

Instrumental Variables Estimation

A. Colin Cameron
Univ. of California, Davis

These slides are part of the set of slides
A. Colin Cameron, Introduction to Causal Methods
<https://cameron.econ.ucdavis.edu/causal/>

March 2023

Introduction

- These slides give an introductory example of instrumental variables (IV) and two-stage least squares (2SLS)
 - ▶ IV is a method for causal inference
 - ▶ it is a general method, but requires existence of a valid instrument
- It relies on the strong exclusion restriction (a nontestable assumption) that the instrument(s) do not belong in the model for the outcome (y) of interest.

- Separately the Stata file `iv.do` implements these methods
 - ▶ using dataset `AED_RETURNSTOSCHOOLING.DTA`
- The data are from chapter 17.4 of A. Colin Cameron (2022) *Analysis of Economics Data: An Introduction to Econometrics* <https://cameron.econ.ucdavis.edu/>.
 - ▶ also analyzed in A. Colin Cameron and Pravin K. Trivedi (2005), *Microeconometrics: Methods and Applications*, chapter 4.9.6.
- The original data source is papers by Jeffrey R. Kling (2001), “Interpreting Instrumental Variables Estimates of the Returns to Schooling,” *Journal of Business and Economic Statistics*, 19, pages 358-364 and by David E. Card (1995), “Using Geographic Variation in College Proximity to Estimate the Return to Schooling,” in *Aspects of Labor Market Behavior: Essays in Honor of John Vanderkamp*, L.N. Christofides, E.K. Grant and R. Swidinsky (Eds.).

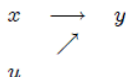
Outline

- 1 Introduction
- 2 Instrumental Variables Estimation
- 3 Example: Returns to Schooling
- 4 Results
- 5 Further Details
- 6 Local average treatment effects - LATE (advanced topic)
- 7 Weak Instruments (advanced topic)
- 8 References

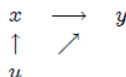
Instrumental Variables Estimation

- Problem: in model $y = \beta_1 + \beta_2 x + u$ we have $E[u|x] \neq 0$
 - ▶ so the error is correlated with the regressor x
 - ▶ then x is called an endogenous variable and OLS is inconsistent.
- Solution: assume there exists an instrument z that
 - ▶ z does not belong in the model for y (crucial exclusion restriction)
 - ▶ z is correlated with x .

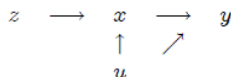
1. OLS consistent



2. OLS inconsistent



3. IV consistent



- Example: in log-wage (y) model treat schooling (x) as endogenous
 - ▶ use distance to closest college as an instrument (z).

Instrumental Variables (continued)

- The instrumental variables (IV) estimator of β_2 is

$$b_{2,IV} = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})}$$

- Note: IV is only possible if one can find a valid instrument.
- Intuitively IV estimates $\frac{\Delta y}{\Delta x} = \frac{\Delta y}{\Delta z} \times \frac{\Delta z}{\Delta x}$ as the ratio $\frac{\Delta y}{\Delta z} / \frac{\Delta x}{\Delta z}$.
 - if a one-unit change in z is associated with
 - a 2 unit increase in x and
 - a 3 unit increase in y
 - then $b_{IV} = 3/2 = 1.5$.
 - and this can be given a causal interpretation of $\frac{\Delta y}{\Delta x} = 1.5$.

IV Exa

- Can extend to multiple regression
 - ▶ exogenous regressors (uncorrelated with u) are instruments for themselves
 - ▶ if more instruments (z) than endogenous regressors (x) then use two-stage least squares (2SLS).
- Suppose $y = \beta_1 + \beta_2 x + \text{other variables} + u$
 - ▶ x is correlated with u while the other variables are uncorrelated with u
 - ▶ z is one or more instruments that are correlated with x but do not directly determine y .
- Then the IV estimator can be computed in two stages
 - ▶ 1. OLS regress x on z and the other variables
 - ▶ Get the prediction \hat{x} from this regressio.
 - ▶ 2. OLS regress y on \hat{x} and the other variables.

IV Example: Returns to schooling

- Does more years of schooling cause higher earnings.
- Model: $y = \beta_1 + \beta_2 x + \text{other controls} + u$.
- Dataset AED_RETURNSTOSCHOOLING.DTA has 1976 data on 3,010 males aged 24 to 34 years old.
- Outcome variable $y = \text{wage76} = \text{log hourly wage}$
- Endogenous regressor $x = \text{grade76} = \text{highest grade completed}$.
- Instrument $z = \text{co14} = \text{indicator for four-year college in county of residence}$.
- Exogenous regressors - here just age for simplicity.
- Nontestable exclusion restriction - having a four-year college in county of residence (z) does not directly affect wage (y)
 - ▶ after controlling for other variables in the model.
- Relevance - need co14 (z) to be correlated with grade76 (x)
 - ▶ after controlling for other variables in the model.

Data Summary

- We have

```
. sum wage76 grade76 col4 age76
```

Variable	Obs	Mean	Std. dev.	Min	Max
wage76	3,010	1.656664	.443798	0	3.1797
grade76	3,010	13.26346	2.676913	1	18
col4	3,010	.6820598	.4657535	0	1
age76	3,010	28.1196	3.137004	24	34

OLS and IV estimates

- First OLS of wage76 on grade76 and age76.
- Then IV of wage76 on grade76 and age76 with col4 an instrument for grade76 (and age76 an instrument for itself).

* OLS and IV estimates

```
reg wage76 grade76 age76, vce(robust)
```

```
estimates store OLS
```

```
ivregress 2sls wage76 age76 (grade76 = col4), vce(robust)
```

```
estimates store IV
```

```
estimates table OLS IV, b(%8.4f) se t(%8.2f) stats(N r2)
```

Results

- IV estimate of grade76 is much larger - a 17% return.
- IV standard error of grade76 is much larger
 - ▶ but grade76 is still statistically significant at level 0.05.

Variable	OLS	IV
grade76	0.0525	0.1740
	0.0028	0.0242
	18.87	7.18
age76	0.0407	0.0416
	0.0024	0.0030
	16.98	13.77
_cons	-0.1831	-1.8196
	0.0773	0.3345
	-2.37	-5.44
N	3010	3010
r2	0.1813	.

Legend: b/se/t

Further Details

- In practice we would add more control variables than just age.
- If we had more than one instrument we use two-stage least squares.
- If we had more than one endogenous regressor then we need at least as many instruments as the number of endogenous regressors.
- An advanced method interprets IV as estimating a local average treatment effects (LATE).
- In applications with a weak instrument we need to use nonstandard inference method
 - ▶ this is usually not a problem for time series examples
 - ▶ but is often a problem with individual cross-section examples
 - ▶ see final section.

Local Average Treatment Effects (advanced topic)

- Consider instrumental variables (IV) estimator in model $y_i = \beta_1 + \gamma d_i$ where z_i is instrument for x_i .
- This model restricts constant treatment effect γ for all individuals.
- Instead allow different (heterogeneous) treatment effects γ_i .
- Specialize to a binary treatment D and suppose for simplicity that higher value of Z makes selection into treatment ($D = 1$) more likely.
- Distinguish between four types of people:
 - ▶ Always-takers chose treatment ($D = 1$) regardless of the value of Z
 - ▶ Never-takers never chose treatment ($D = 0$) regardless of the value of Z
 - ▶ Compliers are induced into treatment so $D = 1$ when $Z = 1$ and $D = 0$ when $Z = 0$
 - ▶ Defiers are induced away from treatment so $D = 0$ when $Z = 1$ and $D = 1$ when $Z = 0$.
- Then, under the crucial and nontestable assumption that there are no defiers, also called the monotonicity assumption, the IV estimator estimates the average treatment effect for compliers.

Weak instruments (advanced topic)

- An instrument z for endogenous regressor x is weak if it is weakly correlated with z after controlling for other variables.
- A diagnostic is to do OLS of $x_j = \alpha_1 + \alpha_2 z_j + \text{other controls} + v_j$
 - ▶ this is called the first-stage regression
 - ▶ if the t statistic for test that $\alpha_2 = 0$ is low then the instrument is weak
 - ▶ there is no clear value of how low is low but definitely $|t| < 3$ is a serious problem.
- Here the instrument is unlikely to be weak as $t = 7.80$

```
. regress grade76 col4 age76, vce(robust) noheader
```

grade76	Robust					
	Coefficient	std. err.	t	P> t	[95% conf. interval]	
col4	.832565	.1067308	7.80	0.000	.6232922	1.041838
age76	-.0126164	.0156219	-0.81	0.419	-.0432471	.0180142
_cons	13.05037	.4366304	29.89	0.000	12.19424	13.90649

Weak instruments (continued)

- With a weak instrument the usual asymptotic theory for inference can fail even in large samples
 - ▶ though with infinite amount of data IV is still consistent.
- Instead use an alternative method - the Anderson-Rubin Wald test and confidence interval
 - ▶ this requires a specialized command.
- Here this alternative gives a similar 95% confidence interval for β_{grade76} as the instrument was not weak.

Weak instrument robust tests and confidence sets for linear IV with robust VCE
 $H_0: \beta[\text{wage76}:\text{grade76}] = 0$

Test	Statistic	p-value	95% Confidence Set
AR	chi2(1) = 75.66	Prob > chi2 = 0.0000	[.132709, .230591]
Wald	chi2(1) = 51.53	Prob > chi2 = 0.0000	[.126472, .221475]

Note: Wald test not robust to weak instruments. Confidence sets estimated for 100

References for IV

- Basic instrumental variables is presented in many texts.
- The following present LATE in addition to IV.
- Joshua D. Angrist and Jörn-Steffen Pischke (2015), *Mastering Metrics*, Princeton University Press, chapter 3.
- Cunningham, Scott (2021), *Causal Inference: The MixTape*, Yale UP, chapter 7.
- A. Colin Cameron and Pravin K. Trivedi (2022), *Microeconometrics using Stata: Volumes 1 and 2, Second Edition*, Stata Press, chapter 7 and 25.5.
- Joshua D. Angrist and Jörn-Steffen Pischke (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, chapter 4.
- A. Colin Cameron and Pravin K. Trivedi (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press, chapter 25.7.
- Jeffrey M. Wooldridge, (2010), *Econometric Analysis of Cross Section and Panel Data, Second Edition*, MIT Press, chapters 5 and 21.4.
- Guido W. Imbens and Donald B. Rubin (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press, chapters 23-25.

References on IV (continued)

- These books by non-economists are similar to *Mastering Metrics* in accessibility.
- Stephen L. Morgan and Christopher Winship (2015), *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, Second edition, Cambridge University Press, chapter 9.
- Richard J. Murnane and John B. Willett (2010), *Methods Matter: Improving Causal Inference in Educational and Social Science Research*, Oxford University Press, chapters 10-11.
- Andrew Gelman, Jennifer Hill and Aki Vehtari (2022), *Regression and Other Stories*, Cambridge University Press, chapter 21.1-21.2.
- This econometrics article reviews inference with weak instruments.
- Isaiah I. Andrews, James H. Stock and L. Sun (2019) "Weak instruments in instrumental variables regression: Theory and practice," *Annual Review of Economics*, 11, pages 727–753.