INSTRUMENTAL VARIABLES (Take 1): CONSTANT EFFECTS

Organizing IV

I tell the IV story in two iterations, first with constant effects, then in a framework with heterogeneous potential outcomes.

- The constant effects framework focuses attention on the IV solution for selection bias and on essential IV mechanics
- But first: Why do IV?
  - I can't say "because the regressors are correlated with the errors."
  - As we've seen, regressors are uncorrelated with errors by definition
- The (short) regression of schooling on wages produces residuals uncorrelated with schooling (that's how the good lord made 'em)
- The problem, therefore, must be that the regression you've got is not the regression you want (and that's your fault!)

IV Goes Long

- Suppose the causal link between schooling and wages can be written $f_i(s) = \alpha + \rho s + \eta_i$
- Imagine a vector of control variables, A_i, called "ability"; write

$$\eta_i = A_i'\gamma + v_i$$

where γ is a vector of pop. reg. coefficients, so v_i and A_i are uncorrelated *by construction*

- We'd happily include ability in the regression of wages on schooling, producing this long regression:

$$\mathrm{Y}_i = \alpha + \rho \mathrm{S}_i + A_i'\gamma + v_i \tag{1}$$

The error term here is the random part of potential outcomes, v_i, left over after controlling for A_i

- If $E[\mathrm{S}_i v_i] = 0$, a version of the CIA, the population regression of Y_i on S_i and A_i identifies ρ. That's like saying: "A_i is the *only* reason schooling is correlated with potential outcomes."

IV and OVB

- *IV allows us to estimate the long-regression coefficient, ρ, when A_i is unobserved.*

  The instrument, Z_i, is assumed to be: (1) correlated with the causal variable of interest, S_i; and (2) uncorrelated with potential outcomes
- Here, "uncorrelated with potentials" means $Cov(\eta_i, \mathrm{Z}_i) = 0$, or, equivalently, Z_i is uncorrelated with both A_i and v_i
  - This is a version of the *exclusion restriction*: Z_i can be said to be excluded from the causal model of interest
- Given the exclusion restriction, it follows from equation (1) that

$$\begin{aligned} \rho &= \frac{Cov(\mathrm{Y}_i, \mathrm{Z}_i)}{Cov(\mathrm{S}_i, \mathrm{Z}_i)} = \frac{Cov(\mathrm{Y}_i, \mathrm{Z}_i)/V(\mathrm{Z}_i)}{Cov(\mathrm{S}_i, \mathrm{Z}_i)/V(\mathrm{Z}_i)} \\ &= \frac{\text{"}RF\text{"}}{\text{"}1st\text{"}} \end{aligned} \tag{2}$$

- The IV *estimator* is the sample analog of (2)

Two-stage least squares (2SLS)

- In practice, we do IV by doing 2SLS. This allows us to add covariates (controls) and combine multiple instruments. Returning to the schooling example, a causal model with covariates is

$$\mathrm{Y}_i = \alpha' \mathrm{X}_i + \rho \mathrm{S}_i + \eta_i, \tag{3}$$

where η_i is the compound error term, $A_i\gamma + v_i$. The first stage and reduced form are

$$\begin{aligned} \mathrm{S}_i &= \mathrm{X}_i' \pi_{10} + \pi_{11} \mathrm{Z}_i + \xi_{1i} && (4) \\ \mathrm{Y}_i &= \mathrm{X}_i' \pi_{20} + \pi_{21} \mathrm{Z}_i + \xi_{2i} && (5) \end{aligned}$$

- The reduced form is obtained by substituting (4) into (3):

$$\begin{aligned} \mathrm{Y}_i &= \alpha' \mathrm{X}_i + \rho[\mathrm{X}_i' \pi_{10} + \pi_{11} \mathrm{Z}_i] + \rho \xi_{1i} + \eta_i && (6) \\ &= \mathrm{X}_i' [\alpha + \rho \pi_{10}] + \rho \pi_{11} \mathrm{Z}_i + [\rho \xi_{1i} + \eta_i] \\ &= \mathrm{X}_i' \pi_{20} + \pi_{21} \mathrm{Z}_i + \xi_{2i} \end{aligned}$$

2SLS Notes

- Again, it's all about the ratio of RF to 1st:

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

  In simultaneous squations models, the sample analog of this ratio is called an *Indirect Least Squares* (ILS) estimator of ρ

- Where does *two-stage least squares* come from? Write the first stage as the sum of fitted values plus first-stage residuals:

$$\mathrm{S}_i = \mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i + \xi_{1i} = \hat{s}_i + \xi_{1i}$$

2SLS estimates of (3) can be constructed by substituting first-stage fitted values for S_i in (3):

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \rho\hat{s}_i + [\eta_i + \rho\xi_{1i}], \tag{7}$$

and using OLS to estimate this "second stage" (a version of eq. 6)

- In practice, we let Stata do it: "manual 2SLS" doesn't get the standard errors right

2SLS example: Angrist and Krueger (1991)

- AK-91 argue that because children born in late-quarters start school younger, they are kept in school longer by birthday-based compulsory schooling laws
- There's a powerful first stage supporting this: Schooling tends to be higher for late-quarter births; this is driven by high school and not college, consistent with the CSL story
- The QOB first stage and reduced form are plotted in **Figure 4.1.1**
- The corresponding 2SLS estimates appear in **Table 4.1.1**
  - 2SLS matches the QOB pattern earnings (the RF) to the QOB pattern in schooling (the first stage).
  - The exogenous covariates include year-of-birth and state-of-birth dummies, as well as linear and quadratic functions of age in quarters
- QOB Questioned: Bound, Jaeger, and Baker (1995) and Buckles and Hungerman (2008) argued QOB is correlated with maternal characteristics. **Allowing for this** fails to overturn AK conclusions

2SLS is a many-splendored thing

- 2SLS is the same as IV where the instrument is $\hat{s}_i^*$, the residual from a regression of $\hat{s}_i$ on X_i
- One-instrument 2SLS equals IV, where the instrument is $\tilde{z}_i$, the residual from a regression of Z_i on the covs, X_i
- One-instrument 2SLS equals indirect least squares (ILS), that is, the ratio of reduced form to first stage coefficients on the instrument. In other words,

$$\begin{aligned}\frac{Cov(\mathrm{Y}_i, \hat{s}_i^*)}{V(\hat{s}_i^*)} &= \frac{Cov(\mathrm{Y}_i, \hat{s}_i^*)}{Cov(\mathrm{S}_i, \hat{s}_i^*)} \\ &= \frac{Cov(\mathrm{Y}_i, \tilde{z}_i)}{Cov(\mathrm{S}_i, \tilde{z}_i)} = \frac{\pi_{21}}{\pi_{11}}\end{aligned}$$

- With more than one instrument, 2SLS is a weighted average of the one-at-time (just-identified) estimates (In a linear homoskedastic constant-effects model, this is efficient)

Multi-Instrument 2SLS

- Let

$$\rho_j = \frac{Cov(\mathrm{Y}_i, \mathrm{Z}_{ji})}{Cov(\mathrm{D}_i, \mathrm{Z}_{ji})}; j = 1, 2$$

  denote two IV estimands using Z_{1i} and Z_{2i} to instrument D_i.
- The 2SLS estimand is

$$\rho_{2SLS} = \psi\rho_1 + (1 - \psi)\rho_2,$$

  where ψ is a number between zero and one that depends on the relative strength of the instruments in the first stage.
- Angrist and Evans (1998) use twins and sex-mix instruments
  - Using a twins-2 instrument alone, the IV estimate of the effect of a third child on female labor force participation is -.084 (s.e.=.017). The corresponding samesex estimate is -.138 (s.e.=.029).
  - Using both instruments produces a 2SLS estimate of -.098 (.015).
  - The 2SLS weight in this case is .74 for twins, .26 for samesex, due to the stronger twins first stage.

2SLS Mistakes

2SLS . . . so simple a fool can do it . . .
and many do!

What can go wrong?

- As explained in MHE 4.6.1, three mistakes have yet to be relegated to the dustbin of IV history:
  - Manual 2SLS
  - Covariate ambivalence
  - Forbidden regressions (from the left and the right)
- These can be interpreted as the result of failed attempts to get round hard-wired 2SLS protocols
- Avoid temptation: let Stata do it!

2SLS Lingo

- These terms come to us from simultaneous equations modeling, the intellectual birthplace of IV:
  - *Endogenous variables* are the dependent variable and the independent variable(s) to be instrumented; in a simultaneous equations model, endogenous variables are determined by solving the system
  - To *treat an independent variable as endogenous* is to instrument it, i.e., to replace it with fitted values in the 2SLS second stage
  - *Exogenous variables* include *covariates* (not instrumented) and the excluded instruments themselves. In a simultaneous equations model, exogenous variables are determined outside the system
- In any IV study, variables are either: dependent or (other) endogenous variables, instruments, or covariates
- If you're unsure what's what, or find yourself asking variables to play more than one role . . . seek counseling

The Wald estimator

- How were Vietnam-era vets affected by their service?
- Let D_i indicate veterans. A causal constant-effects model is:

$$\mathrm{Y}_i = \alpha + \rho \mathrm{D}_i + \eta_i, \tag{8}$$

where η_i and D_i may be correlated. B/c Z_i is a dummy,

$$\frac{Cov(\mathrm{Y}_i, \mathrm{Z}_i)}{V(\mathrm{Z}_i)} = E[\mathrm{Y}_i|\mathrm{Z}_i = 1] - E[\mathrm{Y}_i|\mathrm{Z}_i = 0],$$

with an analogous formula for $\frac{Cov(\mathrm{D}_i, \mathrm{Z}_i)}{V(\mathrm{Z}_i)}$. It follows that,

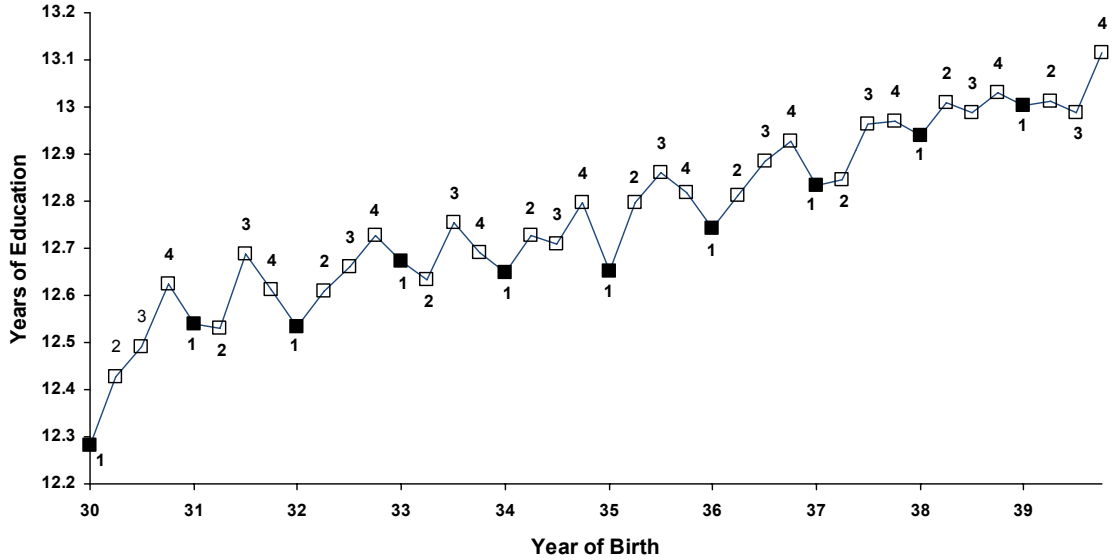
$$\rho = \frac{Cov(\mathrm{Y}_i, \mathrm{Z}_i)}{Cov(\mathrm{D}_i, \mathrm{Z}_i)} = \frac{E[\mathrm{Y}_i|\mathrm{Z}_i = 1] - E[\mathrm{Y}_i|\mathrm{Z}_i = 0]}{E[\mathrm{D}_i|\mathrm{Z}_i = 1] - E[\mathrm{D}_i|\mathrm{Z}_i = 0]} \tag{9}$$

- A direct route uses (8) and $E[\eta_i|\mathrm{Z}_i] = 0$:

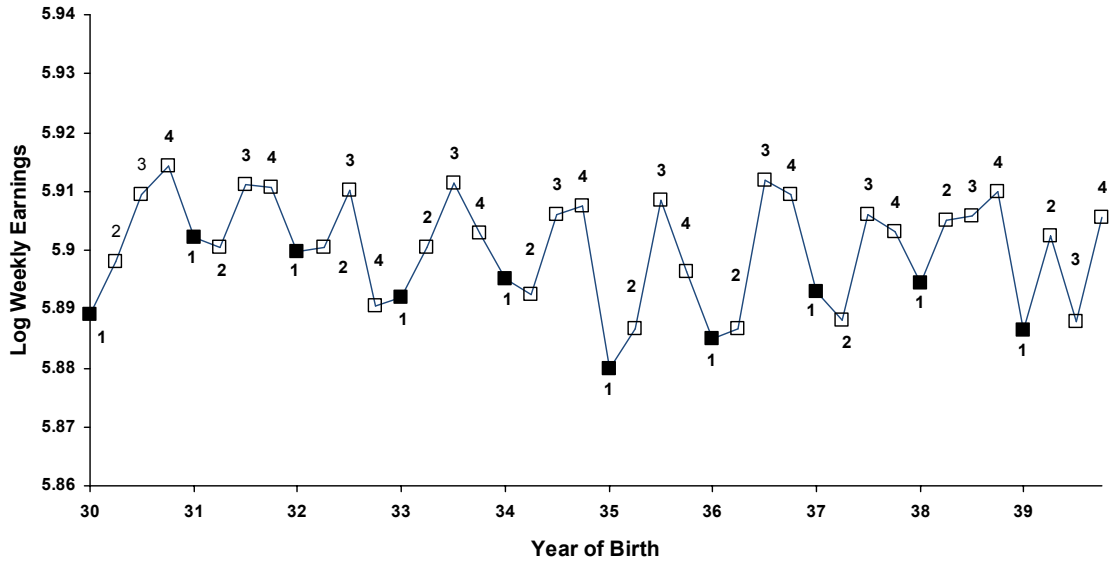
$$E[\mathrm{Y}_i|\mathrm{Z}_i] = \alpha + \rho E[\mathrm{D}_i|\mathrm{Z}_i] \tag{10}$$

Solving this for ρ produces (9)

A. Average Education by Quarter of Birth (first stage)



B. Average Weekly Wage by Quarter of Birth (reduced form)



© Princeton University Press. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

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	.071	.067	.102	.13	.104	.108	.087	.057
	(.0004)	(.0004)	(.024)	(.020)	(.026)	(.020)	(.016)	(.029)
Exogenous Covariates								
Age (in quarters)								✓
Age (in quarters) squared								✓
9 year-of-birth dummies		✓			✓	✓	✓	✓
50 state-of-birth dummies		✓			✓	✓	✓	✓
Instruments								
dummy for QOB = 1			✓	✓	✓	✓	✓	✓
dummy for QOB = 2				✓		✓	✓	✓
dummy for QOB = 3				✓		✓	✓	✓
QOB dummies interacted with year-of-birth dummies (30 instruments total)							✓	✓

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the Angrist and Krueger (1991) 1980 census sample. This sample includes native-born men, born 1930–39, with positive earnings and nonallocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses. QOB denotes quarter of birth.

© Princeton University Press. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

TABLE 6. IV REGRESSIONS ON RETURNS TO EDUCATION: RESULTS FROM THE CENSUS

	Wages, Logged: QOB Instruments		**Wages, Logged: Year*QOB Instruments**		**Wages, in Levels: QOB Instruments**		**Wages, in Levels: Year*QOB Instruments**	
Years of Education	0.103	0.147	0.075	0.09	33.16	49.16	24.13	30.94
	[0.083]	[0.081]	[0.040]	[0.040]	[24.21]	[23.5]	[11.55]	[10.59]
Family Controls?	No	Yes	No	Yes	No	Yes	No	Yes
Instruments	QOB	QOB	YOB*QOB	YOB*QOB	QOB	QOB	YOB*QOB	YOB*QOB
Age Controls?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Dummies?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummies?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weights?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Robust standard errors in brackets. Observations are county-of-birth/quarter-of-birth/year-of-birth cells and all regressions weight by total individuals reporting positive earnings in a cell. The dependent variable in the first two pairs of regressions is the log of average wages in a cell, in the last two pairs of regressions it is the average of cell wages in levels. Regressions are from cohorts of males born between 1944 and 1960; see Table 5 for a description of family characteristic and wage and age variables.

Courtesy of Kasey Buckles and Daniel M. Hungerman. Used with permission.