

Instrumental Variables

Nathaniel Beck

Department of Politics, NYU, New York, NY 10012, nathaniel.beck@nyu.edu

QII - Week 2

1

Endogeneity and OLS

- Begin with a simple linear equation, but let x be a single endogenous variable (can have more, but notation more complex) and W be one or more exogenous variables (doesn't matter how many)

$$\mathbf{y} = \mathbf{x}\beta + \mathbf{W}\gamma + \varepsilon \quad (1)$$

- Endogeneity means that x and the errors are not independent
- A multiple equation structural model would give this, since the errors in the equations are allowed to be correlated
- Thus OLS is no longer unbiased, nor consistent
- The assumption of exogeneity was critical in our proofs that OLS is unbiased and consistent
- This does not mean OLS is terrible; need to discuss how it compares to other estimators after seeing those estimators
- A little bit of endogeneity is not a disaster; small correlation between x and the errors induces small amount of bias

2

- C&T have a nice way to think about - we care about $\frac{dy}{dx}$ item but if a change in x is associated with both a direct change in y and an indirect change through the errors,

- then have

$$\frac{dy}{dx} = \beta + \frac{de}{dx} \quad (2)$$

- but OLS assumes the second term is zero

3

Instrumental variables

- An instrument is some variable, z , that is correlated with x but not with the error (so is exogenous but correlated with an endogenous x)
- Examples - rainfall for economic growth (dv is civil war)
- AJR Settler mortality for economic institutions (dv is growth)
- Parent's Party ID for respondent's (dv is vote)
- How do you know that something is a valid instrument - theory?
- But for rainfall, are pretty sure it is

4

- To see why this works, note we can consistently (by OLS) estimate $\frac{dy}{dz}$
- since z is exogenous (but the impact of z on y is not of direct interest)
- We can also consistently estimate $\frac{dx}{dz}$
- But the impact of z on x is not of direct interest
- Let us assume we have a simple bivariate regression of y on scalar x which is endogenous

$$\frac{dy}{dx} = \frac{\frac{dy}{dz}}{\frac{dx}{dz}} \quad (3)$$

$$= \frac{\frac{\sum z_i y_i}{\sum z_i^2}}{\frac{\sum z_i x_i}{\sum z_i^2}} \quad (4)$$

$$= \frac{\sum z_i y_i}{\sum z_i x_i} \quad (5)$$

5

- If we have a multiple regression, let \mathbf{z} be a vector of instruments
- Note for exogenous x , the instrument is just x
- $\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$
- Where \mathbf{X} is the usual data matrix, \mathbf{y} is the usual vector of the dep var, and \mathbf{Z} is a matrix the same size as \mathbf{X} , but replacing the endogenous columns of \mathbf{X} with the corresponding instruments

6

IV is unbiased

- As with OLS, we just compute the expectation and plug in for the true \mathbf{y}

$$E(\hat{\beta}_{IV}) = E((\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}) \quad (6)$$

$$= E((\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'(\mathbf{X}\beta + \varepsilon)) \quad (7)$$

$$= \beta + E((\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\varepsilon) \quad (8)$$

$$= \beta \quad (9)$$

- The second term is zero by the same argument as day one, that is by the assumption that the instruments and the error are uncorrelated
- $\mathbf{Z}'\mathbf{X}$ is non-singular by the assumption that the instruments are correlated with the variables they instrument for

7

IV is consistent and asymptotically normal

- Consistency depends on the plim of the second term in Eq. 8
- By the same assumptions that worked for OLS, this goes to zero and so IV is consistent
- Again assuming that the limit of $\mathbf{Z}'\mathbf{X}$ is finite
- And by the same argument as for OLS, we get asymptotic normality
- (with a variance-covariance matrix now having $\mathbf{Z}'\mathbf{X}$ terms in place of $\mathbf{X}'\mathbf{X}$ terms)
- Under assumption of heteroskedasticity, the VCV matrix of the IV estimator is
- $\sigma^2(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{X})^{-1}$

8

How know an instrument is 'good'

- Can look at data to see correlation of x and z
- CANNOT see if z and error are uncorrelated (z is exogenous)
- Need an argument
- Eg rainfall for growth is good in that rainfall must be exogenous (or nothing is!)
- Then examine whether rainfall and growth are correlated - by direct computation

9

Efficiency and MSE

- Unbiased just tells you on average your estimator is fine
- But not that any given sample will give you an estimator that is close to what it is estimating
- Consistency says that eventually everything is perfect
- But what happens before you reach asymptopia?
- If we have two unbiased estimators, would prefer the one with smaller variance
- Think of a single parameter (generalization to a vector of parameters is pretty easy though not totally trivial)
- For two unbiased estimators, $\hat{\beta}_1$ is more efficient than $\hat{\beta}_2$ if $\text{VAR}(\hat{\beta}_1) < \text{VAR}(\hat{\beta}_2)$
- Need to do this theoretically in that do not observe more than one value of an estimator!
- B in BLUE says OLS has least variance in the class of LINEAR UNBIASED estimators

10

- What we really care about is how far, on average, our estimator differs from truth
- Using a mean square error criteria, our interest is then in
- $E(\hat{\beta} - \beta)^2$
- Call this MSE (mean square error) - sometime use square root of this (RMSE) so in unit of β and not its square
- If $\hat{\beta}$ is an unbiased estimator, $E(\hat{\beta}) = \beta$ so this just the variance
- if not unbiased, simple algebra shows
- $E(\hat{\beta} - \beta)^2 = \text{VAR}(\hat{\beta}) + E(\hat{\beta} - \beta)^2$
- If unbiased second term is zero
- A biased estimator can have lower variance
- Bias-Variance trade-off

11

Weak instruments

- IV can have high variance
- Note - its variance has $(\mathbf{Z}'\mathbf{X})^{-1}$ terms
- This is just the covariance between the instrument and what it is instrumenting
- Standardize this and we have the correlation
- If correlation of x and z is low, when we invert, get a big number
- Thus a weak instrument (low correlation with x) yields unbiased, consistent but large MSE
- If correlation is low enough can have worse MSE than OLS, even though OLS is biased and inconsistent
- With infinite N IV wins, but we do not live in that world
- The problem is exacerbated if the instrument is a bit correlated with the errors (as it can be, though rainfall should be okay)
- Thus you need to make sure your instruments are reasonably correlated with what they instrument

12

- Lots of rules of thumb, but depends on N, etc. All related to R^2 of x and z, but no cutoff as to when instrument is “too weak”
- But worry when instrument and x are purely related
- Eg closeness of a college and educational level attained
- Oh, for an experiment!!!!

13

IV solution for simeq

- Suppose one has an identified or overidentified set of structural equations
- Say in the first equation we have y_2 as a regressor, some exogenous regressors, but two exogenous variables in the system that are excluded from the first equation

$$y_1 = \beta_1 y_2 + \gamma_{1,1} x_1 + \epsilon_1 \quad (10)$$

$$y_2 = \beta_2 y_1 + \gamma_{2,2} x_2 + \gamma_{2,3} x_3 + \epsilon_2 \quad (11)$$

- Second equation is just identified, first is overidentified
- Estimate each equation by IV

14

- What could we use as instruments?
- We saw we could compute the reduced form
- We could then estimate the parms of $y_2 = d_1x_1 + d_2x_2 + d_3x_3$
- Note that even though x_2 and x_3 do not affect y directly
- They do affect y indirectly, reduced form does not care
- So after running the first stage regression, compute \hat{y}_2
- This is a linear function of the exogenous variables, so is exogenous
- It is related to y_2 by the construction of the reduced form
- So use \hat{y}_2 as the instrument for y_2
- Hence called two stage least squares
- Is good insofar as system is good and y_2 well predicted by the exogenous variables
- Lots of language and slightly old fashioned ways of thinking about this, but is EXACTLY IV
- Will derive 4.53 on whiteboard

15

Old fashioned

- When computers were slow back in the dark ages
- One could do 2SLS by first computing instruments as in IV
- But with some easy matrix manipulations, the IV estimator is just the regression of y on the instruments (doing OLS)
- Follows from idempotency of the hat matrix
- So could just do two rounds of OLS, which was great in 1953 (Theil)
- even useful in 1970, but not today
- Also SE's were wrong from this procedure
- So just think of this as IV
- Will also see can do maximum likelihood, but that is next week

16

- Systems of equations were all what people worked on in the 1950's (Cowles Foundation)
- Incredibly heavy lifting
- But hard to find systems that could be identified
- See this most clearly in large macroeconomic models
- Pretty much dead
- Sims, Macroeconomics and Reality, AER, 1980
- Political economy - resurgence of IV (almost de rigueur)
- But still question of identification and setting up of an implicit system to find iv's