IV Estimation and its Limitations:

Weak Instruments and Weakly Endogeneous Regressors

Laura Mayoral

IAE, Barcelona GSE and University of Gothenburg

Gothenburg, May 2015

Roadmap of the course

- Introduction. Why do we need IVs?
- IV estimation: 2SLS (brief overview), GMM and LIML (even briefer!)
- Consequences of the failure of the assumptions behind IV estimation. Irrelevant/Weak instruments and Endogeneous/weakly endogenous instruments
- Detection of weak instruments
- Inference robust to weak instruments

1. Introduction

1.1 Causality versus correlation

- Goal: Estimate the causal effect of some independent variable(s), X, on a dependent variable, Y.

- OLS on a linear model relating Y and X:

$$Y = X\beta + U$$

- We obtain: $\widehat{\beta}_{ols}, \hat{\Sigma}_{\hat{\beta}}$

- Two key questions

- Under what circumstances $\widehat{\beta}_{ols}$ measures the causal effect of X on Y?

- How shall we interpret $\widehat{\beta}_{ols}$ when these conditions do not hold?

1.2. Endogeneity

- Consider the simplest case: X contains a constant and 1 variable. Then,

$$\widehat{\beta}_{ols} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'U \quad (1)$$

$$\xrightarrow{p} \beta + \frac{cov(X,U)}{var(X)} \quad (2)$$

- Consistency of $\widehat{\beta}_{ols}$ requires $cov(X,U) = 0$
- If $cov(X,U) \neq 0 \longrightarrow$ X is endogeneous.
- If X is endogenous, what is $\widehat{\beta}_{ols}$ measuring?

1.3. Main causes of Endogeneity

1. Omitted variables

2. Errors-in-variables (measurement error)

3. Simultaneous Causality

i) Omitted variables

- Problem:

You should estimate

$$Y = X\beta + Z\delta + U$$

But you estimate

$$Y = X\beta + U^*$$

where $U^* = Z\delta + U$.

- If $cov(X, Z) \neq 0 \longrightarrow$ X is endogeneous.

i) Omitted variables

- Problem:

You should estimate

$$Y = X\beta + Z\delta + U$$

But you estimate

$$Y = X\beta + U^*$$

where $U^* = Z\delta + U$.

- If $cov(X, Z) \neq 0 \longrightarrow$ X is endogeneous.
- The bias:

$$\begin{aligned} \widehat{\beta}_{ols} &= (X'X)^{-1}X'Y \\ &= \beta + (X'X)^{-1}X'Z\delta + (X'X)^{-1}X'U \\ &\xrightarrow{p} \beta + E(X'X)^{-1}E(X'Z)\delta \end{aligned}$$

Example

- Effect of an additional year of schooling on wages

$$log(wage_i) = \beta_0 + \beta_1 tenure_i + \beta_2 educ_i + u_i$$

- innate ability, IA, is a missing variable.
- *educ* and IA are likely to be correlated.
- *educ* is endogenous
- $\widehat{\beta}_{2_{ols}}$ is likely to overestimate or underestimate β_2?

Solutions to omitted variables

- include Z in the regression.
- If Z is not available but a *proxy for Z, Z*, include it.*
- *Example: wage vs. years of education. Proxy: IQ score as a proxy for IA.*
- *Panel Data*
- *Find an instrument for X and estimate your model using IV methods.*

ii) Simultaneous Causality

- *Problem: Direction of causality, X* $\Longrightarrow$ *Y and Y* $\Longrightarrow$ *X.*
- *X is endogeneous because:*

$$\begin{aligned} Y &= X\beta + U \\ X &= Y\delta + V \end{aligned}$$

- $cov(X, U) \neq 0$ *because X is a function of Y and Y is correlated with U!*

Example *Elasticity of demand for wheat*

- *Price and Quantities are jointly determined through the interaction of the supply and demand curves*

$$log(Q_i) = \beta_0 + \beta_1 log(P_i) + U_i$$

- *P is endogenous*

Solution

- *Instrumental variables.*

- *other solutions: randomized controlled experiment with no reverse causality.*

iii) Errors-in-variables

- *Problem:* $Y = X\beta + U$ *but X is measured with error,* X^*.
- *why errors? typographical errors, survey data, etc.*
- *You estimate* $Y = X^*\beta + U^*$

$$Y = X^*\beta + ((X - X^*)\beta + U)$$

- X^* *is endogeneous if* X^* *is correlated with the measurement error* $e = (X - X^*)$

Solutions to measurement error

- *Get a more accurate measure of X.*
- *Use instrumental variables.*

1.4. A general solution to endogeneity: Instrumental variables

- *Suppose you suspect that X is endogenous so you want to instrument it.*

- *Simplest case: 1 endogeneous variable, 1 IV.*

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

  - X_i *endogenous*

- *Simplest approach: use 2SLS to estimate* β

Two stage least squares, I

- *First stage:*
- *decompose X in two parts*
  - *a "problematic" component: correlated with U*
  - *a "problem free" component: uncorrelated with U*

Two stage least squares, I

- *First stage:*
- *decompose X in two parts*
  - *a "problematic" component: correlated with U*
  - *a "problem free" component: uncorrelated with U*
- *Second stage*
- *regress Y on the "problem free" component of X.*

Two stage least squares, II

- *2 OLS regressions*

- *First stage:* $X_i = \pi_0 + \pi_1 Z_i + \nu_i \Longrightarrow \hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$

- *Second stage:* $Y_i = \hat{\beta}_0 + \hat{\beta}_1 \hat{X}_i + \hat{u}_i$

- *It is easy to show that*

$$\hat{\beta}_{1_{2sls}} = \frac{\widehat{cov(Z_i, Y_i)}}{\widehat{cov(Z_i, X_i)}}$$

Two stage least squares, III

- *If the sample size is large enough (under suitable assumptions), sample moments are "close" to the population moments:*

$$\frac{cov(Z_i, Y_i)}{cov(Z_i, X_i)} = \frac{\beta_1 cov(Z_i, X_i) + cov(Z_i, U_i)}{cov(Z_i, X_i)} =$$

$$= \beta_1 + \frac{cov(Z_i, U_i)}{cov(Z_i, X_i)}$$

- *For $\hat{\beta}_{1_{2sls}}$ to approach β_1 as the sample size gets large, we need*

$$\frac{cov(Z_i, U_i)}{cov(Z_i, X_i)} = 0$$

Two stage least squares, IV: instrument validity

- *Z is a valid instrument for X if*

- *Instrument relevance:* $cov(Z_i, X_i) \neq 0$ *(relevance condition)*

- *Instrument exogeneity:* $cov(Z_i, U_i) = 0$ *(exclusion condition)*

Example I: 2SLS using Stata

- *Estimating the elasticity of demand of Flaxseed*

- *First application of IV regression (1926)*

- *annual data, 1904-23*

$$log(Q_i) = \beta_0 + \beta_1 log(P_i) + \beta_2 year_i + u_i$$

- β_1= % *increase in Q if P increases 1%.*

- *IV: (log) rainfall ("supply shifter")*

```
. ivregress 2sls loutput year (lprice = lrainfall), r first;

First-stage regressions
-----------------------

                                                  Number of obs   =         22
                                                  F(   2,     19) =      11.58
                                                  Prob > F        =     0.0005
                                                  R-squared       =     0.4939
                                                  Adj R-squared   =     0.4407
                                                  Root MSE        =     0.1479

------------------------------------------------------------------------------
             |               Robust
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   .0088182    .003869     2.28   0.034     .0007202    .0169161
   lrainfall |   -.484159   .1444125    -3.35   0.003    -.7864178   -.1819002
       _cons |  -11.25359    7.46855    -1.51   0.148    -26.88545    4.378263
------------------------------------------------------------------------------
```

```
Instrumental variables (2SLS) regression              Number of obs =      22
                                                      Wald chi2(2)  =   14.80
                                                      Prob > chi2   =  0.0006
                                                      R-squared     =  0.4412
                                                      Root MSE      =  .32487

------------------------------------------------------------------------------
             |               Robust
     loutput |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lprice |  -.7417028   .3427025    -2.16   0.030    -1.413387   -.0700182
        year |  -.0342571   .0151899    -2.26   0.024    -.0640287   -.0044855
       _cons |   72.15604   28.46158     2.54   0.011     16.37236    127.9397
------------------------------------------------------------------------------
Instrumented:  lprice
Instruments:   year lrainfall
```

Course overview, I

This course focuses on understanding the behaviour of IV estimators when the above assumptions fail or almost fail.

- *If the relevance restriction fails (or is close to failing)*

$\longrightarrow$ *Z is irrelevant (or weak).*

- *If the exclusion restriction fails (or is close to failing)*

$\longrightarrow$ *Z is endogenous (or weakly endogenous).*

Course overview, II

We'll focus on

- *How to estimate β using IVs.*
- *What happens with your estimators and test of hypothesis when*
  - *Z is weak (the literature is ample here)*
  - *Z is (weakly) endogenous (the literature is much more limited here)*
  - *Z is weak and endogenous (this case is still largely unexplored)*
- *How to detect weak instruments*
- *How to carry out inference when instruments are weak*

- *The literature typically treats the relevance and the exogeneity of the instruments as two different problems.*

- *But recall,*

$$\hat{\beta}_{1_{2sls}} = \beta_1 + \frac{\widehat{cov(Z_i, U_i)}}{\widehat{cov(Z_i, X_i)}}$$

- *Thus, what matters is what happens with the ratio of these two quantities.*

- *When searching for IVs, we often face a tradeoff:*

  - *'More exogenous' instruments (i.e., less correlated with the error term) are often weaker. And 'stronger' instruments are usually more at risk of being 'more' endogenous (i.e., more correlated with the error term).*

- *Our conclusions we'll hopefully shed light on how to look for IVs in applied work.*

2. IV estimation

- A bit of history: http://scholar.harvard.edu/stock/content/history-iv-regression

Philip Wright (1861-1934)
Economist, teacher, poet
MA Harvard, Econ, 1887
Lecturer, Harvard, 1913-1917

Sewall Wright (1889-1988)
genetic statistician
ScD Harvard, Biology, 1915
Prof., U. Chicago, 1930-1954

- *The Wrights (father and son) were interested in how to estimate the slope of supply and demand curves of agricultural products when observed prices and quantities are determined by the intersection of the supply and demand lines*

- *Wright (1928) first solves the problem. He notes that if there is an observed variable that shifts supply but not demand, this variable could be used to estimate the slope of the demand curve.*

Different IV estimation procedures

- *Two stage least squares (2SLS). Very popular, mainly because it's efficient (in the class of IV estimators) under conditional homokedasticity.*

- *GMM (Hansen, 1982): more efficient than 2SLS under heterokedasticity. See Wooldridge 2010 (Chapter 8).*

- *k-estimators: LIML, Fuller K-estimators,... (OLS and 2SLS also belong to this group).*

- *Limited Information maximum likelihood (LIML): Less precise than 2SLS but less biased when instruments are weak (we'll come back to this point).*

In this short course we'll mainly focus on 2SLS. See Andrews and Stock (2005) and Stock and Yogo (2005) for a more general treatment.

2SLS: general treatment

- *Consider the following model:*

$$y = X\beta_0 + \epsilon; \tag{3}$$

$$\tag{4}$$

where,

- X *is a* $N \times K$ *matrix, several components of X can be correlated with* ϵ *(multiple endogenous regressors).*
- Z *is a* $N \times L$ *matrix containing the exogenous variables in X plus the instruments.*
- *The number of instruments can be larger than the number of endogenous regressors, so* $L \geq K$

- *The 2SLS estimate of β is given by*

$$\hat{\beta}_{2sls} = (X'P_zX)^{-1}(X'P_zy);\ P_z = Z(Z'Z)^{-1}Z'$$

If both Z and X are $N \times 1$ (simplest case–as above)

$$\hat{\beta}_{2sls} = \frac{Z'y}{Z'X} \longrightarrow (\hat{\beta}_{2sls} - \beta_0) = \frac{Z'\epsilon}{Z'X}$$

Asymptotic properties

Consistency

Two standard conditions are needed (same as before!)

$$1. \quad E(Z'\epsilon) = 0, \ \textit{(exclusion condition)}$$
$$2. \quad rank(E(Z'X)) = K, \textit{(relevance condition)}$$

Condition 1 implies that Z is exogeneous, condition 2, that it is "strong".

- *Under these conditions, $\hat{\beta}_{2sls}$ is consistent:*

$$\hat{\beta}_{2sls} - \beta_0 \xrightarrow{p} 0$$

Asymptotic distribution

- *By the CLT, under conditional homokedasticity of ϵ,*

$$T^{1/2}(\hat{\beta}_{2sls} - \beta_0) \quad \xrightarrow{d} \quad N(0, \sigma_\epsilon^2 (E(X'Z)(E(Z'Z)^{-1})E(Z'X))^{-1})$$

If Z and X are univariate and $X = \Pi Z + v$

$$T^{1/2}(\hat{\beta}_{2sls} - \beta_0) \quad \xrightarrow{d} \quad N(0, \sigma_\epsilon^2 \frac{E(Z'Z)}{E(Z'X)^2}) = N(0, \frac{\sigma_\epsilon^2}{E(Z'Z)\Pi^2})$$

- *Remark: the asymptotic variance is closely related to a measure of the 'strength' of the instrument, μ^2 (we'll come back to this point latter on), where*

$$\mu^2 = \frac{E(Z'Z)\Pi^2}{\sigma_v^2}$$

Other IV estimation procedures

- *2SLS is very popular mainly because (under homokesdasticity of ϵ) is efficient (i.e., smallest variance) in the class of all instrumental variables estimators using instruments linear in Z.*
- *GMM (Hansen 1982): more efficient than 2SLS if homokedasticity fails. See Wooldridge, Chapter 8.*
- *Limited Information maximum likelihood (Anderson and Rubin, 1949).*
- *LIML is a linear combination of the OLS and 2SLS estimate (with weights depending on the data), and they the weights happen to be such that they (approximately) eliminate the 2SLS bias.*
- *Although less precise than 2SLS in general, under deviations from the standard assumptions, LIML behaves much better than 2SLS (we'll come back to this).*

Wrapping up

- *We are usually interested in estimating causal effects*
- *Independent variables are often endogenous due to different reasons*
- *IV estimation is a general solution to the problem above*
- *When carefully implemented, IV is great!*
- *But unfortunately, there are some problems as well...*

Why should we not always use IV?

- *It's difficult to find good IVs.*

- *Even if instruments are good*

- *IV estimators are biased and their finite-sample properties are often problematic $\longrightarrow$ most of the justification for the use of IV is asymptotic.*

- *Performance in small samples may be poor. (Remember: OLS is consistent AND unbiased if X is exogenous)*

- *Even if instruments are good:*

- *the precision of IV estimates is lower than that of OLS estimates (i.e., standard errors are larger) .*

- *IV estimators have good properties ONLY if instruments are both relevant and strong.*

- *But what happens otherwise?*

A few examples using STATA

- You can use the built-in command ivreg or the user-written ivreg2.

- They are similar, but the latter provides more complete output.

- A few examples using ivreg (from Wooldridge)

http://fmwww.bc.edu/gstat/examples/wooldridge/wooldridge15.html

- And some more

http://www.nuff.ox.ac.uk/teaching/economics/bond/IV%20Estimati
20Using%20Stata.pdf