

BS2248: Topic 4 Instrumental Variables (IV)

David Morris

Economics, Finance and Entrepreneurship

d.morris5@aston.ac.uk

Today's Plan

- ① Intuition behind Instrumental Variables (IV)
- ② When is IV used?
- ③ Nuts and Bolts
 - Assumptions that underpin IV
 - Carrying out IV in STATA
- ④ Look at a famous example of IV

OLS is not always appropriate

In these cases other techniques can help us to gain a reliable estimate

While these fancy techniques are useful in some situations, they also have more assumptions underlying them

There is always a trade-off

- No free lunch in economics!

Why IV?

Suppose we want to look at the returns to education...

Classic way to estimate wages is the Mincer equation is:

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{education} + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + \varepsilon$$

Is it reasonable in this to expect the underlying assumptions of OLS to hold?

In particular we are assuming:

$$E(\varepsilon \mid \text{education, experience}) = 0$$

In words, we are assuming the value of the error term is independent of education and experience

This will not be the case if there are unobserved determinants of wages that are correlated with how much schooling someone gets

Are there?

- Ability/ Intelligence?
- Work ethic?
- Health?
- Social Network?

If these play a role in determining health and education and we do not control for them

$E(\varepsilon \mid \text{education, experience}) = 0$ will **NOT** be true !

- The education variable would be endogenous and the result estimate would be biased

Endogenous?

Just said that education would be endogenous

What does endogenous mean?

Endogenous

Endogeneity problems occur when an explanatory variable is correlated with the error term

- ① Can be an uncontrolled confounder causing both independent and dependent variables of a model (the education case)
- ② or a loop of causality between the independent and dependent variables of a model (productivity)

Opposite is exogenous

Back to the problem....

So just control for them right?

Very hard to measure some of these things!!!

- *Ability/ Intelligence?*
- *Work ethic?*
- *Health?*
- *Social Network?*

If we can't measure them...they are taken into account in the error term but that will break the aforementioned assumption of OLS

IV allows us to account for this endogeneity and provide an accurate estimate of the returns of schooling!

Learning Outcomes

Understand why we are using models...

- ① Problem with OLS that endogenous variables will mean that one of Gauss-Markov assumptions does not hold
 - $E(\varepsilon \mid \text{education, experience}) = 0$
- ② OLS estimates will be biased!
- ③ IV can help us remove this problem **IF PROPERLY IMPLEMENTED**

IV

Problem if education is correlated with the error term

So rather than use OLS, we want to just use the part of education the education variable that is not correlated with our error term, ε

$$\log(\text{wage}) = \beta_0 + \beta_1(\text{educ correlated} + \text{educ uncorrelated}) + \beta_2\text{exper} + \beta_3\text{exper}^2 + \varepsilon$$

Rather than the above what we want is:

$$\log(\text{wage}) = \beta_0 + \beta_1\text{educ uncorrelated} + \beta_2\text{exper} + \beta_3\text{exper}^2 + \varepsilon$$

Removed the educ correlated term

Obviously assuming there is part of education which is not correlated with the error term.....

How to do this?

One way to do this is to use an experiment

- Make one half of the sample get an extra year of education and compare the difference in the two samples
- *Not always possible to run an experiment though...*
 - Money
 - Ethics

With IV we are trying to find something that naturally allows us to separate these aspects out

- Instrument for the endogenous variable!

Instrumenting

Key is to instrument the variable that is causing problems (in this case education)

Instrument

A variable that determines the endogenous regressor (education in this case!)

But only affects the dependent variable (wage) through its effect on the independent variables

This means the instrument, Z , should hold the following property:

Z is not correlated with the error term (so that $E(Z | \varepsilon) = 0$)

More formally:

Given the below model,

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + \varepsilon$$

We believe $\text{Cov}(\text{Educ}, \varepsilon) \neq 0$

So we need an instrument, Z , that has the following properties:

- 1 Z is uncorrelated with ε , so that $E(Z | \varepsilon) = 0$
- 2 Z is strongly (or at least weakly) correlated with X (education in this case)
- 3 Z does not have a direct impact on y (it is not an explanatory variable)

Examples of instruments

In simple terms we want an instrument that is highly correlated with education....but not correlated with wages

Examples

- **Distance to school (Card, 1995)**
 - Likely to be correlated with education?
 - Likely to be directly correlated with wage?
 - If labour markets differ around colleges maybe not great.....
- **Quarter of birth (Angrist & Krueger, 1991)**
 - Likely to be correlated with education?
 - Depending on when you are born determines how old you are in the class
 - Oldest kids in the class have traditionally been able to drop out of school at 16, which means this is before the end of their schooling, an option others do not have
 - Likely to be directly correlated with wage?

How does IV work?

While we said earlier that education is composed of two parts, it is better to think of it as being composed of three:

- ① One part is correlated with the error term (troublesome part)
- ② One part is correlated with the instrument (distance to nearest college for example) that is independent of ε
- ③ One part (that we are not concerned with!) that is not correlated to either!

We are just interested in number 2

OLS won't give us number 2, it will give us all three, and thus $E[\hat{\beta}_{1OLS}] \neq \beta_1$ as education is not independent of ε

Use something called 2 Stage Least Squares (2SLS)

Two steps

Step 1: Regress endogenous variable on the instrument and the other variables:

$$\text{Educ} = \gamma_0 + \gamma_1 Z + \gamma_2 \text{exp} + \gamma_3 \text{exp}^2 + v$$

This gives:

$$\hat{\text{Educ}} = \hat{\gamma}_0 + \hat{\gamma}_1 Z + \hat{\gamma}_2 \text{exp} + \hat{\gamma}_3 \text{exp}^2$$

Z , our instrument, partially determines education and thus should give an unbiased estimate of education (should only measure part 2)

$\hat{\text{Educ}}$ is uncorrelated with ε

Step 2: Use the estimated education value, $\hat{\text{Educ}}$, in place of the endogenous education variable

This gives:

$$\log(\text{Wage}) = \beta_0 + \beta_1 \hat{\text{educ}} + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + v$$

$\hat{\beta}_1$ should now be an unbiased estimate of β_1

- *As our education estimate ($\hat{\text{educ}}$) should be independent of the error term, v*

STATA Implementation



STATA Implementation

Step 1:

Regress endogenous variable on the instrument and the other variables:

```
reg educ z exper expersq
```

Step 2:

Use the estimated education value, $\hat{E}duc$, in place of the endogenous education variable

- First need to predict $\hat{E}duc$

```
predict educhat
```

```
reg lwage educhat exper expersq
```

STATA Implementation

STATA has made things a little easier now, and there is a command which will do both steps for you

```
ivregress 2sls lwage exper expersq (educ = proximitytoschool)
```

- Where the element in bracket tells STATA the endogenous variable (educ) and the instrument (proximitytoschool)

Should give the same result as doing this manually

If you struggle with the ivregress command check the help file

- In STATA type: `help ivregress`

More than one instrument?

Typically hard to find an instrument that makes sense

Sometimes we can find two however....

- In this case can use one, or the other, or both
- Best method would be to use both!
 - as long as they both fit the criteria

```
ivregress 2sls lwage exper expersq (educ = one two)
```

Model now said to be over-identified

Testing Instruments



Testing instruments?

In order for a variable, Z , to serve as a valid instrument for endogenous X term the following must be true:

- ① Z is uncorrelated with ε , so that $E(Z | \varepsilon) = 0$
- ② Z is strongly (or at least weakly) correlated with X

Validity of Instruments

We have to use common sense and economic theory to decide if it makes sense to assume $E(Z | \varepsilon) = 0$

- Can't directly test this condition because don't have unbiased estimator for ε
- OLS estimator of ε is presumed biased and the IV estimator of ε depends on the validity of the instrument...

We can test the second assumption though

Testing instruments?

To check the second assumption:

Just need to check that our instrument is significant in the first step of the 2SLS regression

- If it is significant then we know it is strongly related to the endogenous regressor of interest

With two instruments need to test they are both jointly significant....

In STATA: `test instrument1 instrument2`

Testing for endogeneity

If our results from the 2SLS estimates are very different from the OLS estimates it suggests that the OLS estimates are biased

A test for whether the differences between the OLS and 2SLS estimates are statistically significant can be carried out as follows:

- Regress the endogenous variable on all exogenous variables
- Compute the residuals $\hat{v}$ from this regression
- Estimate the model for the dependent variable by OLS, including $\hat{v}$ as an explanatory variable
- Test whether the coefficient for $\hat{v}$ is significantly different from zero using a t-test

Testing for Endogeneity

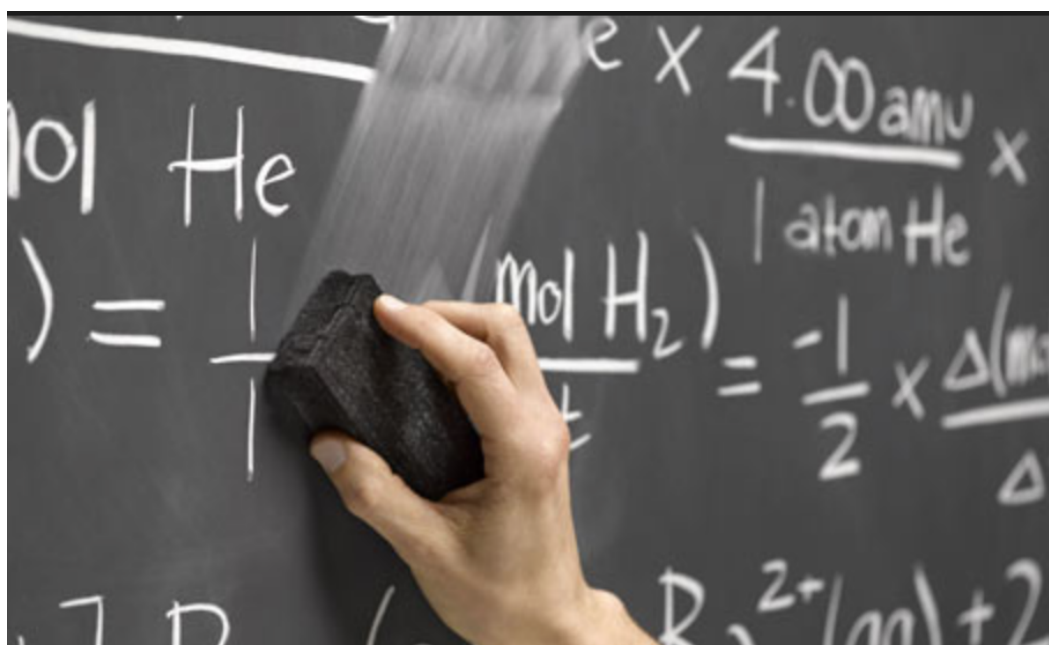
```
regress endogenousvariable instr1 instr2  $\bar{X}$ 
```

```
predict vhat, res
```

```
regress y endogenousvariable vhat  $\bar{X}$ 
```

If the coefficient for $\hat{v}$ is significantly different to zero then it is possible to reject the null that the variable is exogenous

Published Example



Published Example

The example we have been discussing so far is from [Card \(1993\)](#), *“Using geographic variation in college proximity to estimate the return to schooling”*, NBER Working Paper No. 4483.

So we may as well continue with it....

- Again the dataset is on Blackboard if you want to look at it yourself

So the problem is estimating the returns to education, when the education variable is likely to be endogenous

[Card uses proximity to a college as an instrument for education](#)

- ① Likely to be correlated with education levels, as living closer reduces the cost of going
- ② Hopefully having a college is not related to the wage in the given area

[While we have already mentioned this instrument there are other useful instruments that have been used in the literature:](#)

- Mothers education
- Fathers education
- Number of siblings

Card looks at a sample of 3010 working men aged between 24 and 34

Data comes from the 1976 wave of the US National Longitudinal Survey of Young Men

Key Variables:

- The dataset includes a measure of the log of hourly wages in 1976 (**lwage**)
- A measures of years of education (**educ**)
- Years of labour market experience (**exper**)
- The square of years of labour market experience (**expersq**)

Other Control Variables:

- A dummy variable equal to one if the individual is married (**married**)
- A dummy variable equal to one if the individual is black (**black**)
- A dummy variable equal to one if the individual lives in the south in 1976 (**south76**)
- A dummy variable equal to one if the individual lives in a standard metropolitan statistical area in 1976 (**smsa**)
- A dummy variable equal to one if the individual lived in a standard metropolitan statistical area in 1966 (**smsa66**)
- And nine regional dummy variables indicating where the individual lived in 1966 (**reg661 - reg669**)

Instrument:

A dummy variable equal to one if the individual lived close to a college that offered 4 year courses in 1966 (**nearc4**)

Estimation:

The first thing to do is run a standard OLS regression and then we can compare our other results to this

```
regress lwage educ exper expersq black south smsa smsa66 reg661
reg662 reg663 reg664 reg665 reg666 reg667 reg668
```

This should reproduce Table 2, column 2, of Card's paper

OLS Regression

Source	SS	df	MS	Number of obs = 3010		
Model	177.695591	15	11.8463727	F(15, 2994) = 85.48		
Residual	414.946054	2994	.138592536	Prob > F = 0.0000		
Total	592.641645	3009	.196956346	R-squared = 0.2998		
				Adj R-squared = 0.2963		
				Root MSE = .37228		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0746933	.0034983	21.35	0.000	.0678339	.0815527
exper	.084832	.0066242	12.81	0.000	.0718435	.0978205
expersq	-.002287	.0003166	-7.22	0.000	-.0029079	-.0016662
black	-.1990123	.0182483	-10.91	0.000	-.2347927	-.1632318
south	-.147955	.0259799	-5.69	0.000	-.1988952	-.0970148
smsa	.1363845	.0201005	6.79	0.000	.0969724	.1757967
smsa66	.0262417	.0194477	1.35	0.177	-.0118905	.0643739

Cut it off before the region dummies but they are included!

OLS Results

The results from this regression show that:

- ① The education coefficient is positive and significant at the 5% level
 - Suggesting that increased education increases your wage
- ② The experience coefficient is positive and significant at the 5% level
 - Suggesting that more experienced individuals earn higher wages
- ③ The experience squared coefficient is negative and significant at the 5% level
 - Suggesting the returns to experience are decreasing

IV Regression

Let's do this by hand first so we can check it against STATA's command

```
reg educ exper expersq black south smsa smsa66 reg661 reg662  
reg663 reg664 reg665 reg666 reg667 reg668 nearc4
```

```
predict eduhat
```

```
reg lwage eduhat exper expersq black south smsa smsa66 reg661  
reg662 reg663 reg664 reg665 reg666 reg667 reg668
```

IV Regression (Manual)

Source	SS	df	MS	Number of obs = 3010		
Model	115.379173	15	7.69194486	F(15, 2994) = 48.25		
Residual	477.262472	2994	.159406303	Prob > F = 0.0000		
				R-squared = 0.1947		
				Adj R-squared = 0.1907		
Total	592.641645	3009	.196956346	Root MSE = .39926		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educat	.1315037	.0565104	2.33	0.020	.0207006	.2423067
exper	.108271	.0243243	4.45	0.000	.060577	.1559651
expersq	-.0023349	.0003429	-6.81	0.000	-.0030072	-.0016626
black	-.1467759	.0554166	-2.65	0.008	-.2554344	-.0381174
south	-.1446715	.0280524	-5.16	0.000	-.1996755	-.0896676
smsa	.1118084	.032553	3.43	0.001	.0479799	.1756368
smsa66	.0185311	.0222167	0.83	0.404	-.0250303	.0620926

IV Regression

`ivregress 2sls lwage (educ=nearc4) exper expersq black south smsa smsa66 reg661 reg662 reg663 reg664 reg665 reg666 reg667 reg668`

Instrumental variables (2SLS) regression				Number of obs = 3010		
				Wald chi2(15) = 769.20		
				Prob > chi2 = 0.0000		
				R-squared = 0.2382		
				Root MSE = .3873		

lwage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
educ	.1315038	.0548174	2.40	0.016	.0240637	.238944
exper	.1082711	.0235956	4.59	0.000	.0620246	.1545176
expersq	-.0023349	.0003326	-7.02	0.000	-.0029868	-.001683
black	-.1467757	.0537564	-2.73	0.006	-.2521364	-.0414151
south	-.1446715	.027212	-5.32	0.000	-.1980061	-.0913369
smsa	.1118083	.0315777	3.54	0.000	.0499171	.1736995
smsa66	.0185311	.0215511	0.86	0.390	-.0237082	.0607704

Looks the same to me! - Standard errors are a bit tighter though

IV Results

- ① The education coefficient is positive and significant at the 5% level
 - Suggesting that increased education increases your wage
 - This coefficient is larger than before suggesting we were underestimating the impact of education in an OLS model
- ② The experience coefficient is positive and significant at the 5% level
 - Suggesting that more experienced individuals earn higher wages
- ③ The experience squared coefficient is negative and significant at the 5% level
 - Suggesting the returns to experience are decreasing

Further Testing...

- ① From the first regression we can see that our instrument is significant
- ② Cannot test for overidentifying restrictions as we only have one instrument in this case
- ③ Can test for endogeneity though.....
 - `regress educ nearc4 exper expersq black south smsa smsa66 reg661 reg662 reg663 reg664 reg665 reg666 reg667 reg668`
 - `predict vhat, res`
 - `regress lwage educ vhat exper expersq black south smsa smsa66 reg661 reg662 reg663 reg664 reg665 reg666 reg667 reg668`

Further Testing...

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.1315038	.0526906	2.50	0.013	.0281904	.2348172
exper	.1082711	.0226801	4.77	0.000	.0638008	.1527413
expersq	-.0023349	.0003197	-7.30	0.000	-.0029618	-.0017081
black	-.1467758	.0516708	-2.84	0.005	-.2480896	-.045462
south	-.1446715	.0261562	-5.53	0.000	-.1959575	-.0933855
smsa	.1118083	.0303526	3.68	0.000	.0522943	.1713223
smsa66	.0185311	.0207149	0.89	0.371	-.0220858	.0591481
reg661	-.1078142	.0400844	-2.69	0.007	-.1864101	-.0292184
reg662	-.0070465	.0315464	-0.22	0.823	-.0689012	.0548083
reg663	.0404445	.0304662	1.33	0.184	-.0192924	.1001814
reg664	-.0579172	.0360507	-1.61	0.108	-.1286037	.0127694
reg665	.0384577	.0449975	0.85	0.393	-.0497715	.1266868
reg666	.0550887	.0504819	1.09	0.275	-.0438941	.1540715
reg667	.026758	.0468093	0.57	0.568	-.0650238	.1185397
reg668	-.1908912	.0486141	-3.93	0.000	-.2862117	-.0955708
vhat	-.0570621	.0528071	-1.08	0.280	-.1606039	.0464798
_cons	3.773966	.8962812	4.21	0.000	2.016577	5.531356

what is not statistically different to zero, we would conclude that educ is not endogenous

Learning Outcomes

- ① Be able to review the limitations of simple liner regression analysis and the situations where it is not the most appropriate method to use;
 - Why might OLS not give us an unbiased estimate of some relationships?
- ② Have developed an awareness of which models are appropriate to use when an OLS regression may not be robust, and be able to apply these models correctly;
 - How can IV be used to fix this?
- ③ Command an experienced understanding of the empirical approach to economics, and be able to undertake a robust piece of econometric analysis;
 - Estimating these models (chance in the coursework and labs)
 - Do you know what the results mean? Can we test if the model has been applied correctly?
- ④ Be able to explain the covered econometric concepts clearly in concise non-technical language

READING

Have a look at the introductory notes to IV as well as the simulation notes. I have uploaded the Card paper we worked through today as well as another very good piece by two famous economists, Angrist and Krueger

Your questions please....

Glossary

Endogeneity - Occurs when an explanatory variable is correlated with the error term. Normally due to an unobserved variable impacting both the dependent and an independent variable or due to the being a reverse causality link between the dependent and independent variables

Instrument - A variable that determines the endogenous regressor but does not have a link with the dependent variable