Instrumental Variables

Martin J. Conyon

October 2015

Lancaster University

Readings

- Angrist, J.and Krueger, A. (2001) *Instrumental Variables and the Search for Identification*, Journal of Economic Perspectives, Fall .
- Angrist, J and Pischke, J.S. (2009) *Mostly Harmless econometrics*, Princeton University Press.
- See also Steve Pischke's lectures on labor economics for research students – these notes are based on them...

The problem

- The basic problem is to estimate a model with an endogenous variable. The classic example is earnings (y) and education (S) where ability (A) is not observed
- Long form model: $y_i = \alpha + \rho S + \gamma A + \epsilon_i$
- Estimate structural model: $y_i = \alpha + \rho S + e_i$, and so $e_i = \gamma A + \epsilon_i$
- Note: $cov(S, \epsilon) \neq 0$. Estimation leads to OVB.
- IV solution is to find a variable Z_i such that $cov(Z, \epsilon) = 0$ & $cov(Z, S) \neq 0$

Instrument validity

Conditions for a good instrument

1. Random assignment of Z_i (at least as good as)
2. Exclusion restriction: $cov(Z, \epsilon) = 0$. The instrument is not correlated with the disturbance in the long form model
3. Instrument validity: $cov(Z, S) \neq 0$ The instrument is correlated with the endogenous variable.

Numbers 1 and 2 cannot be tested but argued from institutional context. 3 can be tested.

Causal models

Three causal models:

1. The effect of Z_i on S_i
2. The effect of Z_i on y_i
3. The effect of S_i on y_i

Ultimately, we're interested in model 3

Comment

Model estimation:

1. Effect of Z_i on S_i $\Rightarrow$ randomization & instrument validity
2. Effect of Z_i on y_i $\Rightarrow$ randomization & instrument validity
3. Effect of S_i on y_i $\Rightarrow$ randomization, exclusion restriction & instrument validity

Again, model 3 is what we are interested in.

Implementation

The following causal models can be estimated:

1. First stage: $S_i = \pi_{10} + \pi_{11} Z_i + \psi_{1i}$
2. Reduced form: $y_i = \pi_{20} + \pi_{21} Z_i + \psi_{1i}$
3. Structural equation: $y_i = \alpha + \rho S_i + \epsilon_{1i}$

Interested in identifying the causal effect ρ.

Implementation

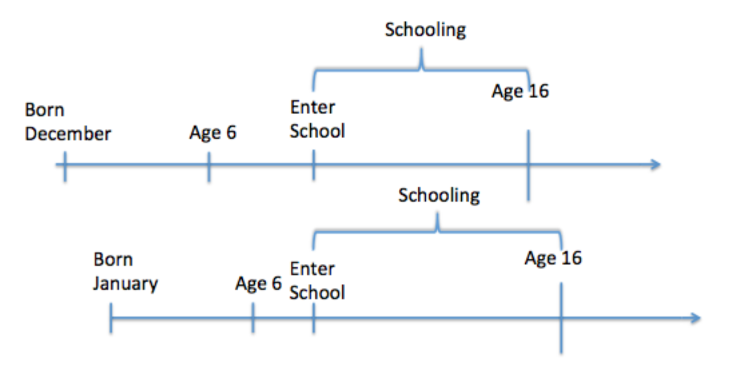
Substitution of first stage into the structural model:

- $y_i = \alpha + \rho S_i + \epsilon_{1i}$
- $y_i = \alpha + \rho[\pi_{10} + \pi_{11} Z_i + \psi_{1i}] + \epsilon_{1i}$
- $y_i = \alpha + \rho\pi_{10} + \rho\pi_{11} Z_i + \rho\psi_{1i} + \epsilon_{1i}$
- $y_i = (\alpha + \rho\pi_{10}) + \rho\pi_{11} Z_i + (\rho\psi_{1i} + \epsilon_{1i})$
- $y_i = \pi_{20} + \pi_{21} Z_i + \epsilon_{2i}$
- where $\pi_{20} = (\alpha + \rho\pi_{10})$ and $\pi_{21} = \rho\pi_{11}$
- $\Rightarrow \frac{\pi_{21}}{\pi_{11}} = \rho$
- $\Rightarrow$ causal effect in the structural eq. (ρ) is the ratio of the coefficient on the reduced form model divided by the coefficient from the first stage

Example

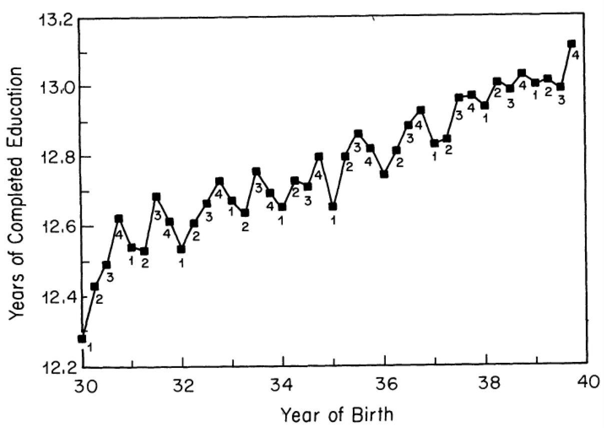
- Angrist & Krueger (1991) Does compulsory school attendance affect schooling and earnings. QJE.
- How does this work? School districts require children to have turned 6 years of age in the year that they enter the school system. Legal dropout rate is exactly 16.
- So, people born earlier in the year will be older when they enter school.
- They will also have fewer years of formal education
- Birth date therefore can be used as a legitimate instrument for years of schooling which is uncorrelated to earning since it is random....

Schooling and birth-quarter



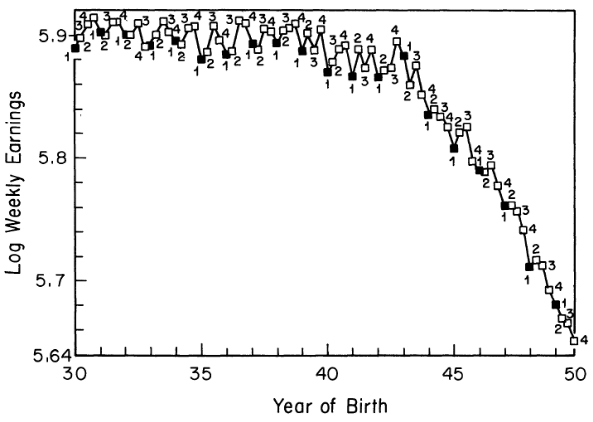
Source: from J. Angrist data

Completed schooling and birth-quarter



Source: from J. Angrist data

Earnings and birth-quarter



Source: from J. Angrist data

Data

- Angrist & Krueger (1991) use 1980 data from the US census.
- The data set contains information on 329,509 men born between 1930 and 1939.
- The men are in the 40s when the sample is taken
- The data set contains information on 1979 earnings (logged), birth quarter, years of schooling
- Data available from Josh Angrist's data archive....

Birth quarter statistics

birth_quarter	Freq.	Percent	Cum.
1	81,671	24.79	24.79
2	80,138	24.32	49.11
3	86,856	26.36	75.47
4	80,844	24.53	100.00
Total	329,509	100.00	

Source: from J. Angrist data

Schooling and birth quarter

Variables	(1) school	(2) school
q2	0.057***	
	(0.016)	
q3	0.117***	
	(0.016)	
q4	0.151***	0.092***
	(0.016)	(0.013)
Constant	12.688***	12.747***
	(0.011)	(0.007)
Observations	329,509	329,509
R-squared	0.000	0.000

Standard errors in parentheses

Earnings and schooling

Variables	(1) lwage (OLS)	(2) lwage (IV)	(3) lwage (IV)
school	0.071***	0.074***	0.103***
	(0.000)	(0.028)	(0.020)
Constant	4.995***	4.955***	4.590***
	(0.004)	(0.358)	(0.249)
Observations	329,509	329,509	329,509
R-squared	0.117	0.117	0.094

Standard errors in parentheses

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Instruments: Col 1: OLS; Col 2: IV using Q4 birth; Col 3: IV using Q3 to Q4

Wald estimator

Consider the model: $y = \beta_0 + \beta_1 x + u$ and let z be a binary ($z = 0$, or $z = 1$) instrumental variable for x. Then the IV estimator is:

$$\hat{\beta}_1 = \hat{\beta}_{1IV} = \frac{\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})},$$

The IV estimator $\hat{\beta}_1$ can also be written as:

$$\hat{\beta}_1 = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0),$$

where $\bar{y}_0$ and $\bar{x}_0$ are the sample averages of y_i over the part of the sample with $z_i = 0$, and the terms $\bar{y}_1$ and $\bar{x}_1$ are the sample averages of y_i over the part of the sample with $z_i = 1$. This is known as the Wald (1940) estimator.

Wald estimator: example

We can use the Angrist data to calculate the Wald estimator. Let y_i be log wages, and S_i years of schooling, and Q4 quarter 4 birth

$$\beta_{Wald} = \frac{Cov(y_i, Q4)}{Cov(S_i, Q4)}$$

$$\beta_{Wald} = \frac{E[y_i|Q4 = 1] - E[y_i|Q4 = 0]}{E[S_i|Q4 = 1] - E[S_i|Q4 = 0]}$$

Wald estimates

Calculate the Wald estimator from the first stage and reduced form regressions.

$S_i = \alpha_{10} + \alpha_{14} Q4_i + \theta_{1i} \Rightarrow$ First Stage regression

$y_i = \alpha_{20} + \alpha_{24} Q4_i + \theta_{2i} \Rightarrow$ Reduced form regression

$E[y_i | Q4 = 1] = \alpha_{20} + \alpha_{24}$ and

$E[y_i | Q4 = 0] = \alpha_{20}$

$\Rightarrow E[y_i | Q4 = 1]$-$E[y_i | Q4 = 0] = \alpha_{24}$ and

Wald estimates

- The Wald estimate is the ratio of these two:

$$\beta_{Wald} = \frac{\alpha_{24}}{\alpha_{14}}$$

- In the case of earnings: the Wald estimator is the difference in average earnings across the two groups divided by the difference in average schooling across the two groups

Reduced form

```
. reg lwage q4 // regress wage on Q4

      Source |       SS           df       MS      Number of obs   =   329509
-------------+----------------------------------   F(  1,329507)   =     6.15
       Model |  2.83199415         1  2.83199415   Prob > F        =   0.0132
    Residual |  151835.039   329507  .460794578    R-squared       =   0.0000
-------------+----------------------------------   Adj R-squared   =   0.0000
       Total |  151837.871   329508  .460801774    Root MSE        =   .67882

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          q4 |   .0068132   .0027482     2.48   0.013     .0014267    .0121996
       _cons |   5.898272   .0013613  4332.90   0.000     5.895604     5.90094
------------------------------------------------------------------------------
```

Source: from J. Angrist data

First stage

```
. reg school q4 // regress school on Q4

      Source |       SS       df       MS              Number of obs =  329509
-------------+------------------------------           F(  1,329507) =   48.09
       Model |  517.7393       1   517.7393            Prob > F      =  0.0000
    Residual | 3547149.92329507  10.7650215            R-squared     =  0.0001
-------------+------------------------------           Adj R-squared =  0.0001
       Total | 3547667.66329508    10.76656            Root MSE      =   3.281

------------------------------------------------------------------------------
      school |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          q4 |   .0921209   .0132834     6.94   0.000     .0660857     .118156
       _cons |   12.74731   .0065796  1937.40   0.000     12.73441    12.76021
------------------------------------------------------------------------------
```

Source: from J. Angrist data

Wald Estimate

- Difference in mean earnings = 0.0068
- Difference in mean schooling = 0.0921
- Ratio of the difference = 0.0068 / 0.0921 = 0.074
- Compare this quantity to the IV estimate arising from the 2SLS method – they are the same...

IV estimates

```
. ivregress 2sls lwage (school= q4)

Instrumental variables (2SLS) regression          Number of obs =   329509
                                                  Wald chi2(1)  =     6.96
                                                  Prob > chi2   =   0.0083
                                                  R-squared     =   0.1171
                                                  Root MSE      =   .63785
```

lwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
school	.0739589	.0280328	2.64	0.008	.0190157	.1289021
_cons	4.955495	.3579775	13.84	0.000	4.253872	5.657118

```
Instrumented:  school
Instruments:   q4
```

Source: from J. Angrist data