Instrumental variables estimates of the returns to schooling

Jörn-Steffen Pischke

LSE

October 12, 2018

Two solutions to the ability bias problem

Remember that we would like to run the *long regression*

$$Y_i = \alpha + \rho S_i + \gamma A_i + e_i.$$

- The **regression solution** was to look for variables which can act as proxies for ability A_i. The key to regression is that we need something that captures *all* the variation in A_i, otherwise we are left with OVB.
- The **IV solution** is to isolate variation in S_i which is unrelated to A_i. The variable which does the "isolating" is the instrumental variable. The good thing is that we just need *some* of the variation in S_i; so the requirements on the instrument seem weaker than those on the regression control A_i. But IV comes with other possible complications.

Conditions for a valid instrument

Call the instrumental variable Z_i. A valid instrument needs to satisfy three conditions:

1. Z_i is as good as randomly assigned.
2. Z_i satisfies the exclusion restriction, i.e. it does not appear as a separate regressor in the long regression we like to run.
3. Z_i affects the endogenous regressor S_i.

Of these only condition 3. can be tested. This is the strength of the first stage. Conditions 1. and 2. have to be argued based on knowledge from outside the data we have.

Three causal effects

There are three causal effects we can think about:

1. The causal effect of Z_i on S_i.
2. The causal effect of Z_i on Y.
3. The causal effect of S_i on Y_i.

The last one is the one we are ultimately interested in, the return to schooling ρ.

Some instrumental variables language

The instrumental variables language comes from old style simultaneous equations models but we can think of it as related to the three causal effects.

- First stage: The regression of schooling on the instrument is called the first stage (causal effect number 1)

$$S_i = \pi_{10} + \pi_{11} Z_i + \xi_{1i}$$

- Reduced form: The regression of earnings on the instrument is called the reduced form (causal effect number 2)

$$Y_i = \pi_{20} + \pi_{21} Z_i + \xi_{2i}.$$

- Structural equation: The regression of earnings on schooling is called the structural equation

$$Y_i = \alpha + \rho S_i + \eta_i,$$

where $\eta_i = \gamma A_i + e_i$, i.e. it is a structural error term, not a regression residual. This involves causal effect number 3.

You get what you pay for

- Conditions 1 and 3 on the instrument are enough to get causal effects 1 and 2. I.e. these conditions are sufficient for the first stage and reduced forms to have a causal interpretation. Often, the reduced form coefficient may be interesting in its own right. For example, the instrument might be a policy variable in which case it is the policy effect.
- To get causal effect 3, i.e. the structural parameter ρ, we also need condition 2, the exclusion restriction. This condition is often the most difficult requirement on an instrument. It is distinct from random assignment, so having experimental variation does not guarantee a valid IV interpretation of the estimates.

Linking the three equations

The coefficients in the three equations are linked. Substitute the first stage into the structural equation:

$$\begin{aligned} Y_i &= \alpha + \rho S_i + \eta_i \\ &= \alpha + \rho \left[\pi_{10} + \pi_{11} Z_i + \xi_{1i} \right] + \eta_i \\ &= (\alpha + \rho \pi_{10}) + \rho \pi_{11} Z_i + (\rho \xi_{1i} + \eta_i) \\ &= \pi_{20} + \pi_{21} Z_i + \xi_{2i}. \end{aligned}$$

Hence, the reduced form coefficients are:

$$\begin{aligned} \pi_{20} &= \alpha + \rho \pi_{10} \\ \pi_{21} &= \rho \pi_{11} \end{aligned}$$

Indirect least squares

It is straightforward to see that

$$\rho = \frac{\pi_{21}}{\pi_{11}},$$

i.e. the IV estimate is equal to the ratio of the reduced form coefficient on the instrument to the first stage coefficient. This is called *indirect least squares*.

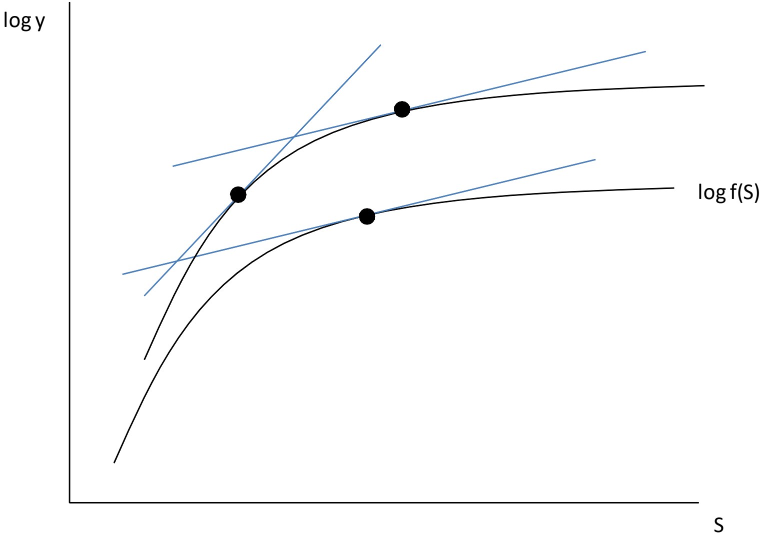
Just identified vs. over-identified models

- Indirect least squares only works when there is one endogenous regressor and one instrument. Such a model is *just identified* (there is only one single solution to get ρ from the first stage and reduced form coefficients).
- If there are multiple instruments for a single endogenous regressor the model is *over-identified*.

$$\begin{aligned} S_i &= \pi_{10} + \pi_{11} Z_{1i} + \pi_{12} Z_{2i} + \xi_{1i} \\ Y_i &= \pi_{20} + \pi_{21} Z_{1i} + \pi_{22} Z_{2i} + \xi_{2i}. \end{aligned}$$

There is no unique way to get ρ from π_{11}, π_{12}, π_{21}, and π_{22}. Two stage least squares (2SLS) is a particular average (it is the optimally weighted GMM estimator for the homoskedastic model).

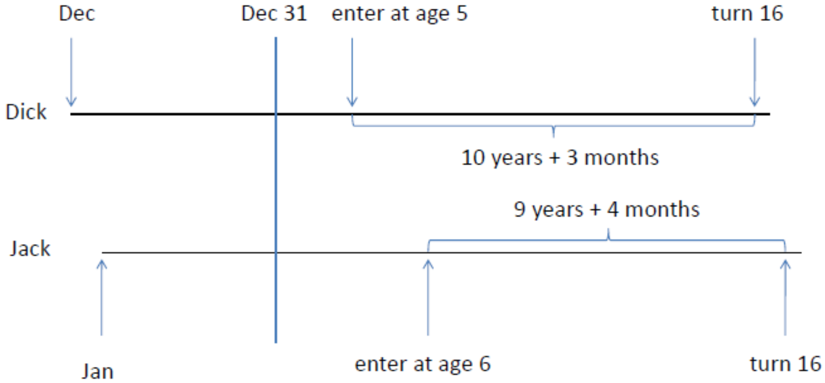
So what's a good instrument?



Angrist and Krueger (1991): US compulsory schooling and school entry rules

- US compulsory schooling laws are in terms of age, not number of years of schooling completed. If compulsory schooling age is 16, you can drop out on your 16th birthday (even if in the middle of the school year).
- School entry is once a year, and cutoffs are based on birthdays.
- The combination of these two generates variation in schooling for those who drop out as soon as they can. This variation depends on birthday within the year.

How compulsory schooling laws and school entry rules interact



Is this a good instrument?

Variation indeed comes from the cost/compulsion side of the schooling problem. Let's check the three conditions

1. Random assignment: Are birthdays random with respect to the counterfactual earnings for different schooling levels?
   - There are small differences in average SES by birthday throughout the year.
2. Do birthdays satisfy the exclusion restriction, or could birthdays affect earnings for other reasons than through their effect on years of schooling?
   - Birthday affects, e.g., age rank in class.
3. Do birthdays indeed affects schooling?
   - Check the first stage.

Angrist and Krueger data

- Data are from the 1980 US Census.
- 329,509 men born 1930 to 1939 (i.e. in their 40s when observed).
- For these men we have year of birth, quarter of birth, years of schooling, and earnings in 1979.

First stage regressions of schooling on quarter of birth

	Dependent variable: schooling			
Regressor	(1)	(2)	(3)	(4)
quarter 2		0.057		0.057
		(0.017)		(0.016)
quarter 3		0.117		0.113
		(0.016)		(0.016)
quarter 4	0.092	0.151	0.091	0.148
	(0.013)	(0.016)	(0.013)	(0.016)
9 year of birth dummies			✓	✓

Reduced form regressions of log wages on quarter of birth

	Dependent variable: log wages			
Regressor	(1)	(2)	(3)	(4)
quarter 2		0.0045 (0.0034)		0.0046 (0.0034)
quarter 3		0.0149 (0.0033)		0.0150 (0.0033)
quarter 4	0.0068 (0.0027)	0.0135 (0.0034)	0.0068 (0.0027)	0.0135 (0.0034)
9 year of birth dummies			✓	✓

IV regressions of log wages on schooling

Coefficient on schooling from an OLS regression: 0.071 (0.0004)

	Dependent variable: log wages			
Regressor	(1)	(2)	(3)	(4)
schooling	0.074 (0.028)	0.103 (0.020)	0.075 (0.028)	0.105 (0.020)
Instruments	Quarter 4	4 quarter dummies	Quarter 4	4 quarter dummies
9 year of birth dummies			✓	✓

IV with a dummy instrument: the Wald Estimator

A way to look at IV with a binary instrument, like the dummy for quarter 4, is the following. With $Q4_i$ a dummy instrument it turns out that

$$\begin{aligned}\beta_{IV} &= \frac{cov\left(\ln Y_i, Q4_i\right)}{cov\left(S_i, Q4_i\right)} \\ &= \frac{E\left[\ln Y_i | Q4_i = 1\right] - E\left[\ln Y_i | Q4_i = 0\right]}{E\left[S_i | Q4_i = 1\right] - E\left[S_i | Q4_i = 0\right]}.\end{aligned}$$

This is called the *Wald* estimator (due to Abraham Wald, 1940).

Constructing the Wald Estimator

The first stage and reduced form are

$$\begin{aligned} S_i &= \pi_{11} + \pi_{14} Q4_i + \xi_{1i}, \\ \ln Y_i &= \pi_{21} + \pi_{24} Q4_i + \xi_{2i}. \end{aligned}$$

Taking expectations conditionally on $Q4_i$ yields

$$\begin{aligned} E\left[\ln Y_i | Q4_i = 1\right] &= \pi_{21} + \pi_{24} \\ E\left[\ln Y_i | Q4_i = 0\right] &= \pi_{21} \end{aligned}$$

Hence, the reduced form coefficients on $Q4_i$ are the differences in group means with the instrument switched on and off:

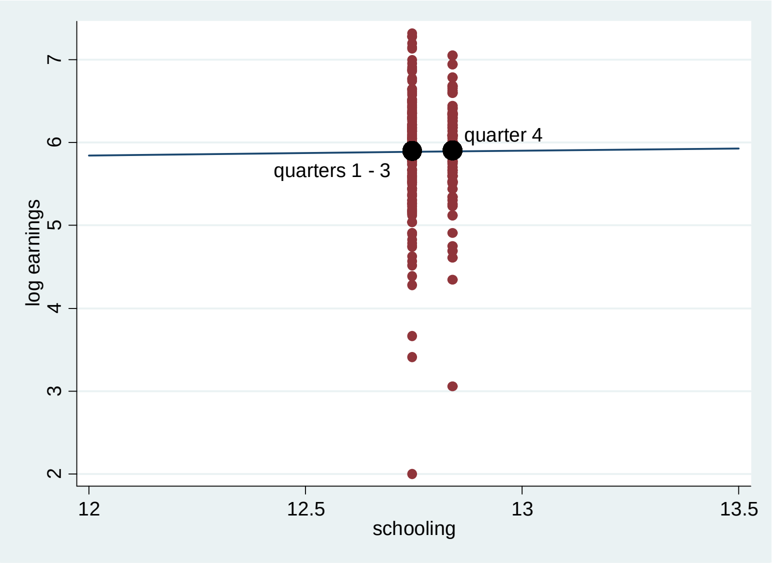
$$\begin{aligned} \pi_{24} &= E\left[\ln Y_i | Q4_i = 1\right] - E\left[\ln Y_i | Q4_i = 0\right] \\ \pi_{14} &= E\left[S_i | Q4_i = 1\right] - E\left[S_i | Q4_i = 0\right] \end{aligned}$$

The IV estimate is the ratio of the two.

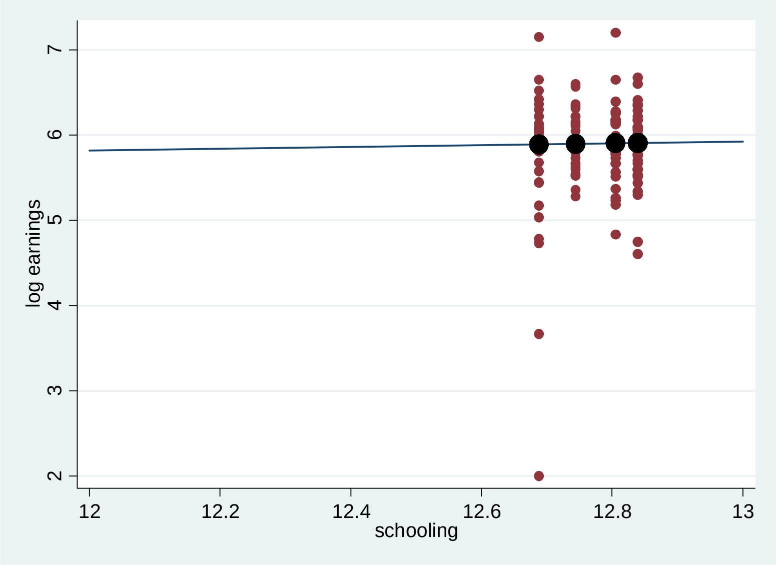
The Wald estimate of the return to schooling

	born quarter 1, 2, 3	born quarter 4	Difference
log earnings	5.8983	5.9051	0.0068 (0.0027)
schooling	12.747	12.839	0.092 (0.013)
Wald estimate			0.074 (0.028)

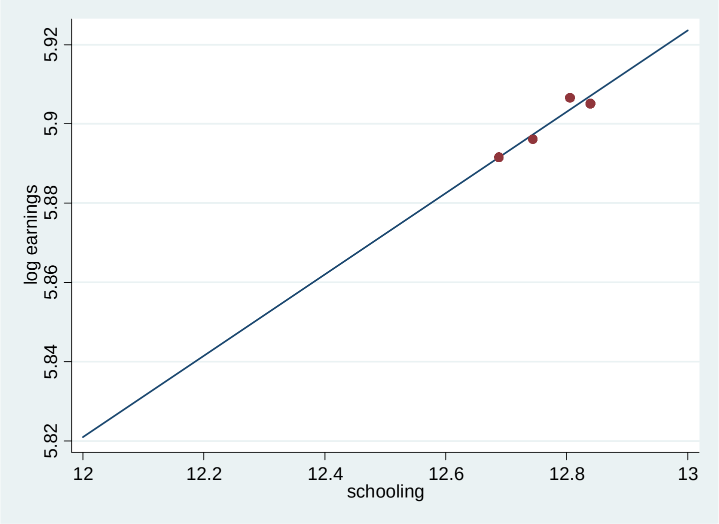
A plot of the Wald estimate of the return to schooling



We can do this with all four quarter dummies



Now the regression line doesn't fit the four means exactly anymore



Grouping is IV

The analogy between IV and the Wald estimator can be extended to IV with multiple dummy instruments. Suppose we have a variable Q_i, for quarter, which takes on four values 1 to 4. Then, using the structural equation

$$\ln Y_i = \alpha + \beta S_i + \eta_i.$$

we get

$$E\left[\ln Y_i | Q_i\right] = \alpha + \beta E\left[S_i | Q_i\right].$$

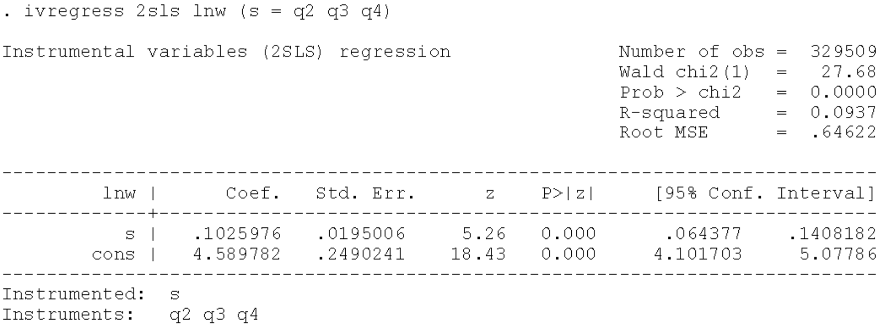
The sample analog of $E\left[\ln Y_i | Q_i = j\right]$ is $\overline{\ln Y}_j$, and similarly with $E\left[S_i | Q_i\right]$. Hence, the IV estimate of log earnings on schooling, instrumenting with a set of mutually exlusive and exhaustive dummy variables is the same as the regression using the group means

$$\overline{\ln Y}_j = \alpha + \beta \overline{S}_j + \overline{\eta}_j$$

weighted by the size of the cells.

So does it work in practice?

Here is 2SLS

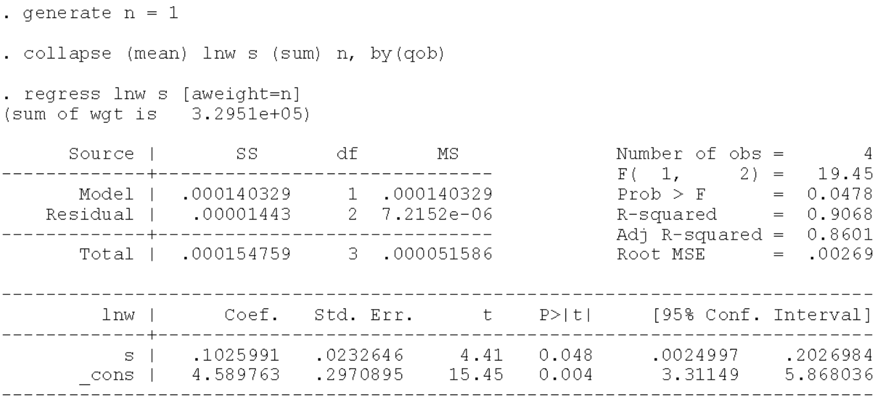


```
. ivregress 2sls lnw (s = q2 q3 q4)

Instrumental variables (2SLS) regression               Number of obs =  329509
                                                       Wald chi2(1)  =   27.68
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.0937
                                                       Root MSE      =  .64622

------------------------------------------------------------------------------
         lnw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |   .1025976   .0195006     5.26   0.000      .064377    .1408182
        cons |   4.589782   .2490241    18.43   0.000     4.101703     5.07786
------------------------------------------------------------------------------
Instrumented:  s
Instruments:   q2 q3 q4
```

And here is grouped data



```
. generate n = 1

. collapse (mean) lnw s (sum) n, by(qob)

. regress lnw s [aweight=n]
(sum of wgt is   3.2951e+05)

      Source |       SS       df       MS              Number of obs =       4
-------------+------------------------------           F(  1,      2) =   19.45
       Model |  .000140329     1  .000140329           Prob > F      =  0.0478
    Residual |   .00001443     2  7.2152e-06           R-squared     =  0.9068
-------------+------------------------------           Adj R-squared =  0.8601
       Total |  .000154759     3  .000051586           Root MSE      =  .00269

------------------------------------------------------------------------------
         lnw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |   .1025991   .0232646     4.41   0.048     .0024997    .2026984
       _cons |   4.589763   .2970895    15.45   0.004      3.31149    5.868036
------------------------------------------------------------------------------
```

A quick derivation

Consider a regression model in matrix notation:

$$\mathbf{y} = \mathbf{x}\beta + \mathbf{e}$$

We have matrix of q exhaustive dummy variable instruments $\mathbf{Z}$. Note that

$$\mathbf{Z} = \begin{bmatrix} \boldsymbol{\iota}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\iota}_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{\iota}_q \end{bmatrix}$$

where $\boldsymbol{\iota}_j$ is a column vector of n_j ones.

A quick derivation

The 2SLS estimator is

$$\beta_{2SLS} = \left[\mathbf{x}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{x}\right]^{-1}\mathbf{x}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{y}.$$

Given the definition of **Z**:

$$\mathbf{Z}'\mathbf{x} = \begin{bmatrix} \overline{x}_1 n_1 \\ \overline{x}_2 n_2 \\ \vdots \\ \overline{x}_q n_q \end{bmatrix}, \mathbf{Z}'\mathbf{y} = \begin{bmatrix} \overline{y}_1 n_1 \\ \overline{y}_2 n_2 \\ \vdots \\ \overline{y}_q n_q \end{bmatrix}, \left(\mathbf{Z}'\mathbf{Z}\right)^{-1} = \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{n_q} \end{bmatrix}$$

and hence

$$\beta_{2SLS} = \frac{\sum_{j=1}^{q} \overline{y}_j \overline{x}_j n_j}{\sum_{j=1}^{q} \overline{x}_j^2 n_j}$$

Aside on grouped data

Discrete regressor: grouping is OLS

Consider a generic regression model

$$Y_i = \alpha + \beta X_i + e_i$$

where X_i is a discrete regressor, taking on J different values. The regression

$$\overline{Y}_j = \alpha + \beta X_j + \overline{e}_j$$

weighted by the cell size is identical to OLS on the micro data (see MHE, section 3.1.2).

Aside on grouped data

Continuous regressor, discrete instrument: grouping is IV

If, on the other hand, X_i is continuous, and we have a discrete instrument, Z_i, which takes on J different values, the IV estimation of the micro data model with $J - 1$ dummy instruments is the same as

$$\overline{Y}_j = \alpha + \beta \overline{X}_j + \overline{e}_j$$

weighted by the cell size.

Aside on grouped data

Why are the two cases different?

- Discrete regressor: the regressor defines the groups. The grouped data regression is OLS because all the variation in X_i is only at the group level. By aggregating Y_i we are not changing anything about X_i.
- Continuous regressor, discrete instrument: now the instrument defines the groups. The grouped data regression is IV because we are changing the variation in X_i by aggregating.