

OUTLINE: Instrumental Variable Estimation

- Estimation:
 - Assumptions: What do we need?
 - 2 Stage Least Squares method
 - Properties of IV estimator
 - Difficulty: finding good instruments
- Testing
 - Endogeneity
 - Exogeneity of instruments
- Examples
 - Returns to education
 - Cross-country regressions

What should you know?

- When IV can/should be used
- How to interpret IV regressions
- “Detect” potential problems with IV regressions

Instrumental Variable estimation (IV)

Estimation method that recognizes the violation of the zero conditional mean assumption

$$y = \beta_0 + \beta_1 x + u$$

Let x be an endogenous explanatory variable

$$\Rightarrow \text{cov}(x, u) \neq 0$$

Is there another variable z with

$$\text{cov}(z, u) = 0 \quad (15.4)$$

(i.e. z is an exogenous variable)

$$\text{cov}(z, x) \text{ highly positive or negative} \quad (15.5)$$

(i.e. x and z are correlated)

$\Rightarrow z$ is an instrumental variable for x

$\Rightarrow z$ can be used to estimate the effect of x on y

NOTE: u is unobserved \Rightarrow cannot test (15.4)

But can test whether x and z are correlated (15.5)

\Rightarrow estimate $x = \pi_0 + \pi_1 z + v$

\Rightarrow reject $H_0: \pi_1 = 0$

2 Stage Least Squares (2SLS)

= other name for Instrumental Variable estimation

~ Intuition behind the IV estimator

- variation in x exists out of part that is correlated with u and a part that is uncorrelated with u
- IV only uses the later part to identify effect of x on y

Two Stage Least Squares

First stage: decomposes x in 2 parts: $x = \pi_0 + \pi_1 z + v$

Second stage: uses only part predicted by z to estimate effect of x on y

Properties of the IV estimator

IV estimator is consistent

Consistency:

weaker criterium than unbiasedness

If an estimator is consistent, the distribution of $\hat{\beta}_j$ becomes more tightly distributed around β_j as the sample size grows

⇒ probability limit of $\hat{\beta}_j$ is β_j

⇒ $\text{plim } \hat{\beta}_j = \beta_j$

⇒ In cases when there exist no unbiased estimator: estimators that are consistent (but not unbiased) are fine if we have large sample sizes.

Note: in small samples can have a large bias (but OLS is biased and inconsistent)

Larger variance than OLS

Variance of the coefficient estimate increases as correlation between x and z decreases

⇒ in bivariate case:

standard error of $\hat{\beta}_j$ is $\frac{\hat{\sigma}^2}{SST_x R_{x,z}^2}$

(if $R_{x,z}^2 = 1 \Rightarrow$ OLS estimator)

R² of IV not meaningful

When x is endogenous, we can't know how much of variation is explained by x

=> don't use R² for F-stats

Finding good instruments

1) Main difficulty: finding z

⇒ z is correlated with x but does not determine y
(since that would imply that (15.4) does not hold)

e.g. in wage equation: instruments for education

- Mother's (or father's) education?
- Number of siblings
- Proximity to school
- Month of birth

e.g. natural experiments
e.g. Vietnam draft lottery
e.g. Twins
e.g. Weather shocks

2) Other difficulty: often low correlation between x and z

⇒ high variance of coefficient estimate
⇒ weak instruments

1 + 2) If z and u are correlated => IV is inconsistent

Small correlation between z and u and between z and x => large inconsistency

⇒ Always need to check t-stat from regressing x on z
i.e. first stage regression.

Example: College proximity as an IV for education

Source: Card.dta (Wooldridge). Data used in D. Card (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*. Ed. L.N. Christophides, E.K. Grant, and R. Swidinsky, 201-222. Toronto: University of Toronto Press.

Description variables:

variable name	type	format	label	variable label
lwage	float	%9.0g		log(wage)
educ	byte	%9.0g		years of schooling, 1976
exper	byte	%9.0g		age - educ - 6
exper ²	int	%9.0g		exper ²
black	byte	%9.0g		=1 if black
smsa	byte	%9.0g		=1 in in SMSA, 1976
south	byte	%9.0g		=1 if in south, 1976
smsa66	byte	%9.0g		=1 if in SMSA, 1966
reg662	byte	%9.0g		=1 for region 2, 1966
..				
reg669	byte	%9.0g		=1 for region 9, 1966
nearc4	byte	%9.0g		=1 if near 4 yr college, 1966

Summary statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	3010	577.2824	262.9583	100	2404
educ	3010	13.26346	2.676913	1	18
exper	3010	8.856146	4.141672	0	23
expersq	3010	95.57907	84.61831	0	529
black	3010	.2335548	.4231624	0	1
smsa	3010	.7129568	.4524571	0	1
south	3010	.4036545	.4907113	0	1
smsa66	3010	.6495017	.4772053	0	1
reg662	3010	.1607973	.367405	0	1
reg663	3010	.1956811	.39679	0	1
reg664	3010	.0641196	.2450066	0	1
reg665	3010	.2083056	.406164	0	1
reg666	3010	.0960133	.2946584	0	1
reg667	3010	.1099668	.3129003	0	1
reg668	3010	.0282392	.165683	0	1
reg669	3010	.0903654	.2867522	0	1
nearc4	3010	.6820598	.4657535	0	1

. regress lwage educ

Source	SS	df	MS	Number of obs = 3010		
Model	58.5153704	1	58.5153704	F(1, 3008) =	329.54	
Residual	534.126274	3008	.177568575	Prob > F =	0.0000	
Total	592.641645	3009	.196956346	R-squared =	0.0987	
				Adj R-squared =	0.0984	
				Root MSE =	.42139	

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0520942	.0028697	18.15	0.000	.0464674	.057721
_cons	5.570882	.0388295	143.47	0.000	5.494747	5.647017

The coefficient of education in the OLS regression suggest that each additional year of education leads to 5.2 % increase in wage, ceteris paribus. However it is possible to think about unobservables that affect both education and wage, e.g. family background. Because of these doubts, we look for an instrument for education. Card suggested using a dummy variable that indicates whether the person lived close to a 4 year college. This dummy might capture both costs of going to college, and family background.

ivreg lwage (educ=nearc4), first

First-stage regressions

Source	SS	df	MS	Number of obs =	3010
Model	448.604204	1	448.604204	F(1, 3008) =	63.91
Residual	21113.4759	3008	7.01910767	Prob > F =	0.0000
				R-squared =	0.0208
				Adj R-squared =	0.0205
Total	21562.0801	3009	7.16586243	Root MSE =	2.6494

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
nearc4	.829019	.1036988	7.99	0.000	.6256912 1.032347
_cons	12.69801	.0856416	148.27	0.000	12.53009 12.86594

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs =	3010
Model	-340.11155	1	-340.11155	F(1, 3008) =	51.17
Residual	932.753194	3008	.310090823	Prob > F =	0.0000
				R-squared =	.
				Adj R-squared =	.
Total	592.641645	3009	.196956346	Root MSE =	.55686

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
educ	.1880626	.0262913	7.15	0.000	.1365118 .2396135
_cons	3.767472	.3488617	10.80	0.000	3.08344 4.451503

Instrumented: educ
 Instruments: nearc4

Interpretation: In the first stage, we are mainly interested in the significance of the nearc4 dummy. The first stage shows a strongly significant coefficient that is positive (as expected), confirming our assumption that cov(z,x) positive. The R-squared is however low, indicating that we will only use a small share of the variation in education, in the second stage. This causes our standard errors in the second stage to be higher. (compare the OLS s.e. with the IV s.e. to confirm this). Nevertheless, the second stage still shows a positive and significant (at the 1%) coefficient for education. An additional year of education is predicted to increase wage by 19%. This is substantially more than the OLS estimate.

IV in multiple regression

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u$$

y_2 : endogenous explanatory variable

hypothesize that y_2 determines y_1 but y_2 is endogenous

⇒ find instrument z_2 (different than exogenous variable z_1)

- $\text{cov}(z_2, u) = 0$
- high correlation z_2 and y_2
 - to test: $y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v$

$$H_0: \pi_2 = 0$$

Notes

- Sometimes all z 's are called instrumental variables
- When more than 1 endogenous explanatory variable
 - ⇒ # instruments \geq # endogenous variables
- Can use more than one instrument for an endogenous variable

Example: continued

```
. reg lwage educ exper expersq black smsa south smsa66 reg662 reg663 reg664
reg665 reg666 reg667 reg668 reg669
```

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
				R-squared =	0.2998
				Adj R-squared =	0.2963
Total	592.641645	3009	.196956346	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
smsa	.1363845	.0201005	6.79	0.000	.0969724 .1757967
south	-.147955	.0259799	-5.69	0.000	-.1988952 -.0970148
smsa66	.0262417	.0194477	1.35	0.177	-.0118905 .0643739
reg662	.0963672	.0358979	2.68	0.007	.0259801 .1667542
reg663	.14454	.0351244	4.12	0.000	.0756696 .2134105
reg664	.0550756	.0416573	1.32	0.186	-.0266043 .1367554
reg665	.1280248	.0418395	3.06	0.002	.0459878 .2100618
reg666	.1405174	.0452469	3.11	0.002	.0517992 .2292356
reg667	.117981	.0448025	2.63	0.008	.0301343 .2058277
reg668	-.0564361	.0512579	-1.10	0.271	-.1569404 .0440682
reg669	.1185698	.0388301	3.05	0.002	.0424335 .194706
_cons	4.620807	.0742327	62.25	0.000	4.475254 4.766359

Interpretation: A whole number of control variables were added in the OLS to reduced the omitted variable problem. The coefficient of education in this OLS regression suggest that each additional year of education leads to 7.4 % increase in wage, ceteris paribus. The OLS controls for experience, ethnicity, living conditions and regional characteristics (south dummy and the regional dummies, most of which are shown to be significant). However it is still possible to think about unobservables that affect both education and wage, e.g. family background. Because of these doubts, we look for an instrument for education. We use again the dummy variable that indicates whether the person lived close to a 4 year college. Using the IV with all the control variables, also makes our assumption of $cov(z,u)=0$ more credible.

```
. ivreg lwage exper expersq black smsa south smsa66 reg662 reg663 reg664 reg665
reg666 reg667 reg668 reg669 (educ =nearc4) , first
```

First-stage regressions

Source	SS	df	MS	Number of obs =	3010
Model	10287.6179	15	685.841194	F(15, 2994) =	182.13
Residual	11274.4622	2994	3.76568542	Prob > F =	0.0000
-----				R-squared =	0.4771
-----				Adj R-squared =	0.4745
Total	21562.0801	3009	7.16586243	Root MSE =	1.9405

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
exper	-.4125334	.0336996	-12.24	0.000	-.4786101 -.3464566
expersq	.0008686	.0016504	0.53	0.599	-.0023674 .0041046
black	-.9355287	.0937348	-9.98	0.000	-1.11932 -.7517377
smsa	.4021825	.1048112	3.84	0.000	.1966732 .6076918
south	-.0516126	.1354284	-0.38	0.703	-.3171548 .2139296
smsa66	.0254805	.1057692	0.24	0.810	-.1819071 .2328682
reg662	-.0786363	.1871154	-0.42	0.674	-.4455241 .2882514
reg663	-.027939	.1833745	-0.15	0.879	-.3874918 .3316138
reg664	.117182	.2172531	0.54	0.590	-.3087984 .5431624
reg665	-.2726165	.2184204	-1.25	0.212	-.7008857 .1556528
reg666	-.3028147	.2370712	-1.28	0.202	-.7676536 .1620242
reg667	-.2168177	.2343879	-0.93	0.355	-.6763953 .2427598
reg668	.5238914	.2674749	1.96	0.050	-.0005617 1.048344
reg669	.210271	.2024568	1.04	0.299	-.1866975 .6072395
nearc4	.3198989	.0878638	3.64	0.000	.1476194 .4921785
_cons	16.63825	.2406297	69.14	0.000	16.16644 17.11007

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs =	3010
Model	141.146813	15	9.40978752	F(15, 2994) =	51.01
Residual	451.494832	2994	.150799877	Prob > F =	0.0000
-----				R-squared =	0.2382
-----				Adj R-squared =	0.2343
Total	592.641645	3009	.196956346	Root MSE =	.38833

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.1315038	.0549637	2.39	0.017	.0237335	.2392742
exper	.1082711	.0236586	4.58	0.000	.0618824	.1546598
expersq	-.0023349	.0003335	-7.00	0.000	-.0029888	-.001681
black	-.1467757	.0538999	-2.72	0.007	-.2524603	-.0410912
smsa	.1118083	.031662	3.53	0.000	.0497269	.1738898
south	-.1446715	.0272846	-5.30	0.000	-.19817	-.091173
smsa66	.0185311	.0216086	0.86	0.391	-.0238381	.0609003
reg662	.1007678	.0376857	2.67	0.008	.0268753	.1746603
reg663	.1482588	.0368141	4.03	0.000	.0760752	.2204423
reg664	.0498971	.0437398	1.14	0.254	-.0358661	.1356602
reg665	.1462719	.0470639	3.11	0.002	.053991	.2385529
reg666	.1629029	.0519096	3.14	0.002	.0611209	.264685
reg667	.1345722	.0494023	2.72	0.006	.0377063	.2314381
reg668	-.083077	.0593314	-1.40	0.162	-.1994113	.0332573
reg669	.1078142	.0418137	2.58	0.010	.0258278	.1898007
_cons	3.666151	.9248295	3.96	0.000	1.852785	5.479517

Instrumented: educ
Instruments: exper expersq black smsa south smsa66 reg662 reg663 reg664
reg665 reg666 reg667 reg668 reg669 nearc4

Interpretation: In the first stage, we are mainly interested in the significance of the nearc4 dummy. The first stage shows a strongly significant coefficient that is positive (as expected), so our assumption that cov(z,x) is positive is confirmed.

The second stage shows a positive and significant (at the 5%) coefficient for education. An additional year of education is predicted to increase wage by 13%, keeping everything else constant. This is substantially more than the OLS estimate.

Testing for endogeneity

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u \quad (15.49)$$

We want to test whether y_2 is exogenous \Rightarrow i.e. $E(u|y_2) = 0$

There are 2 exogenous variables correlated with y_2 and not with u

\Rightarrow First stage: $y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v$

\Rightarrow obtain an estimate of v

\Rightarrow test whether u and v are uncorrelated, by adding estimation of v in (15.49)

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v} + error$$

Test $H_0: \delta_l = 0$

\Rightarrow rejecting $H_0 \Rightarrow y_2$ is endogenous

If more than 1 potentially endogenous

- ⇒ obtain estimate of v for each of them
- ⇒ F-test for joint significance

Notes: Failing to reject might just be a result of weak instrument!
Good to report both OLS and IV estimates

Testing for exogeneity of instruments

Overidentification test:

Is one instrument exogenous assuming that the other one is exogenous?

Again can reject exogeneity, but failing to reject does not mean that instruments are truly exogenous.

Examples of IV in cross country regressions

1) The Colonial Origins of Comparative Development: An Empirical Investigation

Acemoglu, Johnson and Robinson, *American Economic Review*, 2001

Interpretation: The authors use a sample of 64 countries to analyze the effects of property rights security, measured by the average protection against expropriation risk from 85-95, on income per capita on 1995. Property rights security is used as a measurement of the quality of institutions. The OLS coefficient of the expropriation risk variable is shown in Panel C and is significant in all specifications. Because the expropriation risk is potentially endogenous (many unobservables may affect both expropriation risk and income per capita), the authors instrument the expropriation risk, with the log of European settlers mortality. The rationale is that in countries where climatological circumstances caused high mortality among early European settlers, the Europeans decided to only install expropriative regimes and not to install good institutions. These countries still don't have good institutions today. The first stage regression (Panel B) indeed shows a negative correlation between historical mortality and expropriation risk in the late 20th century. The coefficient is significant in all specifications. Panel A shows the second stage relationship between expropriation risk and GDP per capita. A strongly significant positive coefficient is found that is larger than the OLS estimate (this suggests that the OLS estimate is negatively biased). The different columns show robustness tests. The results are robust when different sets of control variables are added. The coefficient of interest also remains positive when different subsets of the data are excluded (although the magnitude of the coefficient changes). This shows that the significant results are driven uniquely by these subsets of countries. (Note: neo-europes stand for US, Canada, Australia and New Zealand).

Acemoglu, Johnson and Robinson, Am. Econ. Review, 2001

TABLE 4—IV REGRESSIONS OF LOG GDP PER CAPITA

	Base sample (1)	Base sample (2)	Base sample without Neo-Europes (3)	Base sample without Neo-Europes (4)	Base sample without Africa (5)	Base sample without Africa (6)	Base sample with continent dummies (7)	Base sample with continent dummies (8)	Base sample, dependent variable is log output per worker (9)
Panel A: Two-Stage Least Squares									
Average protection against expropriation risk 1985–1995	0.94 (0.16)	1.00 (0.22)	1.28 (0.36)	1.21 (0.35)	0.58 (0.10)	0.58 (0.12)	0.98 (0.30)	1.10 (0.46)	0.98 (0.17)
Latitude		-0.65 (1.34)		0.94 (1.46)		0.04 (0.84)		-1.20 (1.8)	
Asia dummy							-0.92 (0.40)	-1.10 (0.52)	
Africa dummy							-0.46 (0.36)	-0.44 (0.42)	
“Other” continent dummy							-0.94 (0.85)	-0.99 (1.0)	
Panel B: First Stage for Average Protection Against Expropriation Risk in 1985–1995									
Log European settler mortality	-0.61 (0.13)	-0.51 (0.14)	-0.39 (0.13)	-0.39 (0.14)	-1.20 (0.22)	-1.10 (0.24)	-0.43 (0.17)	-0.34 (0.18)	-0.63 (0.13)
Latitude		2.00 (1.34)		-0.11 (1.50)		0.99 (1.43)		2.00 (1.40)	
Asia dummy							0.33 (0.49)	0.47 (0.50)	
Africa dummy							-0.27 (0.41)	-0.26 (0.41)	
“Other” continent dummy							1.24 (0.84)	1.1 (0.84)	
R ²	0.27	0.30	0.13	0.13	0.47	0.47	0.30	0.33	0.28
Panel C: Ordinary Least Squares									
Average protection against expropriation risk 1985–1995	0.52 (0.06)	0.47 (0.06)	0.49 (0.08)	0.47 (0.07)	0.48 (0.07)	0.47 (0.07)	0.42 (0.06)	0.40 (0.06)	0.46 (0.06)
Number of observations	64	64	60	60	37	37	64	64	61

Notes: The dependent variable in columns (1)–(8) is log GDP per capita in 1995, PPP basis. The dependent variable in column (9) is log output per worker, from Hall and Jones (1999). “Average protection against expropriation risk 1985–1995” is measured on a scale from 0 to 10, where a higher score means more protection against risk of expropriation of investment by the government, from Political Risk Services. Panel A reports the two-stage least-squares estimates, instrumenting for protection against expropriation risk using log settler mortality; Panel B reports the corresponding first stage. Panel C reports the coefficient from an OLS regression of the dependent variable against average protection against expropriation risk. Standard errors are in parentheses. In regressions with continent dummies, the dummy for America is omitted. See Appendix Table A1 for more detailed variable descriptions and sources.

2) Economic Shocks and Civil Conflict: An Instrumental Variable approach

Miguel, Shanker Satyanath and Ernest Sergenti, 2004, *Journal of Political Economy*, 2004, 112(4), 725-753

Table 4: Economic Growth and Civil Conflict

Explanatory variable	Dependent variable: <u>Civil conflict \geq 25 deaths</u>						<u>Civil conflict \geq 1000 deaths</u>
	Probit (1)	OLS (2)	OLS (3)	OLS (4)	IV-2SLS (5)	IV-2SLS (6)	IV-2SLS (7)
Economic growth rate, t	-0.37 (0.26)	-0.33 (0.26)	-0.15 (0.19)	-0.18 (0.16)	-0.38 (1.38)	-1.13 (1.40)	-1.48* (0.82)
Economic growth rate, t-1	-0.14 (0.23)	-0.08 (0.24)	0.07 (0.19)	0.09 (0.16)	-2.14** (1.03)	-2.53** (1.10)	-0.76 (0.70)
Log(GDP per capita), 1979	-0.067 (0.061)	-0.041 (0.050)	0.093 (0.072)		0.063 (0.080)		
Democracy (Polity IV), t-1	0.001 (0.005)	0.001 (0.005)	0.004 (0.006)		0.005 (0.006)		
Ethno-linguistic fractionalization	0.24 (0.26)	0.23 (0.27)	0.53 (0.41)		0.53 (0.41)		
Religious fractionalization	-0.29 (0.26)	-0.24 (0.24)	-0.08 (0.41)		-0.01 (0.43)		
Oil exporting country	0.02 (0.21)	0.05 (0.21)	-0.19 (0.21)		-0.15 (0.23)		
Log(mountainous)	0.077** (0.041)	0.076* (0.039)	0.06 (0.06)		0.06 (0.06)		
Log (national population), t-1	0.080 (0.051)	0.068 (0.051)	0.232*** (0.081)		0.226** (0.087)		
Country fixed effects	No	No	No	Yes	No	Yes	Yes
Country-specific time trends	No	No	Yes	Yes	Yes	Yes	Yes
R ²	-	0.13	0.52	0.70	-	-	-
Root MSE	-	0.42	0.32	0.26	0.36	0.32	0.24
Number of observations	743	743	743	743	743	743	743

Table 4 Notes: Huber robust standard errors in parentheses. Significantly different than zero at 90% (*), 95% (**), 99% (***) confidence. Regression disturbance terms are clustered at the country level. Regression (1) presents marginal probit effects, evaluated at explanatory variable mean values. The instrumental variables for economic growth in (5)-(7) are (Growth in rainfall, t) and (Growth in rainfall, t-1). A country-specific year time trend is included in all specifications (coefficient estimates not reported), except for (1) and (2), where a linear time trend is included.

Interpretation: this dataset uses data from 41 Sub-Saharan African countries in 19 years (81-99). Hence it is a panel dataset, but we can read the regressions the same way as we would read a cross-sectional regression. The dependent variable is binary, and indicates whether there was a civil conflict. The independent variables of interest are the economic growth rate in the year of the observation and the year before. The authors don't show the first stage, but indicate in the footnote that the instruments used are the growth in rainfall in the year of the observation and the year before. They find significant negative effects of past growth rates on civil conflict in the IV regressions. This contrasts with the insignificant results in the OLS regression.