Micro Econometrics

Sol Yates

February 29, 2024

Contents

Lecture 4: IV p2

Tue 20 Feb 16:07

1 IV

Model $y = x\beta + u$ with x a vector of k exogenous and endogenous regressors and z a vector of m IV's (including the exogenous variable)

1. m = k : the model is just identiified ,we have an instrument for each endogenous variable $\Rightarrow$ use IV
2. m < k : the model is not identified, we do not have enough IVs
3. m > k : the model is over-identified $\rightarrow$ we have too many IVs. Use GIVE / 2SLS

1.1 Case : length(z) = length(x)

Model : $y = x\beta + u$, $x = (q, x_2, \dots, x_k)$ and $z = (1, x_2, \dots, x_{k-1}, z_1)$. We know $Cov(x_j, u) = 0$ for $j = 2, \dots, k-1$ and $Cov(x_k, u \neq 0)$

We have an instrument for x_k :

- Exogenous $Cov(z_1, u) = 0$
- Partial Correlation : $\theta_1 \neq 0$ in $x_k = \delta_1 + \delta_2 x_2 + \dots + \delta_{k-1}x_{k-1} + \theta_1 z_1 + r_k$

Where the moment conditions imply :

$$E[z'u] = E[z'(y - x\beta)] = 0$$

We have one instrument at our disposal for this endogenous regressor, we include the constant and all the exogenous regressors because they can be used for instruments for themselves.

Partial correlation best seen by regressing endogenous regressor x_k on all exogenous variables plus the instrument for x_k and we need the parameter on the instrument θ_1 not to be 0, thus partial correlation, the correlation cannot be 0 after the other effects have been 'netted' out. Different to simple case where sufficient to have covariance between endogenous regressor and instrument $\neq 0$

Exogeneity leads to above expression, plugging in expression for u.

Multiplying the model through with z', taking expectation and using the moment condition :

$$E[z'y] = E[z'x]\beta$$
$$\text{if rank} \quad E[z'x] = k$$
$$\beta = [E[z'x]]^{-1}E[z'y]$$

There is a *unique solution only under full rank* and it can be shown that if we rule out perfect collinearity in z, full rank holds iff $\theta_1 \neq 0$

Given a random sample, we can estimate consistently :

$$\hat{\beta}^{IV} = \left(\frac{1}{n}\sum_{i=1}^{n} z_i'x_i\right))^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i'y_i\right)) = (Z'X)^{-1}Z'Y$$

Where Z and X are $n \times k$ data matrices and y is $N \times 1$

Given the assumptions, this estimator is consistent

1.2 2SLS/GIVE

Case: length(z) $>$ length(x):

the idea is to used the fitted values from the first stage regression of the endogenous regressor on all the exogenous variables (including the instruments) and use them as "instruments" in the IV estimator

$$z = (1, x_1, \ldots, x_{k-1}, z_1, \ldots, z_l) - m = k + I \quad \text{vector for} \quad x_k$$

1. Fitted values from the first stage $\hat{x_i} = (1, x_1, \ldots, x_{k-1}, \hat{x}_k)$

$$\hat{x}_{ik} = \hat{\delta_0} + \hat{\delta_1}x_{i1} + \ldots + \hat{\delta}_{k-1}x_{i,k-1} + \hat{\theta}_1 z_{i1} + \ldots + \hat{\theta}_I z_{il}$$
$$\hat{x_i} = z_i\left(\sum z_i'z_i\right)^{-1} z_i'x_i$$
$$\hat{X} = Z(Z'Z)^{-1}Z'X$$

For this endogenous regressor you have several potential instruments at your disposal, you would then regress on exogenous variables from initial model and instruments. That gives you a vector of instruments that is equal to (including all potential instruments).

Then you start with obtaining fitted values from First Stage (FS) regressing x_k on exogenous regressors δ and instruments z

Using the fitted values as instruments :

$$\hat{\beta}^{IV} = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{x_i}'x_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{x_i}y_i\right) = (\hat{X}'X)^{-1}\hat{X}'Y$$

Then, using calculus, we can show that $\hat{X}'X = \hat{X}'\hat{X}$ and hence

$$\hat{\beta}^{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$$

Which is the GIVE / 2SLS estimator since it equals the OLS estimator on the fitted values from the first stage

Where we have simply replaced with fitted values, and then replaced in matrix form. We can essentially show this is the OLS estimator on the FS using the fitted values. Then we use this to plug in the IV estimator to obtain the OLS estimator on the fitted values

To obtain the β

1. First Stage: Obtain the fitted values $\hat{x}_k$ from the regression x_k on $1, x_1, \ldots, x_{k-1}, z_1, \ldots, z_l$
2. Second Stage: Run the OLS regression: y on $1 + x_1, \ldots, x_{k-1} + \hat{x_k}$

- However, omitting the exogenous regressors in the first stage is easily done and will lead to inconsistency
- And, SE obtained from the second step are incorrect

Testing for rank condition : $H_0 : \theta_1 = \ldots = \theta_l = 0$ vs at least one θ_s for $s - 1, \ldots, l$ is non zero.

1.3 Properties

Here x is 1 x k and generally includes unity, several elements of x may be endogenous, while z includes any exogenous variable

Assumption 1 :

1.4 Group Mean Estimator

In some situations, have instruments that can be changed into 2 groups, water (of birth/ financial year). 'Chop instrument into groups' like Moulton problem/structure.

It can be shown that group mean estimator is IV, a weighted least squares regression, where it is sufficient to know size of groups and means, do regression and obtain estimator that is equivalent to an IV estimator, that is consistent despite the fact we have an endogenous variable.

- where x_{ig} is endogenous
- we have g *moment conditions*, if this is IV, we know exogeneity must hold, whether group 1 or 2, the error term conditional on this group needs to be equal to 0, this must hold for all groups. Essentially, we have g different groups this is really $E[y_{ig}|z_g = I] = \beta_0 + \beta_1 E[x_{ig}|i : z_g = I]$
- To estimate an expectation, we replace with an average, since this is conditional, to estimate the expectation for the first group (born in the 1-st quarter), we take the average for the first group (conditional average by restricting to the first group) - taking the means of all the groups
- we do the same for $\overline{x}$ and intuitively obtain

$$\overline{y_g} = \beta_0 + \beta_1 \overline{x_g} + \overline{u}_g$$

Doing this is the same as using dummy variables for quarter of birth in 2SLS regression, **thus** group means are consistent.

Exercise 1. Group mean estimator - what happens if the number of groups = 2 If we have dummy variable, we obtain the Wald estimator (last week). Our instrument, we obtain the same expression In order to derive,

1. regress x on dummies, x can only belong to 1, so fitted values are sample means of dependent variable x_{ig}. fitted values $\hat{x_{ig}}$ are means $\overline{x}_g$
2. Apply OLS on this after we have found fitted values, the predicted values are our sample means

Angrist and Krueger

- Does compulsory school attendance affect schooling and earnings
- using *quarter of birth* as an instrument
  1. Exclusion : Season of birth is a natural experiment and hence unrelated to innate ability, motivation or family connections
  2. Relevance : In the US, children were allowed to drop out at 16. Since the age of starting school differs, children have different lengths of schooling when they turn 16
- Potentially weak instrument and potential reasons why quarter of birth might be somewhat correlated with the error

TABLE 4.1.2
Wald estimates of the returns to schooling using quarter-of-birth instruments

	(1) Born in 1st Quarter of Year	(2) Born in 4th Quarter of Year	(3) Difference (Std. Error) (1) − (2)
ln (weekly wage)	5.892	5.905	−.0135 (.0034)
Years of education	12.688	12.839	−.151 (.016)
Wald estimate of return to education			.089 (.021)
OLS estimate of return to education			.070 (.0005)

Notes: From Angrist and Imbens (1995). The sample includes native-born men with positive earnings from the 1930–39 birth cohorts in the 1980 census 5 percent file. The sample size is 162,515.

Figure 1: Wald estimates of IV - weak/exclusion violated?

can't test by how much IV exclusion is violated, it might be best to use OLS, but in the same sense it may be incorrect - can we search for better instrument? Or,

TABLE 4.1.1 2SLS estimates of the economic returns to schooling

	OLS		2SLS					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education	0.071	0.067	0.102	0.13	0.104	0.108	0.087	0.057
	(.0004)	(.0004)	(0.024)	(0.020)	(0.026)	(0.020)	(0.016)	(0.029)
Exogenous Covariates								
Age (in qtrs)								√
Age (in qtrs) squared								√
9 yr-birth dummies		√			√	√	√	√
50 state-of-birth dummies		√			√	√	√	√
Extra Instruments								
Dummy for QoB=1			√	√	√	√	√	√
Dummy for QoB=2				√		√	√	√
Dummy for QoB=3				√		√	√	√
QoB dummies interacted with year-of-birth dummies							√	√
No covariates k	2	61	2	2	61	61	61	63
No instruments m	2	61	2	4	61	63	93	95

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the the Angrist and Krueger (1991) 1980 Census sample. This sample includes native-born men, born 1930-1939, with positive earnings and non-allocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses. QoB denotes quarter of birth.

Figure 2: 2SLS estimates of economic returns to schooling

Inflated standard errors : $0.\frac{0.021}{0.0005} = 42$, even though the estimate is significant due to a large sample size, the 9% CI is large.

Problem also of a small $R^2_{x,z}$: the *instruments are weak*, it might be better to use OLS instead of IV including more instruments and covariates

- Reduces SE but comes to the cost of potentially having a weak instrument
- col 3 : just identified 1 instrument
- col 4 : over identified (3 QoB instruments)
- col 5/6 : + 59 covariates to 3/4 : m - k = 0 (or 2 resp)
- col 7 : + 30 Ifs (m-k =32)
- col 8 : + age and age^2 to x and z ($m - k = 32$)

But there is potential to test for over identifying restrictions, using the Sargon test

Why β larger? Asymptotic variance depends on error variance, depends on the $R^2_{x,z}$ from first stage regression (x on instruments, will be higher if instruments highly relevant and vice versa).

Exercise 2. Consequences of weak instruments

1. High SE
2. slight violation of exclusion restrictions leads to large bias

Exercise 3. Discuss intuition behind Hausman Test Exogenous regressor then both (testing for whether x endogenous) OLS and IV consistent (if we have valid instrument) If we have an endogenous regressor and OLS is not consistent, difference does not converge to 0 any more, test start follows chi-squared distribution with k degrees of freedom, to test for endogeneity 2 main assumptions, 1 test for, 1 assume

- relies on valid instrument, otherwise $\hat{\beta}$ would not converge at all, this is almost critical assumption in Hausman test

Tutorial 1. IV and simultaneity bias we have expression for y_1 and y_2, we are going to replace this equation, since our asymptotic bias will sum to .? asymptotic bias $= \frac{Cov(u_1, y_2)}{V(y_2)}$ Then we plug long covariance into $Cov(u_1, y_2)$ We know $cov(u_1, u_2) = 0$ and $cov(u_1, z_2) = 0$, but the problem is the variance is typically positive but depending on assumptions we can determine direction of bias based upon α_2 We show that in the formula we replace by y_1

Tutorial 2. IV is asymptotically unbiased that is $plim\ \tilde{\alpha_1} = \alpha_1$ (the IV estimator)

Tutorial 3. why can z_2 not be used as an instrument for y_1 to estimate α_2 the slope of the supply curve It is under-identified, we don't have an endogenous variable at our disposal, we don't have a shock for $y_1 \rightarrow$ we don't have an instrument, two endogenous variables require 2 instruments

Tutorial 4. Exercise 1 (tut3) Regression of log wage on education, estimate using OLS, internet, do you expect OLS to be trustworthy? Education on wage includes ability and motivation etc explaining the wages, that are correlate with education $\rightarrow$ OVB. We expect a positive omitted variable bias, since ability is likely correlated with log wages Testing relevance condition by FS regression : running education on number of siblings, this is significantly different from 0 and f-stat $>$10 Then running IV regression, we find the instrument has strong enough F-stat, but it could be that the exclusion restriction is violated, before we found coefficient of 0.059, with IV we find 0.122 (12%), which is higher than OLS, revealing inconsistency already, perhaps our assumption about *exogeneity* is not fulfilled.

Tutorial 5. Exercise 2 using sibs as iv is not same as plugging sibs into education (as in proxy), we find very different result from our IV estimator, that is big diff from 0 controlling for. Education and birth quarter negatively correlated? b) c) again, we get an increase than the OLS estimator, and larger than when we used siblings as IV. But do we have similar concerns now using birth order Is birth order endogenous? Like the number of siblings? The decision to have children might be related to budget constraints etc. d) identification assumption $\log(wage) = \beta_0 + \beta_1$ test whether π_2 is significantly different from 0, if we estimate our IV, we need to include all exogenous variables as instruments, we estimate a different coefficient.

did

Lecture 5: Causal Inference

Tue 27 Feb 16:02

[Lecture PDF] [L5-LATE-BLoom]

Motivation : Program Evaluation

Binary treatment on a set of outcomes, lets say the effect of having internet at home on school grades, however this would of course run into selection bias since the decision to have internet at home might depend on other unobserved factors (income etc).

Thus, program evaluation is often about how to overcome the problem of selection bias, using the potential outcomes framework allows us to illustrate this.

Exercise 4. Do hospitals make people healthier? Q : Do hospitals make people healthier? If we have data on the following questions :

1. In the last 12 months have you spent a night in hospital?
2. what would you rate you health 1-5 (being excellent)

Group	Sample Size	Mean Health	SE
Hospital	7,774	3.21	0.014
No hospital	90,049	3.93	0.003

Figure 3: Naive Hospital Comparison

This naive comparison of individuals hospitalised and not, a difference of 0.72 suggest that non hospitalised people are healthier Thus, can we ask does going to the hospital make people sick? Maybe in some cases, but the main problem is *self selection*

- People who decide to go to the hospital are less healthy to begin with
- Even if the treatment works, such individuals won't be healthier than those who do not go to the hospital

We can formalise this with the *Potential Outcomes framework*

Treatment Allocation and Outcomes

- start with single unit I
- denote the outcome of interest by Y and treatment variable D
  - D = 1 the individual is *treated*
  - D= 0 the individual is *not treated* (control)
- Typical assumption is that one individual can have 2 states
  1. Y(1) - the potential outcome if I receives treatment
  2. Y(0) - the potential outcome if I 'would not' recieve the treatment (control)
- Individual Causal Effect of the treatment for observation I:

$$Y(1) - Y(0)$$

- The *problem of causal inference* - is that it is impossible to observe **both** potential outcomes at the same time, only one is realised → thus is is impossible to observe the causal effect

Stable Unit Treatment Value Assumption

- generalisation to n units $i = 1, 2, \dots, n$
- Let D_i be the treatment for unit i
- Each unit can be exposed to the two treatments : *the problem is that in principle the potential outcomes can depend on the treatment of all units*

- *thus we make the assumption* that the potential outcome for unit i depends only on the treatment received by unit i and not on the allocation of other individuals
- Denote $D_{-i} = (D_j) : j \neq i$ treatment status of all other individuals in the population. Then SUTVA states
$$[Y_i(1), Y_i(0) \perp D_{-i}$$
- aka the 'no interference assumption'
- However, this might be violated if individuals intact
- There cannot be contagion between individuals

1.5 RCTs - imperfect compliance

Encouragement design

$$Y = Y(0) + D[Y(1) - Y(0)]$$
$$D = D(0) + Z$$

$D = 1$AT + C $D = 0$

LATE

Let $Y_i(z, d)$ be the potential outcome for indovidual i with treatment status $D_i = d$ and the assignment $Z_i - z, z, d, \in \{0, 1\}$

1. Independence $[Y_i(z, d)\forall d, z, D_i(1), D_i(0)] \perp Z_i$
2. Exclusion Restriction $Y_i(d, 0) = Y_i(d, 1) = Y_i(d)$
3. First Stage $\text{P}(D_i = 1|Z_i = 1) - \text{P}(D_i = 1|Z_i = 0) > 0$
4. Monotonicity $D_1(1) - D_i(0) \geq 0$ for all I

Exclusion restriction in RCTs can be violated.

FS estimator is equivalent to weld in IV, gives share of compliers in this case since under monotonicity, we assume everybody reacts to treatment in same way - ruling out the existence of defiers.

So the difference between Z = 0, D=1 = D+AT and Z = 1, D=1 = AT + C is

Monotonicity is rather harsh, it has to hold for all individuals. Randomising vaccine may not give to those who want, thus they are defiers.

Monotonicity - it can only be positive in binary world, if end up having defiers fg

Estimand

1.6 Late and Bloom Result

Exercise 5. [**Late-bloom**] Starting from monotonicity, we can write total variation of binary variable, condition this equal to 1 plus same thing conditional on 0, then this term doesn't exist any more, we get rid and our numerator is equal to The switching function exists only due to the exclusion restriction, otherwise we would need to write y as a function observed by both variables **The denominator** - same thing but for FS, replace now here

IV estimator provides more meaningful interpretation, provides average treatment effect for average treated people We take the numerator again Replace the observed outcome with switching equation via the switching equation, then replace switching equation since the exclusion restriction holds We also know only the non-treated PO is realised, by independence since D_i is binary

This is a general framework for a binary IV estimator too. The LATE/ATT is obtained.

Exercise 6. Why is administrative data good (esp for RCTs) In RCT can have access to social security info if firms agree to connect, can obtain nationality of recent hires etc - why is this useful?

Attrition - have reliable data, don't have to trust accounts of data - say measurement error, can undertake longer term RCTs too