it must have a direct link to Y

The two-stage least squares (TSLS) estimator

$$Y = \beta_0 + \beta_1 X + e$$

- Idea: use Z to isolate the variation in X that is not correlated with e (i.e., the exogenous variation in X) and use it to estimate β_1

- TSLS procedure. Run two regressions:
  1. First-stage regression of X on Z

$$X = a_0 + a_1 Z + v$$

Exogenous part is the predicted value, $\widehat{X} = \widehat{a_0} + \widehat{a_1} Z$

2. Second-stage regression of Y on predicted value

$$Y = \beta_0 + \beta_1 \widehat{X} + e$$

Because $\widehat{X}$ is uncorrelated with e (why?), the estimator of β_1 is unbiased

Not to be confused with the Frisch-Waugh theorem

The example we saw of the Frisch-Waugh theorem:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + e$$

1. First regress X on Z

$$X = a_0 + a_1 Z + \widetilde{X}$$

2. Then regress Y on the residual $\widetilde{X}$

$$Y = \beta_0 + \beta_1 \widetilde{X} + b$$

Q: What are the differences?

A:

Frisch-Waugh

- Z is an omitted variable that causes endogeneity if not included
- Z has a direct effect on Y
- It is the residual from the stage 1 regression that eliminates the endogeneity

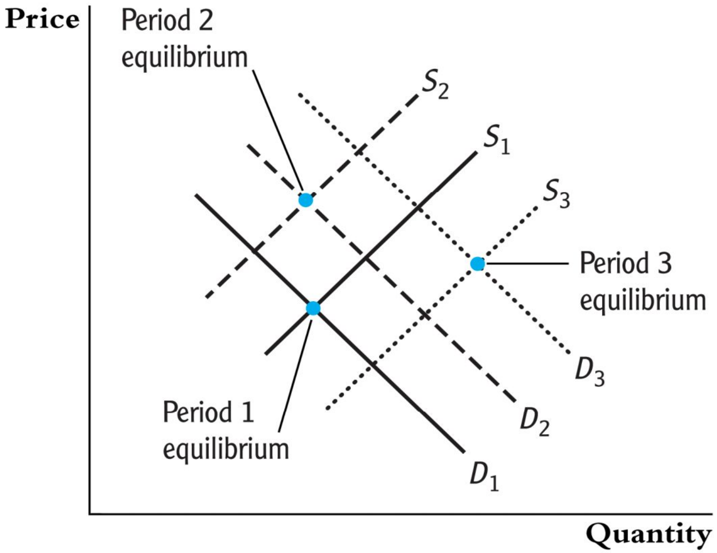
TSLS

- Z is exogenous
- Z has a no direct effect on Y
- It is the predicted value from the stage 1 regression that eliminates the endogeneity

Example 1. Demand estimation

$$\log(Q) = \beta_0 + \beta_1 \log(P) + e$$

- β_1 is demand elasticity = % change in quantity for a 1% change in price (this is the interpretation in a log-log model)
- Data: you observe prices and quantities in different regions or time periods
- We saw that this is an example of simultaneous causality bias: there is a supply curve in the background so P and Q are determined in equilibrium

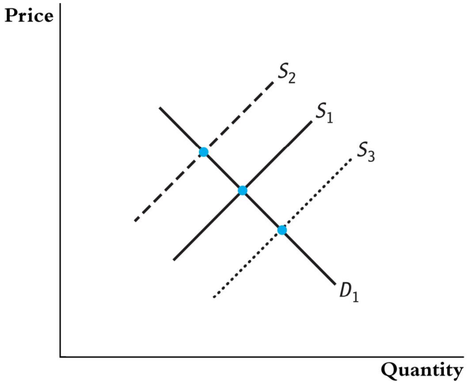


(a) Demand and supply in three time periods

Example 1. Demand estimation

$$\log(Q) = \beta_0 + \beta_1 \log(P) + e$$

- Fitting a regression through the scatterplot gives a biased estimate of elasticity
- But, what if only the supply curve shifted and the demand stayed constant?



(c) Equilibrium price and quantity when only the supply curve shifts

- TSLS estimates the demand curve by isolating shifts in price and quantity that arise from shifts in supply
- Z is a variable that shifts supply but not demand

Instruments for demand estimation

Examples of instruments for the regression

$$\log(Q) = \beta_0 + \beta_1 \log(P) + e$$

1. Suppose the demand is for milk and use Z = rain in dairy-producing regions
   - **Q:** Relevant?

     A: Less rain → less milk produced → higher price
   - **Q:** Exogenous?

     A: Rain doesn’t directly affect how much milk you buy
2. Suppose the demand is for cigarettes and use Z = sales tax in different US states
   - **Q:** Relevant?

     A: Sales tax affects all prices so also cigarette price
   - **Q:** Exogenous?

     A: Sales tax not determined by cigarette market

Another Example

Does putting criminals in jail reduce crime? (Levitt 1996)

- You have data of crime rates and incarceration rates at the state level
- You want to measure how crime rates are affected by incarceration rates
- You also have data on economic conditions, age, and other demographics

What type of bias can we have?

A: Simultaneous bias:

more crime → more prisoners (if police do their job)

more prisoners → less crime

What instrument would you use?

A: "Lawsuits aimed at reducing prison overcrowding"

- Relevance: They slow the growth of prisoner incarcerations
- Exogeneity: Only if overcrowding litigation is induced by prison conditions but not by crime rate or its determinants

Instruments can isolate “as if random” variation

- Say we want to understand the effect of class size on students’ performance

  test_scores= β_0+β_1 class_size + e

  - **Q:** Endogeneity problem in this regression?
  - **A:** OVB, for example due to parental income/involvement
- Suppose you have a natural experiment: an earthquake makes affected districts to double up classrooms in some schools (others have to close)
- Let Z = distance to the epicenter
  - **Q:** Is this a valid instrument?
  - **A:** Relevant because class size is correlated with the distance to the epicenter: districts close to epicenter are most severely affected
  - Exogenous if distance to epicenter is unrelated to any other factors affecting student performance such as being english learner or other disruptive effects of the earthquake on performance
  - **Q:** Can you explain in words how TSLS gives you a causal estimate?
  - **A:** The first stage regression isolates the variation in class size that is “as if randomly assigned”. Using this part in the second stage regression thus gives a causal estimate

Discussion Question

Suppose you wish to measure the impact of smoking on the weight of newborns.

You are planning to use the following model,

$$\log(bw_i) = \beta_0 + \beta_1 male_i + \beta_2 order_i + \beta_3 y_i + \beta_4 cig_i + e_i$$

where

- **bw** is the birth weight,
- **male** is a dummy variable equal to 1 if the baby is a boy or 0 otherwise,
- **order** is the birth order of the child,
- **y** is the log income of the family,
- **cig** is the amount of cigarettes per day smoked during pregnancy,
- **i** indexes the observation
- **β's** are the unknown parameters.

Q1: What could be the problem in using OLS to estimate the above model?

Q2: Suppose you have data on the average price of cigarettes in the state of residence. Would this information help you to identify the true parameters of the model?

Generalizations

- We can have more than one instrument: Z_1,…,Z_m

- We can also add control variables: W_1,…,W_r

1. First-stage regression: regress X on Z_1,…,Z_m and W_1,…,W_r and get predicted value $\widehat{X}$

2. Second-stage regression: regress Y on $\widehat{X}$ and W_1,…,W_r
    - Note: the standard errors from this regression are not the correct ones because they do not account for estimation uncertainty in $\widehat{X}$
    - Traditional softwares/packages account for this when computing standard errors

Generalizations

- We can also have more endogenous regressor: X_1,…,X_k

- In this case, we need that the model is exactly identified or over identified:
  - Exactly identified: when $m=k$, that is, same number of instruments and endogenous regressors
  - Overidentified: when $m>k$
  - Underidentified: when $m<k$

- In the latter case, we cannot estimate all the model parameters

Can you test for instrument validity?

$$Y = \beta_0 + \beta_1 X + e$$

- Requirements for validity when you have multiple instruments Z_1,…,Z_m

  1. Relevance: at least one instrument is correlated with X (a bit more involved when we have multiple endogenous regressors, see Stock & Watson Chapter 12)

  2. Exogeneity: all the instruments are uncorrelated with e

- Can you test these conditions?
  - You can test for relevance by looking at the first-stage regression
  - You can test for exogeneity but only when you have more than one instrument

Testing relevance

$$Y = \beta_0 + \beta_1 X + e$$

- Relevance is akin to sample size: the more relevant the instruments, the more variation in X is explained by the instruments, so that we have more info for the IV regression => More relevant implies more accurate estimator

- First stage regression:

$$X = a_0 + a_1 Z_1 + \cdots + a_m Z_m + v$$

  - The instruments are said to be weak if all the a_1,…, a_m are zero or nearly zero
  - Weak instruments explain little of the variation of X:
    - E.g., something that shifts the supply curve but only by little
  - Weak instruments imply that the TSLS estimator is biased towards the OLS estimator, and usual statistical inference (standard errors, hypothesis tests) is misleading

Testing relevance

$$X = a_0 + a_1 Z_1 + \cdots + a_m Z_m + v$$

- You can test for weak instruments by:
  - Consider the F-statistic testing the hypothesis that a_1,…, a_m are all zero
  - Rule of thumb: if first-stage F-statistic < 10, you have weak instruments
  - Note that simply rejecting the null hypothesis that the coefficients are zero isn’t enough – you need the F-statistic to be large (> 10)
  - There are more formal tests for weak instruments (e.g., built-in in python and R)

Testing relevance

What if you have weak instruments?

- If you have many: a few strong and many weak instruments, drop some of the weak until you get F-stat >10

- Your standard error may increase but keep in mind that the original standard errors were not meaningful

- If the coefficients are exactly identified and you don’t have strong enough instruments:
  - Try to find additional stronger instruments
  - Proceed with weak instruments but use other methods (e.g., Limited information maximum likelihood estimator, LIML)

Testing exogeneity

- You can test for exogeneity only if you have more than one instrument
- The test is called the “J-test of overidentifying restrictions”: it tests that all instruments are exogenous (H_0)
- Intuition:
    - Suppose there are two instruments → could compute two separate TSLS estimates → if estimates are very different something must be wrong: one of the two instruments (or both) must be invalid
    - The J-test makes this comparison in a statistically precise way

Testing exogeneity

More formally:

1. Build:

$$\hat{u}_i^{IV} = Y_i - (\hat{\beta}_0^{IV} + \hat{\beta}_1^{IV} X_{1i} + \cdots + \hat{\beta}_{k+r}^{IV} W_{ri})$$

2. Estimate:

$$\hat{u}_i^{IV} = \delta_0 + \delta_1 Z_{1i} + \cdots + \delta_m Z_{mi} + \delta_{m+1} W_{1i} + \cdots + \delta_{m+r} W_{ri} + e_i$$

3. Test whether the delta’s are zero by computing the F-statistic for: delta_1=...=delta_m=0

Why can’t we do the same approach when m=k (same instruments as endogenous regressors)?

Note: If the J-test rejects, you don’t know which instruments are endogenous. In this case you need to use your knowledge of the problem to decide what is the most appropriate next step (see workshop)

IV regression in Python

- The TSLS estimator is computed in Python using the command *IV2SLS* from linearmodels.iv: “from linearmodels.iv import IV2SLS”
- Syntax is *IV2SLS* (*Y,[const,W], X, Z*) where *X* is the endogenous variable, *Z* the instrument(s) and *W* the control(s)
- Diagnostic tests can be run by
  - iv_reg.first_stage: Reports statistics from the first stage include the partial F-statistic which tests for relevance (weakness of instruments)
  - iv_reg.sargan: This is the J-test of instrument exogeneity. It can only be used if you have more instruments than endogenous regressors. The null hypothesis is of exogeneity

IV regression in R

- The TSLS estimator is computed in R using the command *ivreg*
- Syntax is *ivreg* (*Y ~ X + W* | *W + Z*) where *X* is the endogenous variable, *Z* the instrument(s) and *W* the control(s)
- Diagnostic tests can be run by adding ,*diagnostics = TRUE* inside the *summary* command. This reports:
    - Weak instruments: This tests the null hypothesis that we have weak instruments
    - Wu-Hausman (we didn't discuss this)
    - Sargan: This is the J-test of instrument exogeneity. It can only be used if you have more instruments than endogenous regressors. The null hypothesis is of exogeneity

Attendance