Instrumental Variables

EDLD 650: Week 6

David D. Liebowitz

Agenda

1. Roadmap and goals (9:00-9:10)

2. Discussion Questions (9:10-10:20)

   - Murnane and Willett
   - Angrist et al. (x2)
   - Dee

3. Break (10:20-10:30)

4. Applied instrumental variables (10:30-11:40)

5. Wrap-up (11:40-11:50)

   - DARE #3 prep
   - Plus/deltas

Roadmap

Causal identification problem:
$Y_i(D_i)$: outcome for individual i, where D_i = 1 if treated, 0 if control
$ITE_i = Y_i(1) - Y_i(0)$; but unobserved so:
$$\widehat{ATE} = \left(\sum_{i=1}^{n_1} Y_i \Big/ n_1\right) - \left(\sum_{i=1}^{n_0} Y_i \Big/ n_0\right)$$

Is assignment randomized?

YES

$E[\varepsilon_i|X_i] = E[\varepsilon_i|D_i] = 0$; $\varepsilon_i \perp D_i$

-Check balance: $\bar{X}_{D=1} \cong \bar{X}_{D=0}$
-Estimate effects:
$\bar{Y}_{D=1} - \bar{Y}_{D=0}$; **or**
$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$
$\widehat{\beta_1} = \widehat{ATE} = \widehat{ATT}$
-Sampling approach/External validity
-Power?
-Internal validity:
Spillover (SUTVA)? Hawthorne/John Henry? Non-compliance? Attrition?

ITT v. TOT

NO, exogenous variation in D_i

$E[\varepsilon_i| D_i] \neq 0$

-Difference-in-Differences (DD)
$\widehat{ATE} = (\bar{Y}(1)_{t=1} - \bar{Y}(1)_{t=0}) - (\bar{Y}(0)_{t=1} - \bar{Y}(0)_{t=0})$;
or $Y_{it} = \beta_1 D_{it} + \Gamma_i + \Pi_t + \varepsilon_{it}$
→ D_{it}: 1 if unit under treatment in time t, otherwise 0
Common trends? Simultaneous shocks? Time-varying effects? Weighting?

-Regression Discontinuity (RD)
given forcing var $D_i = (S_i \geq C)$
$Y_i = \beta_0 + \beta_1 S_i + 1(S_i \geq C)\beta_2 + \varepsilon_i$
Smooth and non-manipulated forcing variable? Bandwidth? Functional form? Generalizability/LATE? Sharp or fuzzy?

-Instrumental Variables (IV)
Z_i correl w/ D_i but $Z_i \perp \varepsilon_i$
1st: $D_i = \alpha_0 + \alpha_1 Z_i + \delta_i$
2nd: $Y_i = \beta_0 + \beta_1 \widehat{D_i} + \varepsilon_i$

NO, D_i based on observables

$E[\varepsilon_i| D_i, X_i] = 0$

-OLS
-Multi-level modeling
Random: $Y_{ij} = \gamma_0 + \gamma_1 D_{ij} + (\varepsilon_{ij} + \upsilon_j)$
Fixed: $Y_{ij} = \Sigma_1^J \alpha_j S_{ij} + \gamma_1 D_{ij} + \varepsilon_{ij}$
-Matching estimators
-PSM
-CEM
-IPW

Goals

1. Describe conceptual approach to instrumental variables (IV) analysis
2. Assess validity of IV assumptions in applied context
3. Conduct IV analysis in simplified data and interpret results

So random...

Break

The PACES experiment

- Recall the PACES school voucher experiment (Angrist et al. 2002) from *Methods Matter*, Chapter 11
- Lottery assignment for vouchers to attend private school in Colombia
- What is the **main outcome**?
- What is the **endogenous regressor**?

Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment

By Joshua Angrist, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer*

Estimand of interest: *effect of using financial aid to attend private school*

Let's replicate!

```
paces ← read.csv(here("./data/ch11_PACES.csv"))
DT::datatable(paces[,c(1:7)], fillContainer = FALSE, options =
                list(pageLength = 5))
```

Show 5 entries Search:

	id	won_lottry	male	base_age	finish8th	use_fin_aid	school
1	3	1	0	11	1	1	2
2	4	0	1	11	1	1	2
3	5	0	1	11	1	0	1
4	6	0	0	9	0	0	1
5	10	1	1	11	1	1	5

Showing 1 to 5 of 1,171 entries

Previous 1 2 3 4 5 … 235 Next

First post-randomization task?

Causal identification problem:

$Y_i(D_i)$: outcome for individual i, where D_i = 1 if treated, 0 if control

$ITE_i = Y_i(1) - Y_i(0)$; but unobserved so:

$$\widehat{ATE} = \left(\sum_{i=1}^{n_1} Y_i / n_1\right) - \left(\sum_{i=1}^{n_0} Y_i / n_0\right)$$

Is assignment randomized?

YES

$E[\varepsilon_i|X_i] = E[\varepsilon_i|D_i] = 0$; $\varepsilon_i \perp D_i$

- -Check balance: $\bar{X}_{D=1} \cong \bar{X}_{D=0}$
- -Estimate effects:

  $\bar{Y}_{D=1} - \bar{Y}_{D=0}$; **or**

  $Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$

  $\widehat{\beta_1} = \widehat{ATE} = \widehat{ATT}$
- -Sampling approach/External validity
- -Power?
- -Internal validity:

  Spillover (SUTVA)? Hawthorne/John Henry? Non-compliance? Attrition?

ITT v. TOT

NO, exogenous variation in D_i

$E[\varepsilon_i| D_i] \neq 0$

-Difference-in-Differences (DD)

$\widehat{ATE} = (\bar{Y}(1)_{t=1} - \bar{Y}(1)_{t=0}) - (\bar{Y}(0)_{t=1} - \bar{Y}(0)_{t=0})$;

or $Y_{it} = \beta_1 D_{it} + \Gamma_i + \Pi_t + \varepsilon_{it}$

→ D_{it}: 1 if unit under treatment in time t, otherwise 0

Common trends? Simultaneous shocks? Time-varying effects? Weighting?

-Regression Discontinuity (RD)

given forcing var $D_i = (S_i \geq C)$

$Y_i = \beta_0 + \beta_1 S_i + 1(S_i \geq C)\beta_2 + \varepsilon_i$

Smooth and non-manipulated forcing variable? Bandwidth? Functional form? Generalizability/LATE? Sharp or fuzzy?

-Instrumental Variables (IV)

Z_i correl w/ D_i but $Z_i \perp \varepsilon_i$

1st: $D_i = \alpha_0 + \alpha_1 Z_i + \delta_i$

2nd: $Y_i = \beta_0 + \beta_1 \widehat{D_i} + \varepsilon_i$

NO, D_i based on observables

$E[\varepsilon_i| D_i, X_i] = 0$

- -OLS
- -Multi-level modeling

  Random: $Y_{ij} = \gamma_0 + \gamma_1 D_{ij} + (\varepsilon_{ij} + \upsilon_j)$

  Fixed: $Y_{ij} = \Sigma_1^J \alpha_j S_{ij} + \gamma_1 D_{ij} + \varepsilon_{ij}$
- -Matching estimators
  - -PSM
  - -CEM
  - -IPW

Balance checks

Examine by covariates:

$$\bar{X}_{D=1} \approxeq \bar{X}_{D=0}$$

```r
random <- arsenal::tableby(won_lottry ~ male + base_age, paces)
summary(random)
```

	0 (N=579)	1 (N=592)	Total (N=1171)	p value
male				0.980
Mean (SD)	0.504 (0.500)	0.505 (0.500)	0.505 (0.500)	
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	
base_age				0.422
Mean (SD)	12.036 (1.352)	11.973 (1.343)	12.004 (1.347)	
Range	7.000 - 16.000	9.000 - 17.000	7.000 - 17.000	

Balance checks

Omnibus F-test approach:

```
summary(lm(won_lottry ~ male + base_age, data=paces))
```

```
...
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)  0.609897   0.131294   4.645 3.78e-06 ***
#> male         0.002568   0.029338   0.088    0.930    
#> base_age    -0.008800   0.010894  -0.808    0.419    
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.5005 on 1168 degrees of freedom
#> Multiple R-squared:  0.000559,	Adjusted R-squared:  -0.001152 
#> F-statistic: 0.3266 on 2 and 1168 DF,  p-value: 0.7214
...
```

A naïve estimate of financial aid

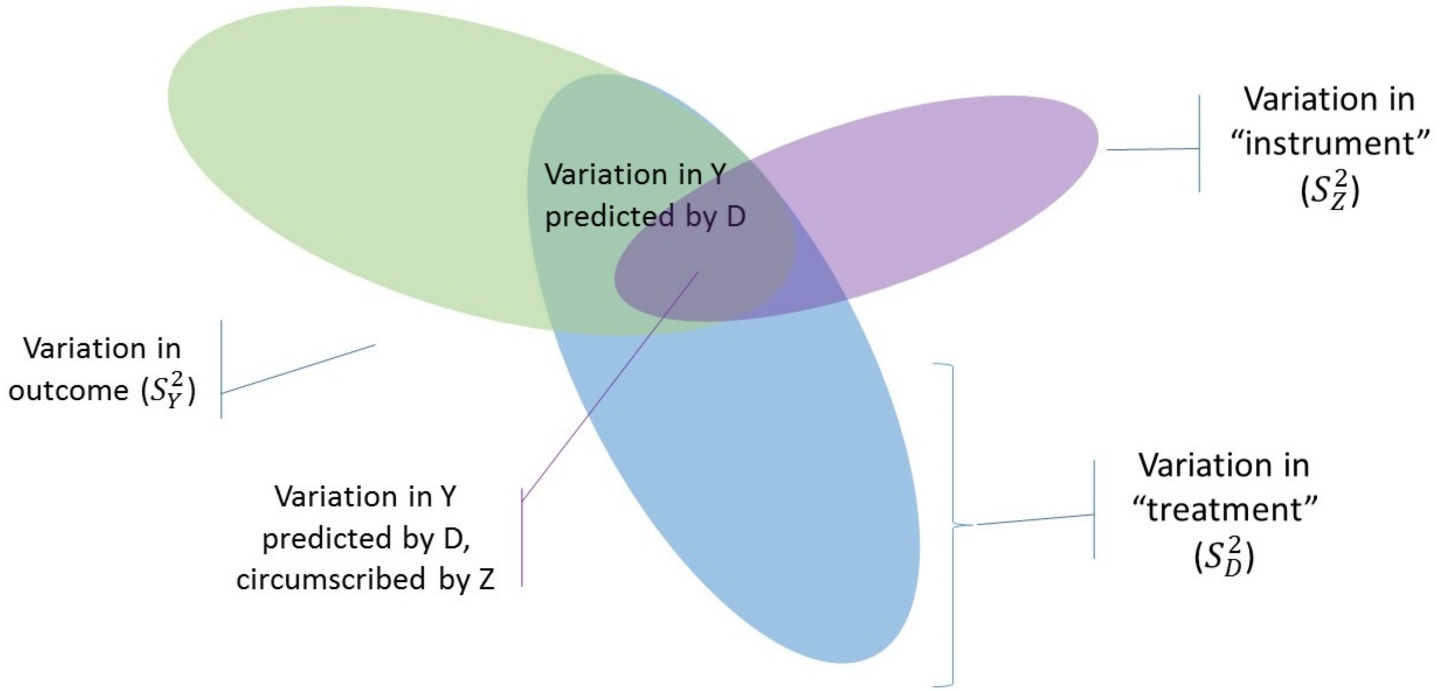
```r
ols1 <- lm(finish8th ~ use_fin_aid, data=paces)
ols2 <- lm(finish8th ~ use_fin_aid + base_age + male, data=paces)
```

	(1)	(2)
use_fin_aid	0.133***	0.121***
	(0.027)	(0.027)
base_age		-0.063***
		(0.010)
male		-0.086**
		(0.026)
Observations	1,171	1,171
R^2	0.020	0.064

What's wrong with the naïve approach?

> Only about 90 percent of lottery winners used the private school voucher to pay for private school and 24 percent of lotter losers found other sources of scholarships for which to pay for private school. There are endogenous differences in the expected outcomes of children from families who chose to both use the voucher and those who secured scholarship funding from sources outside the voucher lottery. The policy relevant question is how a public subsidy of private school might affect educational attainment for children from low-income families in Bogota, Colombia. The naïve approach does not identify these effects but rather the combination of voucher subsidy and endogenous unobservables across families and individuals.

How could IV address?



IV estimate: ratio of area of *overlap of Y and Z* to area of *overlap of D and Z*. Depends entirely on variation in Z that predicts variation in Y and D:

$$\hat{\beta}_1^{IVE} = \frac{S_{YD}}{S_{DZ}}$$

a **Local Average Treatment Effect**

Recall 2SLS set-up

1st stage:

Regress the endogenous treatment (D_i) on instrumental variable (Z_i):

$$D_i = \alpha_0 + \alpha_1 Z_i + \nu_i$$

Obtain the *predicted values* of the treatment $(\hat{D}_i)$ from this fit.

2nd stage:

Regress the outcome (Y_i) on the predicted values of the treatment $(\hat{D}_i)$:

$$Y_i = \beta_0 + \beta_1 \hat{D}_i + \varepsilon_i$$

Think about this in the Colombia PACES experiment context. What is the **main outcome**? What is the **endogenous regressor**? What is the **instrument**? Can you write the two-stage equation without consulting the next slide or book?

The PACES Scholarship

1st stage:

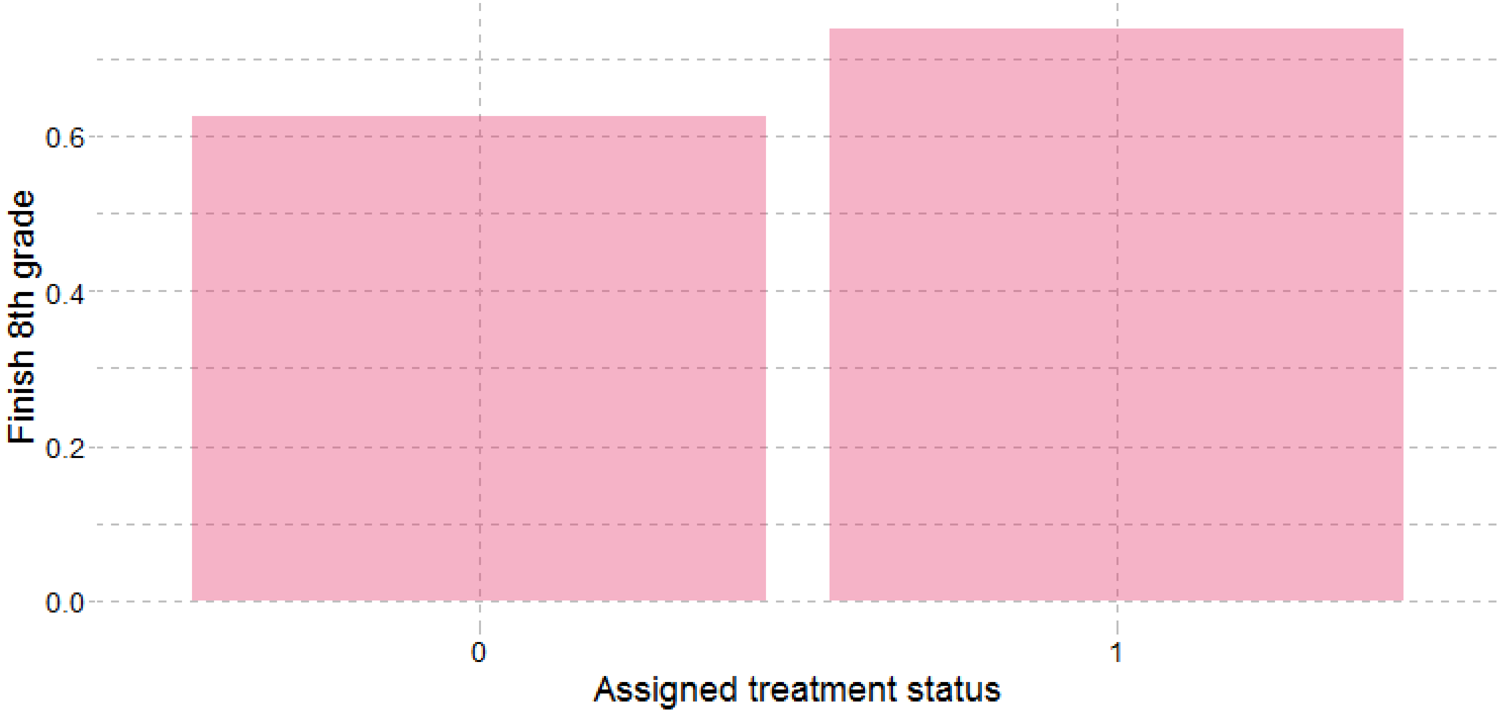
$$USEFINAID_i = \alpha_0 + \alpha_1 WONLOTTERY_i + \nu_i$$

2nd stage:

$$FINISH8TH_i = \beta_0 + \beta_1 USEF\hat{I}NAID_i + \varepsilon_i$$

What is the main outcome? What is the endogenous regressor? What is the instrument? What are the assumptions?

Outcome by lottery status



This represents an important substantive finding... **can you interpret what it is?**

A simple t-test

```r
ttest <- t.test(finish8th ~ won_lottry, data=paces)
ttest
```

```
#> 
#>     Welch Two Sample t-test
#> 
#> data:  finish8th by won_lottry
#> t = -4.1077, df = 1153.5, p-value = 4.279e-05
#> alternative hypothesis: true difference in means between group 0 and grou
#> 95 percent confidence interval:
#>  -0.16441869 -0.05812251
#> sample estimates:
#> mean in group 0 mean in group 1 
#>       0.6252159       0.7364865
```

Can you interpret what this means?

Intent-to-Treat Estimates

```
itt1 <- lm(finish8th ~ won_lottry, data=paces)
itt2 <- lm(finish8th ~ won_lottry + base_age + male, data=paces)
itt3 <- lm(finish8th ~ won_lottry + base_age + male +
              as.factor(school), data=paces)
```

Table 1. Intent-to-Treat Estimates of Winning the PACES lottery on 8th Grade Completion

	Model 1	**Model 2**	**Model 3**
Won Lottery	0.111***	0.107***	0.108***
	(0.027)	(0.026)	(0.027)
Starting Age		-0.065***	-0.064***
		(0.010)	(0.010)
Male		-0.088***	-0.089***
		(0.027)	(0.027)
School Fixed Effects	No	No	Yes
Num.Obs.	1171	1171	1171

Notes: The table displays coefficients from Equation X and standard errors in parentheses.

Intent-to-Treat Estimates

What is our estimand of interest? Do these estimates represent that?

Table 1. Intent-to-Treat Estimates of Winning the PACES lottery on 8th Grade Completion

	Model 1	Model 2	Model 3
Won Lottery	0.111***	0.107***	0.108***
	(0.027)	(0.026)	(0.027)
Starting Age		-0.065***	-0.064***
		(0.010)	(0.010)
Male		-0.088***	-0.089***
		(0.027)	(0.027)
School Fixed Effects	No	No	Yes
Num.Obs.	1171	1171	1171

Notes: The table displays coefficients from Equation X and standard errors in parentheses.

Implementing IV in regression

Reminder of key assumptions:

1. Instrument correlated with endogenous predictor (no **"weak" instruments**)
2. Instrument not correlated with 1st stage residuals $(\sigma_{Z\nu} = 0)$
3. Instrument not correlated with 2nd stage residuals $(\sigma_{Z\varepsilon} = 0)$ and correlated with outcome only via predictor[1]
    - Exclusion restriction means **NO THIRD PATH!**

Practical considerations:

Can implement this various ways. Pedagogically, we'll implement 2SLS using the `fixest` package because it allows straightforward presentation of 1st stage results. This can also be done via `ivreg` and `iv_robust` in R.

[1] Don't forget, **no defiers** too.

IV Estimation

```
# Instrument with no covariates
# With only instrumented predictor and no covariates,
# need to include a "1" in 2nd stage
tot1 <- feols(finish8th ~ 1 | use_fin_aid ~ won_lottry, data=paces)

# Instrument with covariates
# Note that these are automatically included in 1st stage
# Can include multiple instruments and multiple
# endogenous predictors
tot2 <- feols(finish8th ~ base_age + male |
                          use_fin_aid ~ won_lottry, data=paces)
```

IV results - First Stage

```
summary(tot2, stage = 1)
```

```
#> TSLS estimation, Dep. Var.: use_fin_aid, Endo.: use_fin_aid, Instr.: won_
#> First stage: Dep. Var.: use_fin_aid
#> Observations: 1,171
#> Standard-errors: IID
#>              Estimate Std. Error   t value   Pr(>|t|)
#> (Intercept)  0.432760   0.095159  4.547738 5.9880e-06 ***
#> won_lottery  0.674527   0.021014 32.098773  < 2.2e-16 ***
#> base_age    -0.015160   0.007826 -1.937178 5.2965e-02 .
#> male        -0.020257   0.021070 -0.961417 3.3654e-01
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> RMSE: 0.358813   Adj. R2: 0.469577
#> F-test (1st stage): stat = 1,030.3, p < 2.2e-16, on 1 and 1,167 DoF.
```

You will see some rules of thumb about what makes for a strong instrument (e.g., $t_F > 10$), but recent work has found that with t-ratios lower than 100 one should adjust critical value (Lee et al., 2021).

IV results - Second Stage

```
summary(tot2)
```

```
#> TSLS estimation, Dep. Var.: finish8th, Endo.: use_fin_aid, Instr.: won_l
#> Second stage: Dep. Var.: finish8th
#> Observations: 1,171
#> Standard-errors: IID
#>                  Estimate Std. Error  t value   Pr(>|t|)
#> (Intercept)      1.378128   0.123090 11.19614  < 2.2e-16 ***
#> fit_use_fin_aid  0.159000   0.039173  4.05890 5.2589e-05 ***
#> base_age        -0.062157   0.009872 -6.29603 4.3146e-10 ***
#> male            -0.085145   0.026504 -3.21258 1.3515e-03 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> RMSE: 0.451177   Adj. R2: 0.059822
#> F-test (1st stage), use_fin_aid: stat = 1,030.3    , p < 2.2e-16 , on 1 
#>                      Wu-Hausman: stat =     1.78464, p = 0.181841, on 1 
```

Can you interpret what this means?

A taxonomy of IV estimates

```
# Include school fixed effects
tot3 <- feols(finish8th ~ base_age + male | as.factor(school) |
              use_fin_aid ~ won_lottry,
              vcov = "iid",  data=paces)

# Cluster-robust standard errors
tot4 <- feols(finish8th ~ base_age + male | as.factor(school) |
              use_fin_aid ~ won_lottry,
              vcov = ~ school, data=paces)
```

Estimate voucher use effects

Table 2. Instrumental variable estimates of using financial aid to attend private school due to winning the PACES lottery on 8th grade completion

	(1)	(2)	(3)	(4)
Use Fin. Aid	0.165***	0.159***	0.161***	0.161*
	(0.040)	(0.039)	(0.039)	(0.052)
Starting Age		-0.062***	-0.062***	-0.062**
		(0.010)	(0.010)	(0.009)
Male		-0.085**	-0.086**	-0.086
		(0.027)	(0.027)	(0.037)
School FE	No	No	Yes	Yes
Num.Obs.	1171	1171	1171	1171

The table displays coefficients from Equation X and standard errors in parentheses. Model 4 uses cluster-robust standard errors at school level.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

OLS, ITT and TOT estimates

Table 3. Comparison of OLS, ITT and IV estimates of using financial aid to attend private school due to winning the PACES lottery

	(1)	(2)	(3)	(4)	(5)
	OLS	ITT	TOT	TOT	TOT
Use Fin. Aid	0.121***		0.165***	0.159***	0.161*
	(0.027)		(0.040)	(0.039)	(0.052)
Win Lottery		0.107***			
		(0.026)			
School FE	No	No	No	No	Yes
Student Chars.	Yes	Yes	No	Yes	Yes
Clust. SEs	No	No	No	No	Yes
Num.Obs.	1171	1171	1171	1171	1171

The table displays coefficients from Equation X and standard errors in parentheses.

* p < 0.05, ** p < 0.01, *** p < 0.001

Interpretation of results

The naïve OLS estimates understate the effects of a public voucher subsidy for private school attendance for over 125,000 children from low-income families in Bogota, Colombia. Our preferred estimates of the effect of voucher use on eighth-grade completion imply an increase in the on-time completion rate of 16 percentage points.

The estimates of the endogenous relationship between the use of financial aid to attend private school and school attainment (Model 1) imply that students who use any form of external scholarship are 12 percentage points more likely to complete eighth grade. In Model 2, we present results of winning an unbiased lottery to receive vouchers covering slightly more than half the cost of average private school attendance. We find that the offer of the voucher increased eighth-grade completion rates by just less than 11 percentage points. Finally, Models 3-5 present a taxonomy of Treatment-on-the-Treated estimates in which we use the randomized lottery as an instrument for the use of financial aid to attend private school. We find consistent effects 50 percent larger than the Intent-to-Treat estimates. These models are robust to the inclusion of baseline student characteristics, cohort fixed effects, and the clustering of standard errors at the level of randomization (within school).

Synthesis and wrap-up

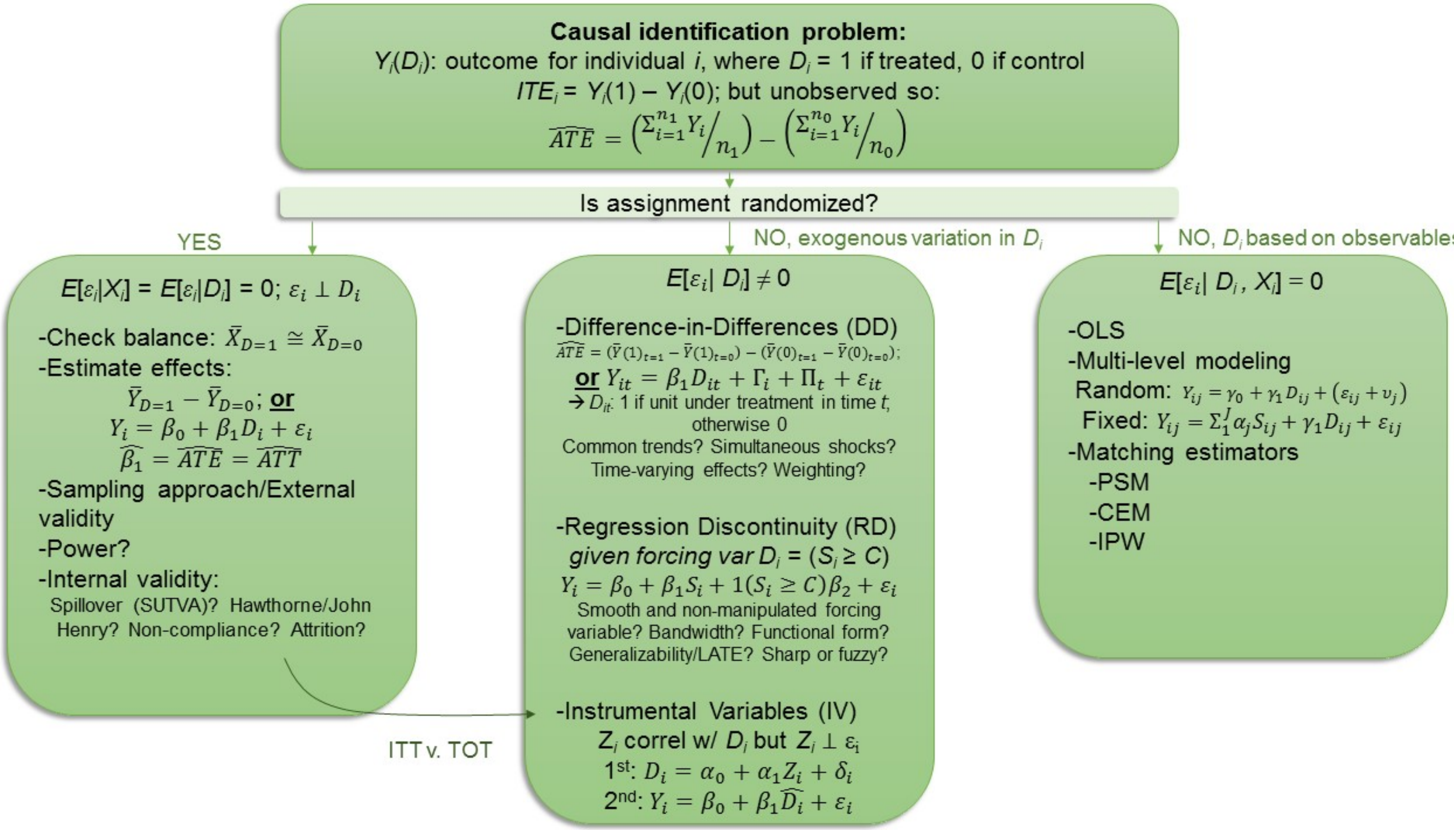
Goals

1. Describe conceptual approach to instrumental variables (IV) analysis

2. Assess validity of IV assumptions in applied context

3. Conduct IV analysis in simplified data and interpret results

Can you explain this figure?



To-Dos

Week 7: Instrumental Variables

Readings:

- Kim, Capotosto, Hartry & Fitzgerald (2011)

Assignments Due

DARE 3

- Due 9:00am, Feb. 14

Final Research Project

- Presentation, March 8
- Paper, March 18 (submit early [March 10] for feedback)

Feedback

Plus/Deltas

Front side of index card

Clear/Murky

On back