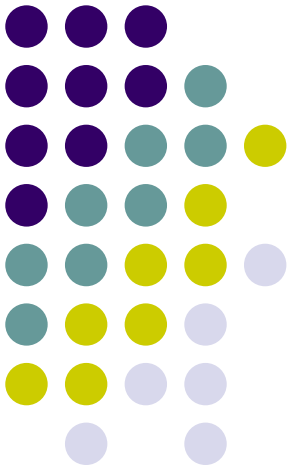
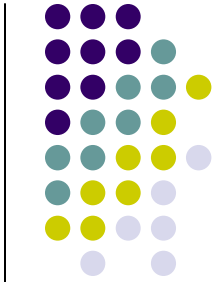
Differences-in-Differences

November 10

Clair



The Big Picture

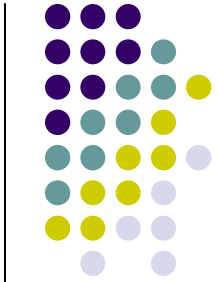


- What is this class really about, anyway?

The Big Picture

- What is this class really about, anyway?
  - Causality

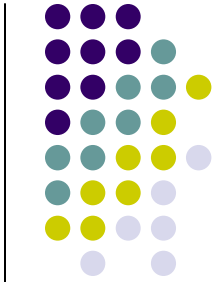
The Big Picture



- What is this class really about, anyway?
  - Causality

- What is our biggest problem?

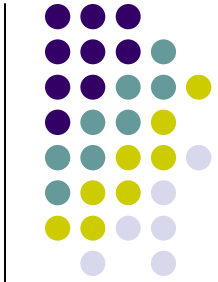
The Big Picture



- What is this class really about, anyway?
  - Causality

- What is our biggest problem?
  - Omitted variable bias

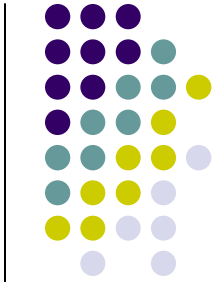
Omitted Variable Bias



- The actual cause is unobserved
  - e.g. higher wages for educated actually caused by motivation, not schooling
- Happens when people get to choose their own level of the “treatment” (broadly construed)
  - Selection bias

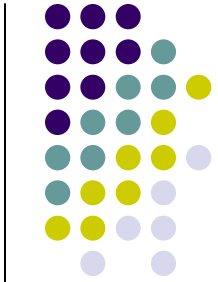
- Non-random program placement
  - Because of someone else’s choice, “control” isn’t a good *counterfactual* for treated

Math Review



(blackboard)

Math Review



for those of you looking at these slides later, here's what we just wrote down:

(1) $Yi = a + bTi + cXi + ei$

(2) $E(Yi \mid Ti=1) - E(Yi \mid Ti=0)$

$= [a + b + cE(Xi \mid Ti=1) + E(ei \mid Ti=1)]$
$\quad - [a + 0 + cE(Xi \mid Ti=0) + E(ei \mid Ti=0)]$

$= b + c\,[E(Xi \mid Ti=1) - E(Xi \mid Ti=0)]$

True effect "Omitted variable/selection bias" term

What if we had data from before the program?

What if we estimated this equation using data from before the program?

(1) $Yi = a + bTi + cXi + ei$

Specifically, what would our estimate of b be?

What if we had data from before the program?

What if we estimated this equation using data from before the program?

(1) $Yi = a + bTi + cXi + ei$

(2) $E(Yi \mid Ti=1) - E(Yi \mid Ti=0)$

$= [a + \mathbf{0} + cE(Xi \mid Ti=1) + E(ei \mid Ti=1)]$
$\quad - [a + 0 + cE(Xi \mid Ti=0) + E(ei \mid Ti=0)]$

$= c [E(Xi \mid Ti=1) - E(Xi \mid Ti=0)]$

"Omitted variable/selection bias" term

ALL THAT'S LEFT IS THE PROBLEMATIC TERM – HOW COULD THIS BE HELPFUL TO US?

Differences-in-Differences (just what it sounds like)

- Use two periods of data
  - add second subscript to denote time

$= \{E(Y_{i1} \mid T_{i1}=1) - E(Y_{i1} \mid T_{i1}=0)\}$ (difference btwn T&C, post)
$- \{E(Y_{i0} \mid T_{i1}=1) - E(Y_{i0} \mid T_{i1}=0)\}$ – *(difference btwn T&C, pre)*

$= b + c\,[E(X_{i1} \mid T_{i1}=1) - E(X_{i1} \mid T_{i1}=0)]$
$- c\,[E(X_{i0} \mid T_{i1}=1) - E(X_{i0} \mid T_{i1}=0)]$

Differences-in-Differences (just what it sounds like)

- Use two periods of data
  - add second subscript to denote time

$= \{E(Y_{i1} \mid T_{i1}=1) - E(Y_{i1} \mid T_{i1}=0)\}$ (difference btwn T&C, post)

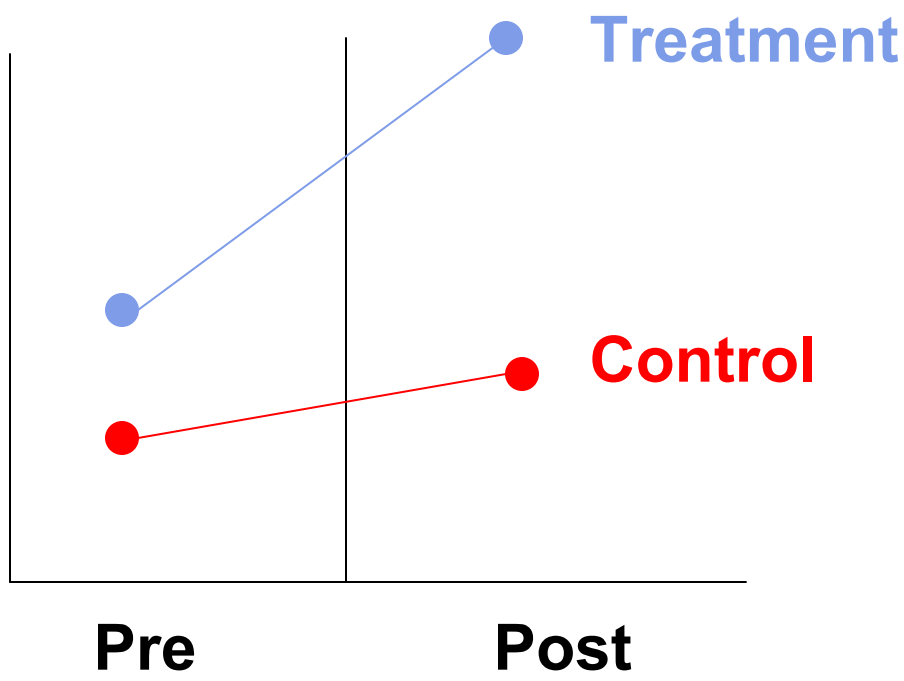
$- \{E(Y_{i0} \mid T_{i1}=1) - E(Y_{i0} \mid T_{i1}=0)\}$ – (difference btwn T&C, pre)

$= b + \cancel{c\,[E(X_{i1} \mid T_{i1}=1) - E(X_{i1} \mid T_{i1}=0)]}$

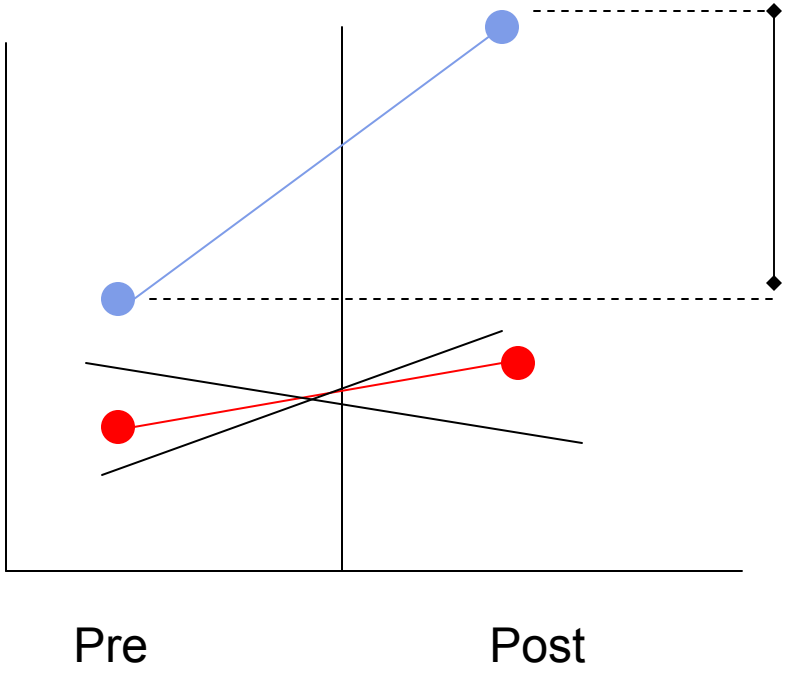
$\cancel{- c\,[E(X_{i0} \mid T_{i1}=1) - E(X_{i0} \mid T_{i1}=0)]}$

$= b$ YAY!

Differences-in-Differences, Graphically

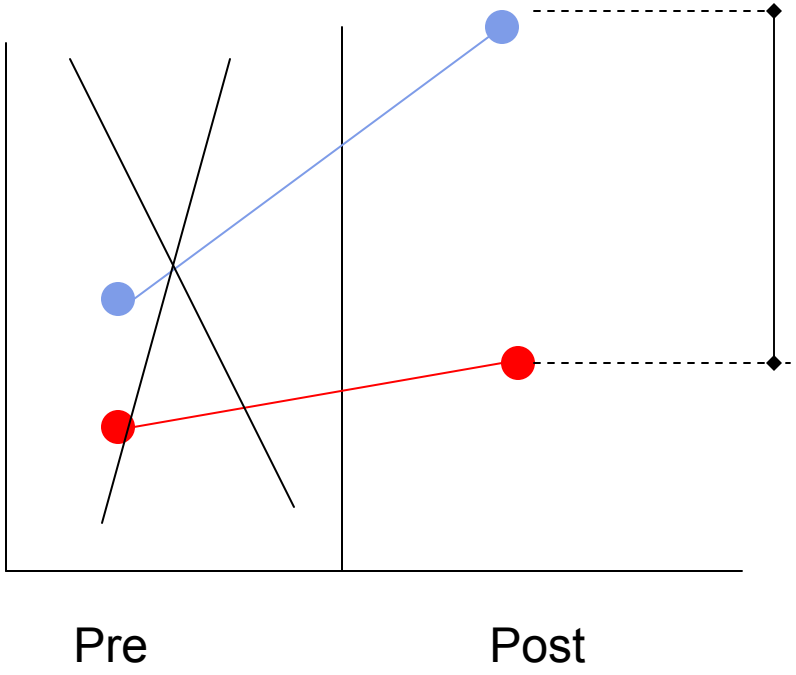


Differences-in-Differences, Graphically



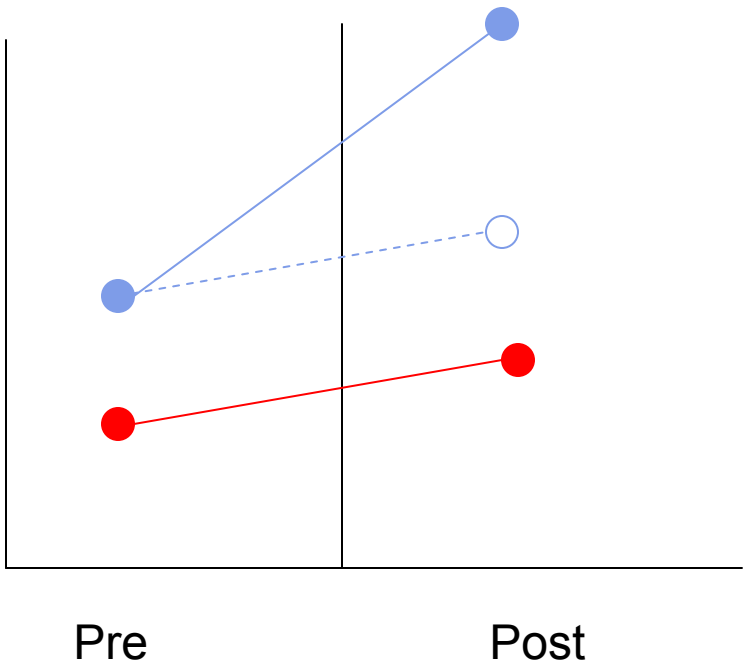
Effect of program using only pre- & post- data from T group (ignoring general time trend).

Differences-in-Differences, Graphically

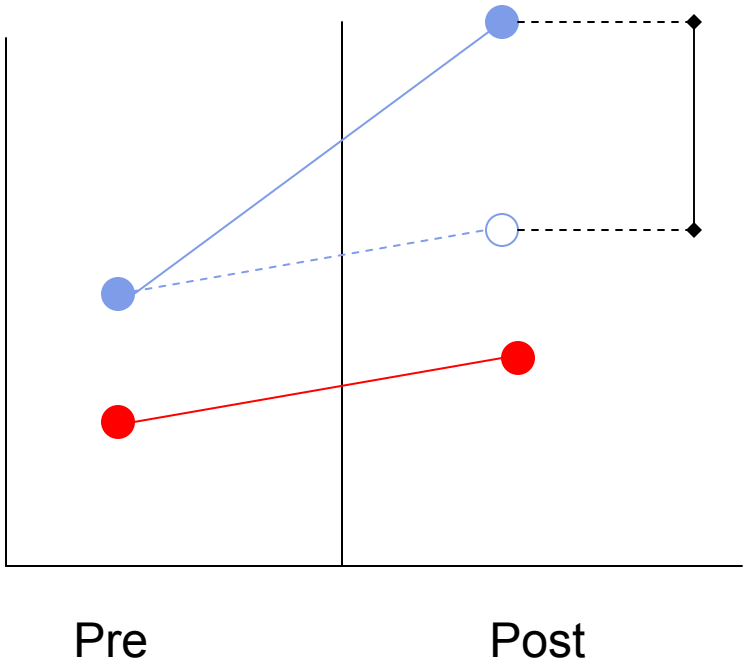


Effect of program using only T & C comparison from post-intervention (ignoring pre-existing differences between T & C groups).

Differences-in-Differences, Graphically



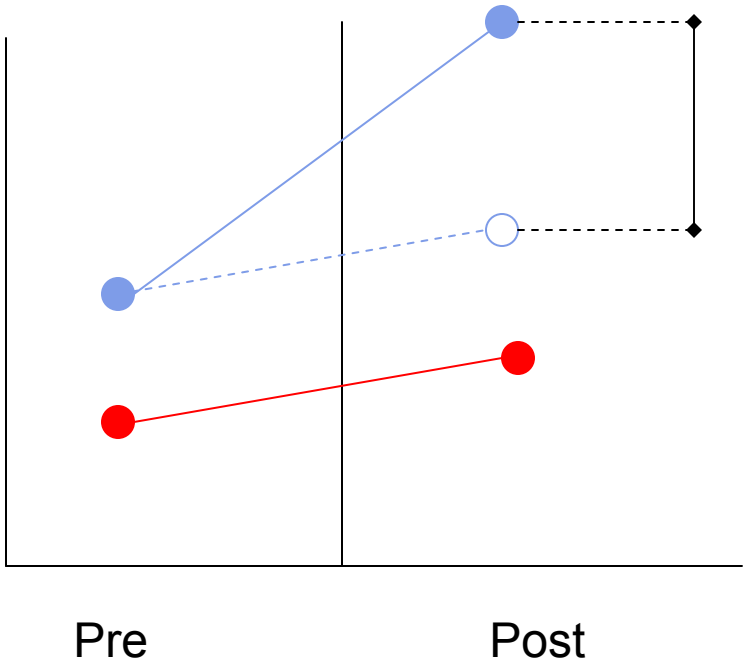
Differences-in-Differences, Graphically



Effect of program difference-in-difference (taking into account pre-existing differences between T & C and general time trend).

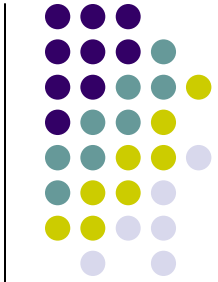
Identifying Assumption

- Whatever happened to the control group over time is what would have happened to the treatment group in the absence of the program.



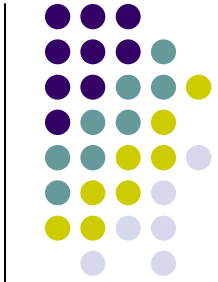
Effect of program difference-in-difference (taking into account pre-existing differences between T & C and general time trend).

Graphing Exercise



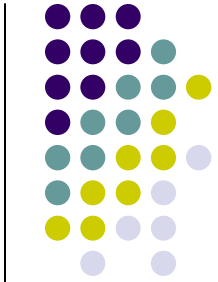
- Which of these programs had no effect?
- Which of these programs look like they were randomly assigned?
- Which of these programs look like they were placed in the areas that needed them most?
- Which of these programs make you wonder if there was some mean reversion going on?

Uses of Diff-in-Diff



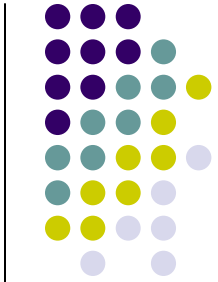
- Simple two-period, two-group comparison
  - very useful in combination with other methods

Uses of Diff-in-Diff



- Simple two-period, two-group comparison
  - very useful in combination with other methods
    - Randomization
    - Regression Discontinuity
    - Matching (propensity score)

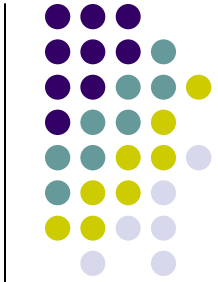
Uses of Diff-in-Diff



- Simple two-period, two-group comparison
  - very useful in combination with other methods
    - Randomization
    - Regression Discontinuity
    - Matching (propensity score)

- Can also do much more complicated “cohort” analysis, comparing many groups over many time periods

The (Simple) Regression



$$Y_{i,t} = a + bTreat_{i,t} + cPost_{i,t} + d(Treat_{i,t}Post_{i,t}) + e_{i,t}$$

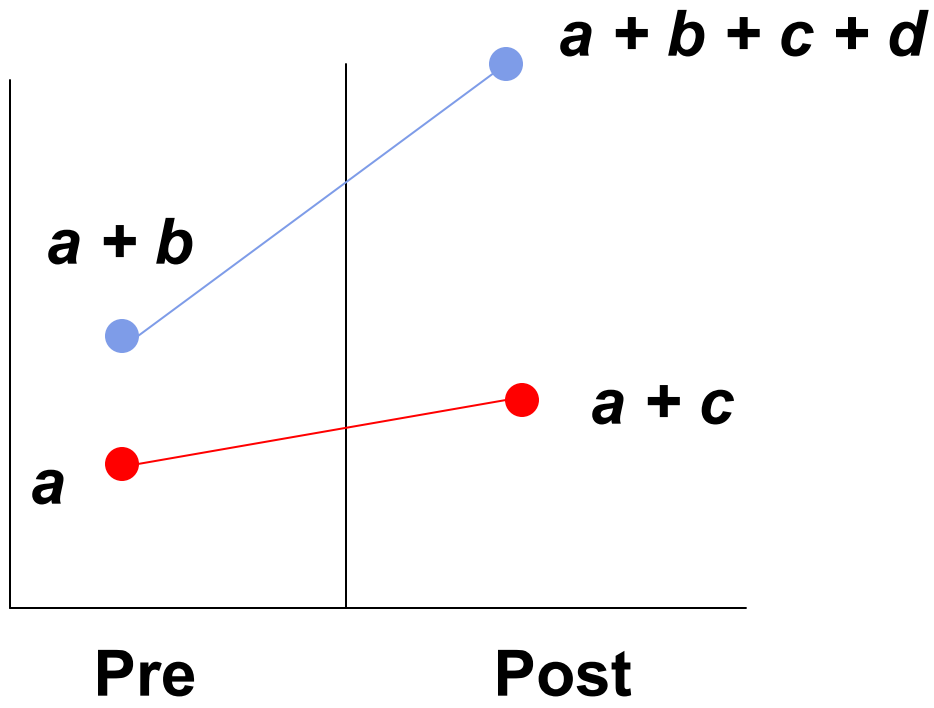
- $Treat_{i,t}$ is a binary indicator ("turns on" from 0 to 1) for being in the treatment group
- $Post_{i,t}$ is a binary indicator for the period after treatment
- and $Treat_{i,t}Post_{i,t}$ is the interaction (product)

Interpretation of *a, b, c, d* is "holding all else constant"

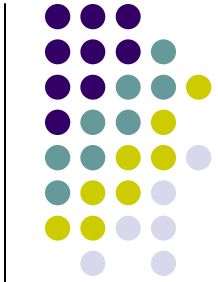
Putting Graph & Regression Together

$$Y_{i,t} = a + bTreat_{i,t} + cPost_{i,t} + d(Treat_{i,t}Post_{i,t}) + e_{i,t}$$

d **is the causal effect of treatment**

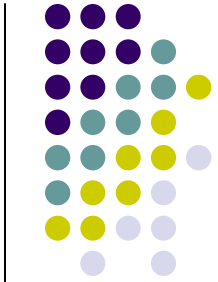


Cohort Analysis



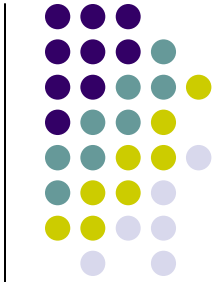
- When you've got richer data, it's not as easy to draw the picture or write the equations
  - cross-section (lots of individuals at one point in time)
  - time-series (one individual over lots of time)
  - repeated cross-section (lots of individuals over several times)
  - ★ panel (lots of individuals, multiple times for each) ★
- Basically, control for each time period and each "group" (fixed effects) – the coefficient on the treatment dummy is the effect you're trying to estimate

DiD Data Requirements



- Either repeated cross-section or panel
- Treatment can't happen for everyone at the same time
- If you believe the identifying assumption, then you can analyze policies ex post
  - Let's us tackle really big questions that we're unlikely to be able to randomize

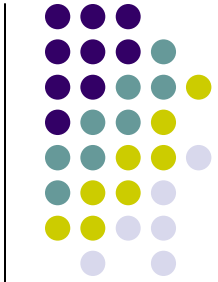
Malaria Eradication in the Americas (Bleakley 2007)



Question: What is the effect of malaria on economic development?

5 types of correlations (remember?):

Malaria Eradication in the Americas (Bleakley 2007)

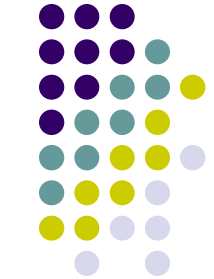


Question: What is the effect of malaria on economic development?

5 types of correlations (remember?):

- Causation
- Reverse causation
- Simultaneity
- Omitted variables
- Spurious correlation

Assumption OK?



- Eradication campaigns not determined by affected regions
- Campaigns made major progress over a short time span (10 years)
- Cross-regional variation in how bad malaria was (ecological differences)

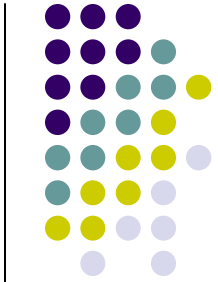
Malaria Eradication in the Americas (Bleakley 2007)

- Treated vs Control – those who were (were not) children in malaria endemic regions

- Pre vs Post – DDT spraying

"In both absolute terms and relative to the comparison group of non-malarious areas, cohorts born after eradication had higher income and literacy as adults than the preceding generations."

Robustness Checks



- If possible, use data on multiple pre-program periods to show that difference between treated & control is stable
  - Not necessary for trends to be parallel, just to know function for each

- If possible, use data on multiple post-program periods to show that unusual difference between treated & control occurs only concurrent with program

- Alternatively, use data on multiple indicators to show that response to program is only manifest for those we expect it to be (e.g. the diff-in-diff estimate of the impact of ITN distribution on diarrhea should be zero)