Applied Microeconometrics II , Lecture 6

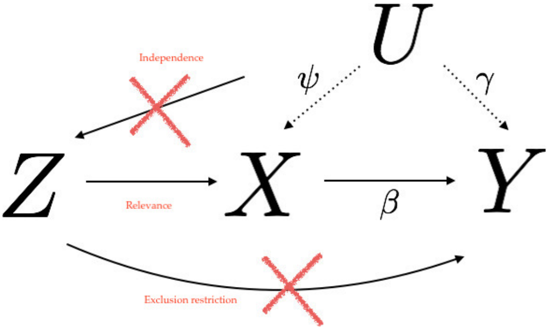
Ciprian Domnisoru
Aalto University

Outline

- Review
- IV as a method to address OVB
- Weak instruments
- IV as a method to addess simultaneity bias
- IV as a method to deal with attrition in randomized experiments
- IV as a method to address measurement error

IV review

- $Y = \alpha + \beta X + u$



-
  - $X = \alpha^{'} + \delta Z + v$
  - $Y = \alpha + \beta_{2SLS}\hat{X} + \epsilon$
- Use only $\hat{X}$, the part of the variation in X that is explained by its correlation with Z, and uncorrelated with U.

IV Regression: 2SLS

- Using regression to form an IV estimate—one binary instrument (Z_i)
- First Method
  1. Estimate the *reduced form* with this regression

     $$Y_i = \alpha_0 + \rho Z_i + e_{0i}$$

     the coefficient in this regression has the interpretation

     $$\rho = E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$$

  2. Estimate the *first stage* with this regression

     $$X_i = \alpha_1 + \phi Z_i + e_{1i}$$

     the coefficient in this regression has the interpretation

     $$\phi = E[X_i|Z_i = 1] - E[X_i|Z_i = 0]$$

  3. Form ratio $\lambda = \frac{\rho}{\phi} = \frac{C(Y_i, Z_i)/V(Z_i)}{C(X_i, Z_i)/V(Z_i)} = \frac{C(Y_i, Z_i)}{C(X_i, Z_i)}$

IV Regression: 2SLS

- Using regression to form an IV estimate—one binary instrument (Z_i)
- Second Method (2SLS)
  1. Estimate the *first stage* with this regression
     $$X_i = \alpha_1 + \phi Z_i + e_{1i}$$
     and form *fitted* values $\hat{X}_i$
  2. Estimate the regression
     $$Y_i = \alpha_2 + \lambda \hat{X}_i + e_{2i}$$
     This results in the coefficient
     $$\begin{aligned}\lambda_{2SLS} &= \frac{C(Y_i, \hat{X}_i)}{V(\hat{X}_i)} = \frac{C(Y_i, \alpha_1 + \phi Z_i)}{V(\alpha_1 + \phi Z_i)} \\ &= \frac{\phi C(Y_i, Z_i)}{\phi^2 V(Z_i)} = \frac{C(Y_i, Z_i)/V(Z_i)}{\phi} = \frac{\rho}{\phi} = \lambda\end{aligned}$$
- Second method is same as the first!

OLS vs. 2SLS bias and weak instruments

- $\hat{\beta}_1^{OLS} = \frac{Cov(X,Y)}{Var(X)} = \beta_1 + \frac{Cov(X,u)}{Var(X)}$
- Notice the bias depends on the exogeneity of X
- $\hat{\beta}_1^{2SLS} = \frac{Cov(Z,Y)}{Cov(Z,X)} = \beta_1 + \frac{Cov(Z,u)}{Cov(Z,X)}$
- The 2SLS bias depends on two conditions: exogeneity and relevance.
- In the presence of weak (low relevance) instruments, the bias in 2SLS can be much larger than the OLS bias.
- To make matters worse, the standard normal asymptotic approximation for the sampling distribution of the 2SLS estimator relies on the correlation between instruments and the endogeneous regressor. If correlation is close to zero, approximation will not be accurate.
- Corrections for this start at F$>$10 in the homoskedastic case, but then critical F can increase when we adjust for heteroskedasticity, clustering, and relax other assumptions...F$> 16 - 25$...$>$100 (Lee et al., 2020. Valid t-ratio Inference for IV).

AR confidence intervals for weak instruments: weakiv

Table 4: Effects of the Berthoin Reform on Educational Attainment and Earnings, Global Polynomial Approach, Comparison with Grenet (2013)

	(1)	(2)	(3)	(4)	(5)
A. Full sample					
First stage	.328***	.248***	.270***	.270***	.222***
	(.050)	(.064)	(.057)	(.057)	(.055)
2SLS estimate	0.054***	0.037*	0.027	0.018	.004
	(.017)	(.019)	(0.027)	(0.023)	(.018)
AR c.i.	[.018,.088]	[-.009, .070]	[-.059, .084]	[-.028,.068]	[-.036,.044]
F-stat	42.94	33.30	14.81	22.36	16.66
Wild bootstrap p-value	0.023	0.123	0.475	0.555	0.843
Obs.	42,214	45,874	54,590	54,590	54,590
B. Parents in lower education occupations					
First stage	.390***	.317***	.308***	.308***	.258***
	(.056)	(.055)	(.081)	(.081)	(.071)
2SLS estimate	0.093***	0.091***	0.065**	0.052**	.048**
	(.024)	(.023)	(.025)	(0.021)	(.019)
AR c.i.	[.047,.141]	[.034, .144]	[.023,.140]	[.016,.116]	[.011, .103]
F-stat	48.35	44.60	14.32	14.32	13.24
Wild bootstrap p-value	0.001	0.000	0.003	0.086	0.118
Obs.	19,949	21,530	26,155	26,155	26,155
Age range	29-49	28-49	28-58	28-58	28-58
Cohorts	1946-1960	1944-1962	1944-1962	1944-1962	1944-1962
Earnings	Monthly	Monthly	Monthly	Hourly	Hourly
Polynomial	Quadratic	Quadratic	Quadratic	Quadratic	Quartic

AR confidence intervals for weak instruments: weakiv command

- The AR statistic is the F-stat testing the hypothesis that the coefficients on Z are 0 in a regression of $Y - X\beta_0$ on Z and other covs. Valid test if instruments are weak.
- Limitation: rejection can arise bc. β_0 is false OR Z is endogenous; less powerful.

Finlay, K., Magnusson, L.M., Schaffer, M.E. 2013. weakiv: Weak-instrument-robust tests and confidence intervals for instrumental-variable (IV) estimation of linear, probit and tobit models.

Weak instrument robust tests and confidence sets for linear IV
H0: beta[learn:agelfted] = 0

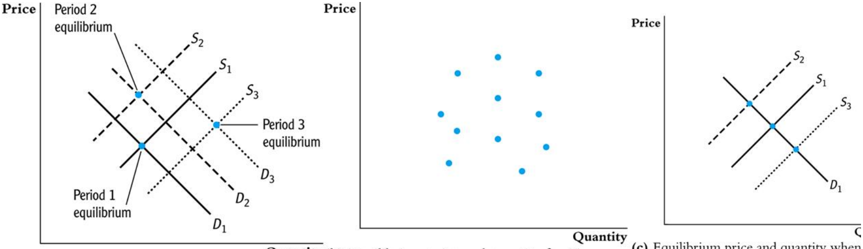
Test	Statistic	p-value	Conf. level	Confidence Set
AR	chi2(1) = 7.32	Prob > chi2 = 0.0068	95%	[.017896, .087861]
Wald	chi2(1) = 9.81	Prob > chi2 = 0.0017	95%	[.020297, .088204]

Confidence sets estimated for 100 points in [-.013657, .122158].
Number of observations N = 13496164.
Notes: Method = **minimum distance/Wald.** Tests robust to heteroskedasticity and clustering on clust.

IV as a method to address simultaneous causality

Simultaneous causality

Simultaneous causality bias in the OLS regression of ln(Supply) on ln(Price) arises because both price and quantity are determined by the interaction of demand and supply.
IV estimates the demand curve by isolating shifts in price and quantity that arise from shifts in supply. Z is a variable that shifts supply but not demand (shifts supply exogeneously).



(a) Demand and supply in three time periods **(b)** Equilibrium price and quantity for 11 time periods **(c)** Equilibrium price and quantity when the supply curve shifts

Simultaneous causality

$ln(Supply) = \beta_0 + \beta_1 ln(Price) + u$
price and quantity are jointly determined by the interactions of supply and demand

Need to find a variable that shifts supply but not demand
Z=rainfall
1) regress ln(price) on rainfall : This isolates changes in log price that arise from supply (part of supply, at least).
2) regress ln(supply) on $\widehat{ln(price)}$: The regression counterpart of using shifts in the supply curve to trace out the demand curve.

Example

Angrist, Graddy, Imbens (2000). The interpretation of instrumental variables estimators in simultaneous equation models with an application to the demand for fish

TABLE 2

Reduced form estimates for log quantity (111 *Obs.*)

Variable	coef	(s.e.)	coef	(s.e.)	coef	(s.e.)	coef	(s.e.)
Stormy	−0·36	(0·15)	−0·38	(0·15)	−0·45	(0·08)	−0·43	(0·17)

TABLE 3

Reduced form estimates for log price (111 *Obs*)

Variable	coef	(s.e.)	coef	(s.e.)	coef	(s.e.)	coef	(s.e.)
Stormy	0·34	(0·07)	0·31	(0·08)	0·44	(0·08)	0·42	(0·08)

−0·11 (0·11)
−0·08 (0·11)

TABLE 5

Two-stage-least-squares estimates of demand function with stormy and mixed as instruments

Variable	est.	(s.e.)	est.	(s.e.)
Av. price effect	−1·01	(0·42)	−0·947	(0·46)

Simultaneous causality

$log(violentcrime) = \alpha_1 + \alpha_2 log(policeforce) + \alpha_3 X_3 + u_1$ and
$log(policeforce) = \beta_1 + \beta_2 log(violentcrime) + \beta_3 W_3 + v_1$

$Y_1 = \alpha_1 + \alpha_2 Y_2 + \alpha_3 X_3 + u_1$
$Y_2 = \beta_1 + \beta_2 Y_1 + \beta_3 W_3 + v_1$

$$Y_2 = \beta_1 + \beta_2(\alpha_1 + \alpha_2 Y_2 + \alpha_3 X_3 + u_1) + \beta_3 W_3 + v_1$$

$$Y_2(1 - \beta_2\alpha_2) = \beta_1 + \beta_2\alpha_1 + \beta_2\alpha_3 X_3 + \beta_3 W_3 + \beta_2 u_1 + v_1$$

$Cov(Y_2, u_1) = \frac{\beta_2 Var(u_1)}{1-\beta_2\alpha_2}$. Hence OLS is biased if $\beta_2 \neq 0$.

Simultaneous causality

$$log(violentcrime) = \alpha_1 + \alpha_2 log(policeforce) + \alpha_3 X_3 + u_1$$

- Use instrument for log(policeforce): indicators for mayoral and gubernatorial election in year T (Levitt, AER 1997).

	Gubernatorial election year ($N = 302$)	Mayoral election year ($N = 391$)	No election ($N = 621$)
Δln Sworn police officers per capita	0.021 (0.006)	0.020 (0.007)	0.000 (0.006)

-

Simultaneous causality

Variable	(1) OLS	(2) OLS	(3) 2SLS	(4) 2SLS	(5) 2SLS
ln Sworn officers per capita	0.28	−0.27	−1.39	−0.90	−0.65
	(0.05)	(0.06)	(0.55)	(0.40)	(0.25)
State unemployment rate	−0.65	−0.25	−0.00	−0.19	−0.13
	(0.40)	(0.31)	(0.36)	(0.33)	(0.32)
ln Public welfare spending per capita	−0.03	−0.03	−0.03	−0.03	−0.02
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
ln Education spending per capita	0.04	0.06	0.02	0.03	0.05
	(0.07)	(0.06)	(0.07)	(0.07)	(0.06)
Percent ages 15–24 in SMSA	1.43	−2.61	−1.47	−2.55	−2.02
	(1.00)	(3.71)	(4.12)	(3.88)	(3.76)
Percent black	0.010	−0.017	−0.034	−0.025	−0.022
	(0.003)	(0.011)	(0.015)	(0.013)	(0.012)
Percent female-headed households	0.003	0.007	0.040	0.023	0.018
	(0.006)	(0.023)	(0.030)	(0.027)	(0.025)
Data differenced?	No	Yes	Yes	Yes	Yes
Instruments:	None	None	Elections	Election * city-size interactions	Election * region interactions

IV as method to deal with attrition

IV in randomized trials

- Individuals may be assigned to treatment (training) but only some actually participate
- Motivation leads to bias in the estimated treatment effect.
- use IV: send a letter encouraging one randomly selected part of the treatment group to participate, control gets no letter
- Z = 1 if a letter is sent, X = 1 if the person followed the training program, Y = 1 if she had found a job after 6 months.
- Remember: IV as Ratio of Coefficients: If you have one endogenous variable X and one instrument Z, you can regress X on Z to get β_{XZ} and regress Y on Z to get β_{YZ}, and the IV estimate $\beta_{IV} = \beta_{YZ}/\beta_{XZ}$.

IV in randomized trials

- In the special case where X, Y and Z are binary (Wald estimator), $\beta_{IV} = \frac{P(Y=1|Z=1)-P(Y=1|Z=0)}{P(X=1|Z=1)-P(X=1|Z=0)}$
- Remember, four categories: always takers (independent of letter) , never takers (regardless of the letter), compliers (only if receive letter) , deniers (they would have participated, but the letter made them change their mind).
- Average treatment effect only for compliers: Local Average Treatment Effect.

IV in randomized trials

- $\beta_{IV} = \frac{P(Y=1|Z=1)-P(Y=1|Z=0)}{P(X=1|Z=1)-P(X=1|Z=0)}$
- Average treatment effect only for compliers: Local Average Treatment Effect.
- % always takers: % of the no letter group which followed the training
- % never takers: % of the letter group which did not follow the training
- % compliers: % of the letter group which followed the training (includes compliers + always takers) - % of the no letter group which followed the training (always takers).
- Monotonicity assumption: no deniers.

- Why is percentage of always takers = the same in the test and in the control group ?

IV in randomized trials : using initial assignment as an instrument

- The Tennessee STAR class size experiement randomly assigned students to small, regular and large classrooms
- Attrition may bias estimates
- Use initial assignment to a type of class as an instrumental variable for class size

IV in randomized trials : using initial assignment as an instrument

achievement would take actual class size into account. A natural model for this situation is a triangular model of student achievement in which the actual number of students in the class is included on the right-hand side, and initial assignment to a class type is used as an instrumental variable for actual class size. Specifically, we estimate the following model by 2SLS:

$$CS_{ics} = \pi_0 + \pi_1 S_{ios} + \pi_2 R_{ios} + \pi_3 X_{ics} + \delta_s + \tau_{ics} \tag{3}$$

$$Y_{ics} = \beta_0 + \beta_1 CS_{ics} + \beta_2 X_{ics} + \alpha_s + \varepsilon_{ics}, \tag{4}$$

where CS_{ics} is the actual number of students in the class, S_{ios} is a dummy variable indicating assignment to a small class the first year the student is observed in the experiment, R_{ios} is a dummy variable indicating assignment to a regular class the first year the

-

IV in randomized trials

TABLE VII
OLS and 2SLS Estimates of Effect of Class Size on Achievement
Dependent Variable: Average Percentile Score on SAT

Grade	OLS	2SLS	Sample size
	(1)	(2)	(3)
K	−.62	−.71	5,861
	(.14)	(.14)	
1	−.85	−.88	6,452
	(.13)	(.16)	
2	−.59	−.67	5,950
	(.12)	(.14)	
3	−.61	−.81	6,109
	(.13)	(.15)	

The coefficient on the actual number of students in each class is reported. All models also control for school effects; student's race, gender, and free lunch status; teacher race, experience, and education. Robust standard errors that allow for correlated errors among students in the same class are reported in parentheses.

▶

Using IV to address measurement error

Measurement Error (ME)

- Suppose you've dreamed of running the regression

$$Y_i = \alpha + \beta S_i^* + e_i$$

  - but data on S_i^* are unavailable
  - you only observe a mismeasured version, S_i

- Write relationship between observed and desired regressor as

$$S_i = S_i^* + m_i$$

  - m_i is the *measurement error* in S_i

Using IV to Address Measurement Error

- Without covariates, the IV formula for the coefficient on S_i in a bivariate regression is

$$\beta_{IV} = \frac{Cov(Y_i, Z_i)}{Cov(S_i, Z_i)}$$

  - where Z_i is the instrument
- Provided the instrument is uncorrelated with the measurement error and the residual, e_i, IV eliminates the bias due to mismeasured S_i

Using IV to Address Measurement Error

- To see why IV works in this context, substitute for Y_i and S_i

$$\begin{aligned}\beta_{IV} = \frac{Cov(Y_i, Z_i)}{Cov(S_i, Z_i)} &= \frac{Cov(\alpha + \beta S_i^* + e_i, Z_i)}{Cov(S_i^* + m_i, Z_i)} \\ &= \frac{\beta Cov(S_i^*, Z_i) + Cov(e_i, Z_i)}{Cov(S_i^*, Z_i) + Cov(m_i, Z_i)}\end{aligned}$$

- Again, provided the instrument is uncorrelated with the measurement error and the residual, IV eliminates the bias due to mismeasured S_i. That is,
  - if $C(e_i, Z_i) = C(m_i, Z_i) = 0$, then

$$\beta_{IV} = \beta \frac{C(S_i^*, Z_i)}{C(S_i^*, Z_i)} = \beta$$

Using IV to Address Measurement Error

- For the problem of measurement error in a regressor, a common choice of instrument (Z_i) is the rank of the mismeasured variable
  - although the mismeasured variable contains an element of measurement error, if that error is relatively small, it will not alter the rank of the observation in the distribution
  - be cautious: mismeasurement can be large in many settings
- Other popular instruments are lagged values of the regressor of interest when it is observed over a number of periods of time
  - the past might explain the present values of the regressor
  - and should affect the outcome only through this channel

Random Measurement Error in the Dependent Variable

- Should we be concerned about bias in this case?
  - NO, there is no bias if measurement error is random, only larger standard errors

- To see why, suppose you've dreamed of running the regression

$$Y_i^* = \alpha + \beta S_i + e_i$$

  - but data on Y_i^* are unavailable
  - you only observe a mismeasured version, Y_i

- Write relationship between observed and desired outcome as

$$Y_i = Y_i^* + m_i$$

  - m_i is the *measurement error* in Y_i

Measurement Error in the Dependent Variable

- The regression equation becomes

$$
\begin{aligned}
Y_i^* &= \alpha + \beta S_i + e_i \\
Y_i^* + m_i &= \alpha + \beta S_i + (e_i + m_i) \\
Y_i &= \alpha + \beta S_i + u_i
\end{aligned}
$$

- Notice we can still run the standard OLS on

$$Y_i = \alpha + \beta S_i + u_i$$

  - and there would be no bias in β
  - but $V(u_i) = V(e_i) + V(m_i) > V(e_i)$

- Because the standard errors of the estimated β depend on $V(u_i)$, then they would be larger than in the dream regression