Instrumental Variables: Introduction

Labor Economics

Instructor: Haoran LEI

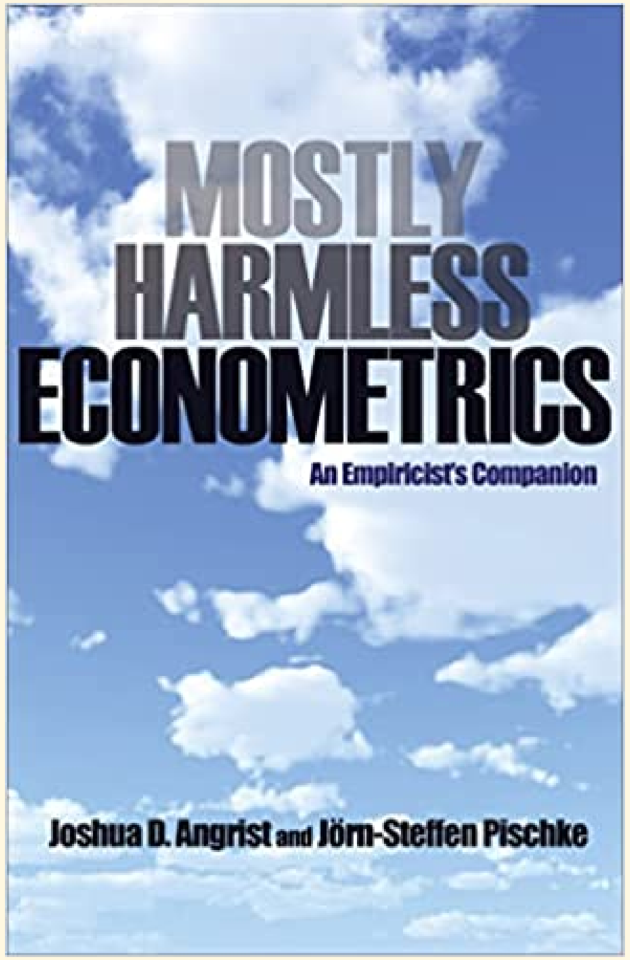
Hunan University

Instrumental variables

- I'll tell *two stories* about IV.
- The first story follows the *traditional approach*: **endogeneity**, **inconsistent LS estimators**, and **two-stage LS**
  - The first story may also serve as a review for undergraduate econometrics.
- The second story takes the *"modern" causality approach*: **confounder**, **causal effect**, and **directed acyclic graph (DAG)**.
  - The second story is gaining more popularity in empirical research and has been used widely in labor economics.

More "hype" of the second story

Angrist and Imbens (together with labor economist Card) won the Nobel Prize in 2021 for their "*methodological contributions to the analysis of **causal relationships.***"



Endogeneity problem

Consider the simple linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where Y_i is individual i's wage and X_i is i's years in education.

Endogeneity problem

Consider the simple linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where Y_i is individual i's wage and X_i is i's years in education.

We say X is **endogenous** if X is correlated with ε:

- Notable reasons for the endogeneity problem include omitted variables and measurement errors.

```r
set.seed(2022); N = 10000
b0 = 0.5; b1 = 1 # coefficients
x = runif(N)
e = rnorm(N)
x = x + e/2 # make x correlated with e
w = b0 + b1*x + e
my_lm = lm(w ~ x)
my_lm$coefficients
```

```
(Intercept)           x
 -0.2498114   2.4963632
```

```r
cor(x,e) #0.8644214
```

Endogeneity leads to inconsistent LS estimators

- LS estimators are **inconsistent** when there's an endogeneity problem.
  - I.e., $\hat{\beta}^{LS}$ does not approach β **even with infinite data**
- In the R simulation, the estimated $\hat{\beta}_1$ is around 2.5 while the truth β_1 is 1.0.
- Intuition for why $\hat{\beta}_1 > \beta_1$: when X is positively correlated with ε, an increase in X has two effects on Y:

1. higher $\beta_1 X_1$
2. higher ε on average

Sources of endogeneity: omitted variables

- Suppose the true relationship is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 A_i + e_i$$

- where A_i stands for individual i'th ability
- However, abilities A_i's are **unobservable**. A is correlated with X
- In the linear model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, the error term $\varepsilon = \beta_2 A + e$ is correlated with X

Instrumental variable

- A common tool to correct for endogeneity is "Instrumental Variable".
- Z is a valid **instrumental variable** for X if

1. X and Z are correlated
2. ε and Z are *not* correlated

- Intuiton: An ideal instrumental variable Z contains all the info in X except those correlated with ε

The Two-Stage Least Squares Estimator (2SLS)

Stage 1:

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

- $\hat{Z}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$ is the component of X_i that is explained by Z_i
- v_i is the component that cannot be explained by Z_i and exhibits correlation with ε_i

The Two-Stage Least Squares Estimator (2SLS)

Stage 2:

$$Y_i = \beta_0 + \beta_1 \hat{Z}_i + \epsilon_i$$

- $\hat{Z}$ is obtained in the first regression and is uncorrelated with ϵ
- We get $\hat{\beta}_1^{2SLS}$, which is a consistent estimator for β_1.

```r
set.seed(2022)
N = 1000
b0 = 0.5; b1 = 1
z = runif(N)
e = rnorm(N)
x = 2*z + e/2
y = b0 + b1*x + e
# 1st stage LS
ls1 = lm(x ~ z)
z_hat = ls1$fitted.values
# 2nd stage LS
ls2 = lm(y ~ z_hat)
ls2$coefficients
```

```
(Intercept)       z_hat
  0.5099503   1.0086850
```

The function `ivreg()` from the package `AER` carries out 2SLS procedure automatically. It is used similarly as lm().

```
library(AER)
ivreg(y ~ x | z)
```

```
Call:
ivreg(formula = y ~ x | z)

Coefficients:
(Intercept)            x
      0.510        1.009
```

Notes on computing standard errors

- Running the individual regressions for each stage of 2SLS using `lm()` leads to the same coefficient estimates as when using `ivreg()`
- However, the standard errors reported for the second-stage regression by `summary(ls2)` are invalid:
  - Special adjusts are needed for using predictions from the first-stage regression $\hat{Z}$ as regressors in 2nd regression.
- `ivreg()` performs the necessary adjustment automatically.

```r
summary(ivreg(y ~ x | z))
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-3.31046 -0.66313 -0.02887  0.71758  3.04277
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.50995    0.06274   8.127 1.29e-15 ***
x            1.00868    0.05404  18.666  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9918 on 998 degrees of freedom
Multiple R-Squared: 0.6181,     Adjusted R-squared: 0.6178
Wald test: 348.4 on 1 and 998 DF,  p-value: < 2.2e-16
```

Recap

- X is *endogenous* if it's correalted with ε. In that case, the LS estimator is not consistent.
- Z is a valid *instrumental variable* for X if

1. Z is correalted with X, and
2. Z is not correlated with ε

- Using 2SLS regression, we
  - first regress Z on X and use the fitted $\hat{Z}$ as a proxy for X
  - Then regress Y on $\hat{Z}$ to get a consistent estimator

General Instrumental Variables Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \beta_{k+1} W_1 + \cdots + \beta_{k+r} W_r + \varepsilon$$

- Y is the dependent variable
- $X_1, \ldots, X_k$ are k variables that are correlated with ε
- $W_1, \ldots, W_r$ are control variables and are uncorrelated with ε
- $Z_1, \ldots, Z_m$ are m instrumental variables

2SLS

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \beta_{k+1} W_1 + \cdots + \beta_{k+r} W_r + \varepsilon$$

1. First-stage regressions

- Regress X_j on all IVs $(Z_1, \ldots, Z_m)$ for all X_j, $j = 1, \ldots, k$.
- Obtain the fitted values $\hat{X}_j$, $j = 1, \ldots, k$.

2. Second-stage regression

- Regress Y on all $(\hat{X}_1, \ldots, \hat{X}_k, W_1, \ldots, W_r)$.
- Obtain the 2SLS estimands: $\hat{\beta}_j$, $j = 1, \ldots, k$.

The art of finding IVs

- Finding valid IVs requires **detailed institutional knowledge** and the **investigation and quantification of the forces** at work *in a particular setting*.
- When Y = "Wage" and X = "Schooling years", IVs =
    - Region and time variation in school construction, Duflo (2001)
    - Distance to college, Card (1995)
    - Quarter of birth, Angrist and Krueger (1991)
    - ... etc. For more examples, see Angrist and Krueger (2001, *JEP*). You may also just google "IV econ papers".

> “ Our view is that progress in the application of instrumental variables methods depends mostly on the gritty work of finding or creating plausible experiments that can be used to measure important economic relationships—what statistician David Freedman (1991) has called “shoe-leather” research. Here the challenges are not primarily technical in the sense of requiring new theorems or estimators. Rather, progress comes from detailed institutional knowledge and the careful investigation and quantification of the forces at work in a particular setting. Of course, such endeavors are not really new. They have always been at the heart of good empirical research. --- Angrist and Krueger (2001, *JEP*) ”