Lecture 21: Endogeneity and Instrumental Variables

Prof. Esther Duflo

14.310x

When regression control will just not do it...

- Back to our effort to establish causality.
- $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$
- We have seen some possible control-variables strategy to get at it: simply controlling for the right set of X_{i2} variable, difference in difference stategies, Regression discontinuity designs....
- But some times none of this will work.
- The necessary control variables may not be available...
- Or a plausible case can be made that the variable X_{i2} can affect Y_i!

Endogeneity

- We talk about endogeneity, when there this mutual relationship. i.e when reasonable case can be made either way...
- Examples:
  - Democracy and Growth
  - Health and Exercise
  - Crosswords and Cognitive decline
  - Prices and quantities

The Benefits of Education

- There is a correlation between education and many outcomes. e.g. knowledge, earnings, fertility, health etc. ..
- What is the possible bias if we interpret the relationship between education and earnings causally?
- Randomly assigning "education" to people is not possible: one's education is closely linked to other aspects of one's person.

Assigning an "instrument"

- You can randomly assign a student to a program which may lead her to get more education. What are examples of interventions like that?
- Then we can exploit the fact that the intervention affects education: if it has no direct impact on earnings (or any other outcomes you want to look at), but you see that it affects the earnings, you can infer that it affected earnings through education, and hence that education affects earning.
- Today, we are going to use this insight to look formally at a tool to use a randomized experiment to estimate a relationship of interest: the method of instrumental variables.

Random Assignment as an Instrumental Variable

- The question: How much does education improve cognitive scores, and wages.
- Notation:

$$Y_i = \alpha + \beta A_i + \epsilon_i$$

  where A_i is Whether individual i goes to secondary school , and Y_i is earnings
- Note that this formulation assumes that the effect of education is the same for all people. We will discuss how to interpret IV when this is not true in a bit.

Randomized Scholarship

- Pascaline Dupas, Michael Kremer and I have conducted a randomized experiment in Ghana which makes it more likely to that students who qualify for high school actually attend: a scholarship program. Scholarship were randomly assigned to students who qualified for secondary school on a basis of a set of competitive test scores but had not yet enrolled.
- Let Z_i be a dummy variable equal to 1 if one is assigned to the treatment group (and were therefore offered the scholarship), 0 otherwise.
- Receiving a scholarship increases the probability to ever enroll in high school by 33% for females, 36% for males.

Scholarship and participation in Senior High School

Table 2: Survey Rate and Educational Outcomes at 5-yr Follow-up

	Female			Male		
	Treatment Mean	Control Mean	Difference (SE)	Treatment Mean	Control Mean	Difference (SE)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Survey Rate						
Surveyed in person in 2013	0.97	0.966	0.004	0.957	0.969	-0.012
			(0.012)			(0.012)
Observations	333	701		345	671	
Panel B. Educational Outcomes						
Ever enrolled in SHS	0.78	0.446	0.334	0.927	0.569	0.358
			(0.032)***			(0.029)***
Completed SHS	0.576	0.244	0.332	0.706	0.323	0.383
			(0.031)***			(0.031)***
Started and Stopped SHS	0.146	0.072	0.073	0.161	0.089	0.071
			(0.02)***			(0.021)***
Enrolled in SHS other than admission SHS[a]	0.015	0.004	0.011	0.012	0.008	0.004
			(0.006)*			(0.006)
Still enrolled in SHS	0.043	0.124	-0.081	0.048	0.149	-0.101
			(0.02)***			(0.021)***
Ever enrolled in SHS track...						
Agricultural Science	0.068	0.012	0.056	0.094	0.055	0.039
			(0.011)***			(0.017)**
Business	0.105	0.081	0.024	0.212	0.142	0.071
			(0.019)			(0.025)***
Technical Skills	0.003	0.001	0.002	0.058	0.048	0.01
			(0.003)			(0.015)
Home Economics	0.186	0.08	0.106	0.015	0	0.015
			(0.021)***			(0.005)***
Visual Arts	0.003	0.009	-0.006	0.109	0.043	0.066

Combining the two: an instrumental variables estimate of the effect of going to school on cognitive scores

Effect of treatment on participation can be measured by :

$$E[A_i|Z_i = 1] - E[A_i|Z_i = 0] \tag{1}$$

Effect of treatment on cognitive test scores can be measured by:

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] \tag{2}$$

Using our expression for Y_i, we have:

$$E[Y_i|Z_i = 1] = \alpha + \beta E[A_i|Z_i = 1] + E[\epsilon_i|Z_i = 1]$$

and:

$$E[Y_i|Z_i = 0] = \alpha + \beta E[A_i|Z_i = 0] + E[\epsilon_i|Z_i = 0]$$

Therefore

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

$$\beta(E[A_i|Z_i = 1] - E[A_i|Z_i = 0])+$$

$$E[\epsilon_i|Z_i = 1] - E[\epsilon_i|Z_i = 0]$$

Now assume about $E[\epsilon_i|Z_i = 1] - E[\epsilon_i|Z_i = 0]$ [we will comment this assumption in a minute]
Putting everything together:

$$\hat{\beta} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[A_i|Z_i = 1] - E[A_i|Z_i = 0]} \tag{3}$$

Three conditions make Z a good instrument

1. It affects A_i : $E[A_i|Z_i = 1] - E[A_i|Z_i = 0]$
2. It is randomly assigned, or as good as randomly assigned, so that $E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$ can be interpreted as the causal effect of Z on Y [when we use RCT as an instrument this is guaranteed, otherwise it needs to be checked]
3. It has no direct effect on Y (exclusion restriction). This may or may not be true, has to be argued on a case by case basis, and cannot be tested.

RCT as IV

$$\hat{\beta} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[A_i|Z_i = 1] - E[A_i|Z_i = 0]}$$

- Careful: never forget to check *both* conditions when thinking about using an instrument. The third condition is often not verified even when the first is.
- For example, in this example, could the scholarships per se be having an effect on cognitive scores?
- If the assumptions are valid: We obtain the effect of health on knowledge by dividing the effect of the program on cognitive scores by the effect of the program on education.

$$\hat{\beta} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[A_i|Z_i = 1] - E[A_i|Z_i = 0]}$$

Equation 1 is the *first stage* relationship (the denominator).
Equation 2 is the *reduced form* relationship (the numerator).
Therefore, the Wald Estimate is the reduced form divided by the first stage. $\hat{\beta}$, given by the equation above, is the *Wald estimate* of the effect of SHS participation on Y_i. It is the simplest form of the instrumental variable estimator (Z_i is our instrument).

Scholarship and participation in Senior High School

Table 2: Survey Rate and Educational Outcomes at 5-yr Follow-up

	Female			Male		
	Treatment Mean	Control Mean	Difference (SE)	Treatment Mean	Control Mean	Difference (SE)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Survey Rate						
Surveyed in person in 2013	0.97	0.966	0.004	0.957	0.969	-0.012
			(0.012)			(0.012)
Observations	333	701		345	671	
Panel B. Educational Outcomes						
Ever enrolled in SHS	0.78	0.446	0.334	0.927	0.569	0.358
			(0.032)***			(0.029)***
Completed SHS	0.576	0.244	0.332	0.706	0.323	0.383
			(0.031)***			(0.031)***
Started and Stopped SHS	0.146	0.072	0.073	0.161	0.089	0.071
			(0.02)***			(0.021)***
Enrolled in SHS other than admission SHS[a]	0.015	0.004	0.011	0.012	0.008	0.004
			(0.006)*			(0.006)
Still enrolled in SHS	0.043	0.124	-0.081	0.048	0.149	-0.101
			(0.02)***			(0.021)***
Ever enrolled in SHS track...						
Agricultural Science	0.068	0.012	0.056	0.094	0.055	0.039
			(0.011)***			(0.017)**
Business	0.105	0.081	0.024	0.212	0.142	0.071
			(0.019)			(0.025)***
Technical Skills	0.003	0.001	0.002	0.058	0.048	0.01
			(0.003)			(0.015)
Home Economics	0.186	0.08	0.106	0.015	0	0.015
			(0.021)***			(0.005)***
Visual Arts	0.003	0.009	-0.006	0.109	0.043	0.066

Scholarship and cognitive test scores

Table 3: Impact on General Intelligence and Cognitive Skills

	Female			Male			All
	Treatment Mean (SD)	Control Mean (SD)	T-C Difference (SE)	Treatment Mean (SD)	Control Mean (SD)	T-C Difference (SE)	T-C Difference (SE)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Scores on General Intelligence Tests							
Memory for Digit Span (Forward)	7.350	7.38	-0.03	7.764	7.714	0.049	0.015
	(2.587)	(2.597)	(0.175)	(2.606)	(2.513)	(0.172)	(0.123)
Memory for Digit Span (Backward)	4.402	4.374	0.028	4.9	4.714	0.186	0.113
	(1.835)	(1.676)	(0.117)	(1.918)	(1.874)	(0.128)	(0.087)
Raven's Progressive Matrices	6.52	6.558	-0.037	7.403	7.368	0.035	0.012
	(2.512)	(2.588)	(0.173)	(2.427)	(2.573)	(0.171)	(0.123)
Panel B. Performance on Reading and Math Skills Test							
Standardized score, Reading Test	0.042	-0.096	0.138	0.242	0.1	0.142	0.143
(6 questions)	(1.019)	(1.047)	(0.07)**	(0.852)	(0.939)	(0.062)**	(0.047)***
Standardized score, Math Test	-0.026	-0.189	0.163	0.272	0.197	0.074	0.124
(10 questions)	(0.998)	(0.990)	(0.067)**	(0.949)	(0.972)	(0.065)	(0.048)***
Total Standardized Score	0.006	-0.174	0.18	0.306	0.181	0.125	0.158
	(1.061)	(1.022)	(0.07)**	(0.912)	(0.944)	(0.063)**	(0.048)***
Full effort on Test (as assessed by surveyor)	0.598	0.558	0.04	0.703	0.663	0.04	0.042
	(0.491)	(0.497)	(0.033)	(0.458)	(0.473)	(0.032)	(0.023)*
Observations	323	679	1002	330	651	981	1983

Source: Duflo et al. (2014) "Estimating the Impact and Cost-Effectiveness of Expanding Access to Secondary Education in Ghana"

Scholarship and earnings –at 24

Panel B. Outcomes at 7-yr follow-up (2015)							
Enrolled in formal study / training	0.098	0.075	0.023	0.081	0.079	0.002	0.012
	(0.298)	(0.264)	(0.019)	(0.273)	(0.269)	(0.018)	(0.013)
Apprentice	0.058	0.087	-0.029	0.039	0.085	-0.046	-0.037
	(0.235)	(0.282)	(0.018)	(0.194)	(0.278)	(0.017)***	(0.012)***
Wage worker	0.202	0.19	0.012	0.359	0.337	0.022	0.02
	(0.402)	(0.393)	(0.027)	(0.481)	(0.473)	(0.032)	(0.021)
Day or Seasonal Laborer	0.031	0.016	0.015	0.084	0.054	0.029	0.023
	(0.173)	(0.125)	(0.010)	(0.278)	(0.227)	(0.016)*	(0.01)**
Farming	0.055	0.084	-0.029	0.093	0.089	0.004	-0.012
	(0.229)	(0.278)	(0.018)	(0.291)	(0.285)	(0.019)	(0.013)
Working for own or family business	0.193	0.225	-0.032	0.087	0.113	-0.026	-0.031
	(0.395)	(0.418)	(0.028)	(0.282)	(0.317)	(0.021)	(0.017)*
No occupation	0.337	0.318	0.02	0.246	0.239	0.007	0.012
	(0.474)	(0.466)	(0.031)	(0.431)	(0.427)	(0.029)	(0.021)
Actively searching for a job	0.336	0.237	0.1	0.31	0.34	-0.03	0.036
	(0.473)	(0.425)	(0.029)***	(0.463)	(0.474)	(0.031)	(0.021)*
If no occupation: actively searching for a jo	0.618	0.452	0.166	0.573	0.677	-0.104	0.053
	(0.488)	(0.499)	(0.058)***	(0.498)	(0.469)	(0.065)	(0.044)
Total earnings last month (GHC)	72.7	58.7	14.0	161.6	154.2	7.5	12.2
	(138.7)	(116.0)	(8.318)*	(227.1)	(229.2)	(15.3)	(9.0)
Earned no money last month	0.613	0.612	0.001	0.407	0.431	-0.023	-0.014
	(0.488)	(0.488)	(0.033)	(0.492)	(0.496)	(0.033)	(0.024)
Observations	326	689	1015	334	662	996	2011

The Wald Estimate

Let us calculate the Wald estimator ourselves, for cognitive scores.

- Effect on cognitive scores? WOMEN 0.185 Standard deviation of tests MEN 0.125 standard deviation.
- Effect on years of schooling? WOMEN 1.227 years MEN 1.115 years
- Effect of years of schooling on cognitive scores? WOMEN MEN
- How about earnings?

The importance of the exclusion restriction

- You can see that even a "small" violation of either of the conditions for the validity of the instrument can result in very large bias. Any bias in the reduced form will be "blown up" when I divide by the first stage difference.
- Let's consider some examples. Valid, no valid? [hint: there is both here!]
  - Doctors are randomly selected to receive advice to remind their patients that it is flu season and they should take a flu shot, can we use it as an instrument for taking the flu shot, so we can estimate the impact of taking the flu shot on sick days?
  - Kids who apply to charter schools are picked randomly if the school is oversubscribed.
  - Kids are in school that are randomly receive deworming pills, I use being in treatment school are instrument for actually being de-wormed (hint: worms are highly contagious).

The interpretation of IV when the treatment effect is not constant

- In reality the effect of going to school on test scores is likely to be different for different children. [remember that when we first introduced causality we did not assume constant treatment effect]
- In that case the simple calculation we just did does not apply
- Yet, under a fairly mild assumption, the Wald estimate still has a causal interpretation, which is in fact quite intuitive: it captures the effect of the treatment on those who are compelled by the instrument to get treated: this is the *Local Average Treatment Effect*, or LATE.

Hmm... what?

- Back to our school example, and consider the binary decision of going to high school or not.
- There could in principle be 4 groups of kids:
  1. Those who would go to high school anyways *Always Takers*
  2. Those who would not go to high school even when offered a scholarship *Never Takers*
  3. Those who would not go if not offered, but go if offered *Compliers*
  4. Those who would go if not offered, but not go if offered *Defiers*
- Now that last group is a bit weird, no? So let's assume it does not exist
- In that case, just a few easy lines of Algebra that I will spare you are enough to prove that the Wald estimate is the effect for the compliers. The trick to the proof is that the first stage for the other two groups is zero: only the compliers contribute to the first stage.

From the Wald Estimate to two state Least Squares

- Instead of computing differences in means and taking the ratio, we could have couched this in a regression framework.
- First stage , $\widehat{\pi_1}$ in the equation: $A_i = \pi_0 + \pi_1 Z_i + \upsilon_i$
- Reduced form : $\widehat{\gamma_1}$ in the equation $Y_i = \gamma_0 + \gamma_1 Z_i + \omega_i$
- Two stage least square: Run the first stage, and take the fitted values $\widehat{A_i}$,
- Then, in the second stage, run: $Y_i = \alpha + \beta \widehat{A_i} + \epsilon_i$

The two stage least squares and the Wald estimates are identical

$$\widehat{\beta} = \frac{Cov(Y_i, \widehat{A}_i)}{Var(\widehat{A}_i)}$$

$$= \frac{Cov(Y_i, \pi_0 + \pi_1 Z_i)}{Var(\pi_0 + \pi_1 Z_i)}$$

$$= \frac{\pi_1 Cov(Y_i, Z_i)}{\pi_1^2 Var(Z_i)} = \frac{\gamma_1}{\pi_1}$$

More generally: two stage least squares

Consider the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

where X_{i1} a vector of endogenous variables, and X_{2i} some variables that you assume are not endogenous (control variables).
Look for an instrument. In this case you will need at least one instrument, you could have more. Denote Z the matrix $(Z_1, ...Z_k, X_2)$ [in other words the control variables, which do not need to be instrumented, are part of the matrix of instruments.

Intuitive steps:

- First stage: $X_{1i} = \pi_o + \pi_1 Z_1 + \pi_2 Z_2 + \cdots + \pi_k Z_k + X_{2i} + \omega_i$
- Second stage: $Y = \beta_0 + \beta_1 \widehat{X_{1i}} + X_{i2} + \epsilon_i$

In practice if you do that point estimates will be correct, but the standard errors and all the tests will be wrong (because you have estimated your first stage, rather than knowing it, and the standard errors must reflect this uncertainty).

Two stage least square in reality

- In reality, you run two stage least square in one stage,
- Specify your Y, your X, your Z
- If Z and X have the same number of variable (e.g. if you have chosen one instrument for one endogenous variables, and included the control variable in the matrix of instruments), then the 2SLS formula is:

$$\widehat{\beta} = (Z'X)^{-1}Z'Y$$

and the variance is

$$Var(IV) = \sigma^2(Z'X)^{-1}Z'Z(Z'X)^{-1}$$

- If there are more instrument than endogenous variable, the formula is a big longer, but the idea remains just the same: it will project the X onto the Z and take the projected value .

IV in R: Test scores on SHS completion, females

```
Call:
ivreg(formula = total_score ~ shs_complete + region.f | treatment +
    region.f, data = data_female)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8707 -0.5674  0.1525  0.7328  2.3780

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.44686    0.09991  -4.473 8.62e-06 ***
shs_complete  0.64215    0.21502   2.986 0.002891 **
region.f2    -0.25133    0.09557  -2.630 0.008677 **
region.f3    -0.12955    0.08839  -1.466 0.143033
region.f4     0.26889    0.07946   3.384 0.000743 ***
region.f5     0.38052    0.12033   3.162 0.001612 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9315 on 989 degrees of freedom
Multiple R-Squared: 0.2206,    Adjusted R-squared: 0.2167
Wald test: 11.39 on 5 and 989 DF,  p-value: 1.051e-10
```

IV in R: Test scores on SHS completion, females with controls

```
Call:
ivreg(formula = total_score ~ . - treatment | . - shs_complete,
    data = data_female_controls)

Residuals:
     Min       1Q   Median       3Q      Max
-2.78926 -0.58436  0.05948  0.65158  2.51616

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.21286    0.68435  -0.311  0.75584
shs_complete     0.63134    0.23362   2.702  0.00701 **
region.f2       -0.18561    0.09969  -1.862  0.06294 .
region.f3       -0.13692    0.08934  -1.533  0.12573
region.f4        0.26212    0.08236   3.182  0.00151 **
region.f5        0.33882    0.12379   2.737  0.00632 **
age             -0.08152    0.02868  -2.842  0.00458 **
base_bece_score  2.37519    0.44933   5.286 1.56e-07 ***
hhh_highest_edu  0.04214    0.02164   1.948  0.05175 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9118 on 912 degrees of freedom
Multiple R-Squared: 0.2705,    Adjusted R-squared: 0.2641
Wald test: 22.14 on 8 and 912 DF,  p-value: < 2.2e-16
```

IV in R: Test scores on SHS completion, males

```
Call:
ivreg(formula = total_score ~ shs_complete + region.f | treatment +
    region.f, data = data_male)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8406 -0.5252  0.1207  0.6293  2.2500

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.08608    0.11700  -0.736 0.462114
shs_complete  0.47801    0.20212   2.365 0.018227 *
region.f2    -0.16118    0.09187  -1.754 0.079679 .
region.f3    -0.15501    0.07680  -2.018 0.043817 *
region.f4     0.25938    0.07158   3.624 0.000306 ***
region.f5     0.32959    0.11178   2.949 0.003268 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8388 on 968 degrees of freedom
Multiple R-Squared: 0.1514,    Adjusted R-squared: 0.147
Wald test: 11.21 on 5 and 968 DF,  p-value: 1.613e-10
```

IV in R: Test scores on SHS completion, males with controls

```
Call:
ivreg(formula = total_score ~ . - treatment | . - shs_complete,
    data = data_male_controls)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0122 -0.5004  0.1246  0.6095  2.3681

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.28464    0.53472   0.532 0.594639
shs_complete     0.42590    0.21476   1.983 0.047656 *
region.f2       -0.08622    0.09835  -0.877 0.380910
region.f3       -0.15138    0.07939  -1.907 0.056868 .
region.f4        0.24712    0.07496   3.297 0.001016 **
region.f5        0.31591    0.11466   2.755 0.005986 **
age             -0.07252    0.02007  -3.613 0.000319 ***
base_bece_score  1.85472    0.39272   4.723  2.7e-06 ***
hhh_highest_edu  0.03314    0.02007   1.651 0.099120 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8343 on 898 degrees of freedom
Multiple R-Squared: 0.1891,    Adjusted R-squared: 0.1818
Wald test: 15.38 on 8 and 898 DF,  p-value: < 2.2e-16
```

IV in R: Test scores on SHS completion, R code

```
## Load packages, load in the data
library("AER")
setwd("/Users/MaddieDuhon/Dropbox (MIT)/14.31 edX/Exercises/Ghana")
data<-read.csv("Ghana_data.csv")

## Prep region fixed effects
data$region.f <- factor(data$region)

## Subset male and female
data_female<-subset(data,gender==0)
data_male<-subset(data,gender==1)

## Subset data to include controls
controls <- c("total_score", "shs_complete", "treatment", "region.f", "age", "base_bece_score", "hhh_highest_edu")
data_female_controls<-data_female[controls]
data_male_controls<-data_male[controls]

## IV Regression: Standardized test scores on secondary schooling completion
## Female (without and with controls)
summary(ivreg(total_score ~ shs_complete +region.f | treatment+region.f, data=data_female))
summary(ivreg(total_score ~ .-treatment | .-shs_complete, data=data_female_controls))

## Males (without and with controls)
summary(ivreg(total_score ~ shs_complete +region.f | treatment+region.f, data=data_male))
summary(ivreg(total_score ~ .-treatment | .-shs_complete, data=data_male_controls))
```

14.310x Data Analysis for Social Scientists
Spring 2023