Econometrics Review: PHS 2000B

Maggie McConnell

Spring 2023

Department of Global Health and Population
Harvard T. H. Chan School of Public Health

Plan of presentation

1. Recap objectives of econometrics
2. Discussion of internal and external validity
3. Using a method
4. Review of Instrumental variables
5. Review of Interrupted Time Series

What is the point of econometrics?

- The primary objective of econometrics is to identify *causal effects*

- What differentiates econometrics from other causal inference approaches is extremely limited tolerance for confounding
  - Very unwilling to imagine that we could model sources of confounding
  - Human behavior is unpredictable and interesting and full of agency – people can choose how to react to policies/program and anticipate policy changes
  - Many of the things that economists care about are hard to measure/not typically available in datasets
    - Your beliefs, preferences, motivations, past experiences, etc
  - Economists love to imagine alternative theories – plausible or not!
    - Very hard to imagine you could model/"control for" all possible theories of behavior

Econometrics: focus on internal validity

- Ultimately economists require methods that ensure internal validity
  - Sometimes external validity is less of a focus
- Internal validity
  - The estimates are close to the true parameter values in the sample for which we have data
  - If methods are not internally valid, economists not interested in results
- External validity
  - The estimates are close to the parameter values in a population when they are used to predict outcomes
  - Generalizability and transportability
- One thing to practice as you study different methods is knowing how to differentiate between concerns related to internal vs external validity

Internal validity

- Internal validity usually requires some assumptions
- However, only some of the assumptions are testable
  - Sometimes when we can't directly test we can provide transparent visualizations that can increase our confidence in assumptions
- Where a choice of methods is available, economists always prefer the method with fewer untestable assumptions
- New methods sometimes have "hidden" assumptions that are stated in highly technical language --may take time and examples to uncover plausibility
  - Example – the discovery that two-way fixed effects models in difference in difference estimates using already treated units as controls

External validity

- When we use methods we have discussed we get estimates of “treatment effects” for specific sample at a particular time under certain conditions
- How much can we generalize these findings? what about similar interventions/policies? differences in implementation?
- While trials provide excellent *internal validity* – they may have external validity challenges
  - Special inclusion criteria: i.e. restricting trials to populations that are not likely to move away, those above a certain age, etc
  - Trial conditions that minimize attrition may limit external validity
    - i.e. many trials provide substantial financial payments for ongoing participation
- Quasi-experimental techniques we have learned may be
  - more externally valid or
    - For example if they take place in more naturalistic or policy-relevant settings, use naturally occurring administrative records that don’t require follow-up data collection and include data from a wide variety of states and settings
  - less externally valid
    - For example when they take place at a specific time/setting or the estimated treatment effect is only valid for a narrow group (i.e. those affected by an instrumental variable or near a threshold)

Improving Internal and External Validity

- Empirical work usually does not settle a research question – new papers continue to be published
- Contribution is to advance knowledge – can we improve internal and external validity?
- Some questions are hard to answer – can we improve on existing estimates even if we do not have perfection?

Learning to use a method

- Assumptions: we should know and state the assumptions of the method that are required to establish internal validity of estimates
  - Testable assumptions should be tested, proceeding only if these tests are passed
  - Untestable assumptions should be stated and their plausibility discussed
- Additional assumptions are required for valid confidence intervals, and these should also be stated and tested
- Analysis should reflect recent developments – awareness of advances in tests of assumption validity, estimation techniques, options for estimation, post estimation presentation of results
- Increasingly, success in publishing relies on making compelling and transparent figures to go along with your estimates

Returning to our original example

- We started with a motivating example using some evidence from a randomized controlled trial of intensive nurse home visiting during pregnancy/early childhood
- A trial achieves internal validity by comparing those assigned to treatment to those assigned to control
- Let's imagine that we wanted to consider the impact of the program on infant mortality
- Trials are often considered "gold standard" of evidence but there are challenges
  - *Internal validity*
    - attrition that may be different across treatment / control, challenges with power for rare outcomes, possible concerns about spillovers
  - *External validity*
    - does selection process for trial choose those who would typically participate, is the program practiced as it would be in more naturalistic settings

Other causal inference strategies

- What if we couldn't randomize?
- We have covered several *other* econometric strategies you can use
- Suppose we knew that there was a lot of variation in how much OB-GYNs serving Medicaid patients referred to nurse home visiting services
  - Instrumental variables approach
- Suppose there was policy that allowed families with an infant born weighing <1600 grams to receive nurse home based services?
  - Regression discontinuity approach
- Suppose we had only a time series of infant mortality and we knew that starting in a specific month all Medicaid eligible families were provided with home visiting services
  - Interrupted time series
- Suppose had data on infant mortality in all states and we know that universal access to home visiting services delivered by Medicaid rolled out at different times in different states
  - Difference in differences approach

Instrumental Variables

When does this work?

- There are FOUR key requirements for IV inference

1. **Relevance:** The instrument must be highly correlated with the variable of interest
2. **Excludability:** The instrument must NOT have a direct causal effect on the outcome Y
3. **Independence:** The instrument cannot be correlated with any other unobserved determinant of the outcome Y
4. **Homogeneity:** The treatment effect is homogenous across all study subjects
   - If we think this isn't plausible, we can still estimate LOCAL average treatment effects relevant for those affected by the instrument if we assume monotonicity

Condition 1: Relevance

- The relevance assumption is relatively easy to both satisfy and verify
- In most contexts, other factors are correlated with the variable of interest
- The “strength” of the instrument is determined by its predictive power in a multivariate setting
- A good instrument explains a significance part of the variation in the endogenous variable *once we control for all other factors in our empirical model*

Conditions 2/3: Excludability/Independence

- The excludability/independence conditions are more complex since orthogonality to (i.e., being uncorrelated with) the error requires that:
  - The instrument does not have a direct effect on the dependent variable
  - The instrument is not correlated with other omitted variables that have a direct effect
  - The instrument is not correlated with measurement error
  - The instrument is not "caused" by the dependent variable
  - This assumption is not testable
    - We can only discuss its conceptual plausibility

IV estimation in practice using Two Stage Least Squares

- In multivariate models, we may have several endogenous, exogenous, and instrumental variables

- IV estimation is done in two steps:
  - **Step 1:** regress the endogenous variables on the instrument and predict the "exogenous" part of the variation (first stage)
  - **Step 2:** regress the dependent variable (outcome of interest) on the predicted values from the first stage to get the (ideally) unbiased IV estimator (second stage)
- Since the two-stage approach is fairly standard, IV estimation is also referred to as Two Stage Least Squares estimation (2SLS, or TSLS)
- We can also look at what economists often call the "reduced form"
  - The impact of the *instrument* on the outcome of interest

Application – home visiting

- *Research question*: What is impact of home visiting on infant mortality?
- *Setting*: Suppose we knew that there was a lot of variation in how much OB-GYNs serving Medicaid patients referred to nurse home visiting services

- Why not just compare those who receive home visiting services with those who do not? i.e. why do we need an IV strategy?
- What could you use as an instrument to study ?
  - which OB you had during your pregnancy
- How plausible are the requirements for internal validity in this case?
  - What kind of evidence would you want to see to be more confident about internal validity?

- How would we estimate?
  - *First stage*: impact of OB assignment on probability of participating in home visiting
  - *Second stage*: regress infant mortality on fitted values from first stage
  - *Reduced form*: impact of OB assignment on infant mortality

- Assuming there is heterogeneity, who are the treatment effects relevant for?
  - Those who participate in home visiting because their OB refers them

Regression Discontinuity

Under what conditions can we estimate RDD?

Conditions

- A continuous eligibility index: a continuous measure on which the population of interest is ranked (i.e. test score, health indicator, age, vote share)
  - Often called “running variable”

- A clearly defined cutoff point: a point on the index above or below which the population is determined to be eligible for the program
  - Sometimes referred to as the “threshold”

What assumptions required for RDD to be valid?

Key assumption

The only thing that differs systematically for groups above and below the threshold is the likelihood of program participation/eligibility

- We'd expect continuity of outcome at threshold if untreated by the program – no natural discontinuity in the outcome at the threshold

Implications of this assumption

- Individuals can't be able to manipulate the running variable in order to increase the chances of being included / excluded
- Individuals close to the cutoff point should be very similar, on average, in observed and unobserved characteristics

RDD, RCTs, and IV estimation

- Under ideal conditions, RDDs are very similar conceptually to RCTs: if it is true that treatment assignment is close to “random” within the analyzed range, then RDDs become a quasi-experiment

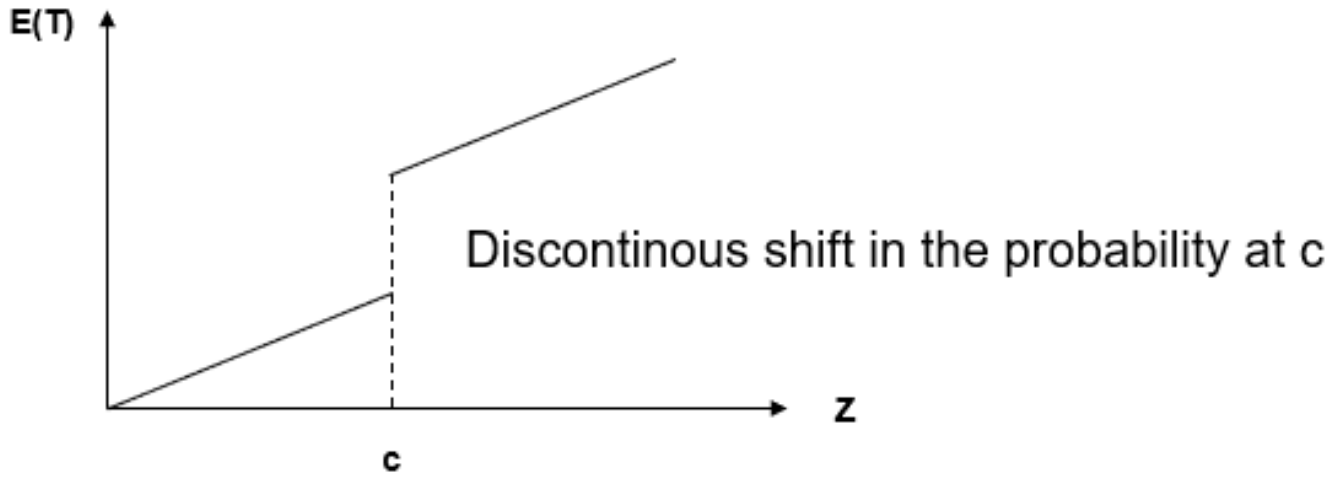
- RDDs can also be seen as a special case of IV estimation
  - Under the RDD assumption, we know that there is an increase in the treatment at the cutoff, and we explore this exogenous variation in the treatment to identify the causal effect of interest

- **Key question:** How plausible is it that the cutoff affects the outcome only through treatment?
  - Recall, this is the similar to the exclusion assumption in an IV setting

The critical threshold

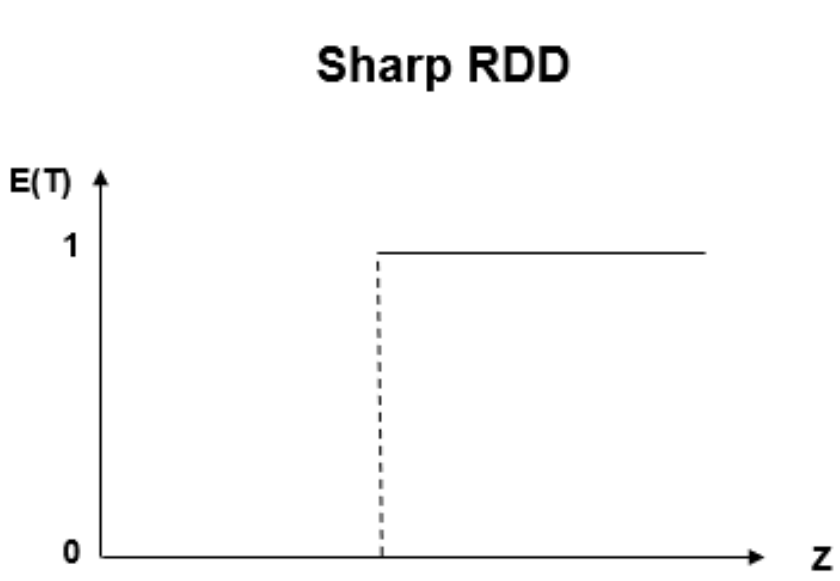
- Assume that there is a continuous variable z which determines eligibility for treatment discontinuously at some cutoff (c)

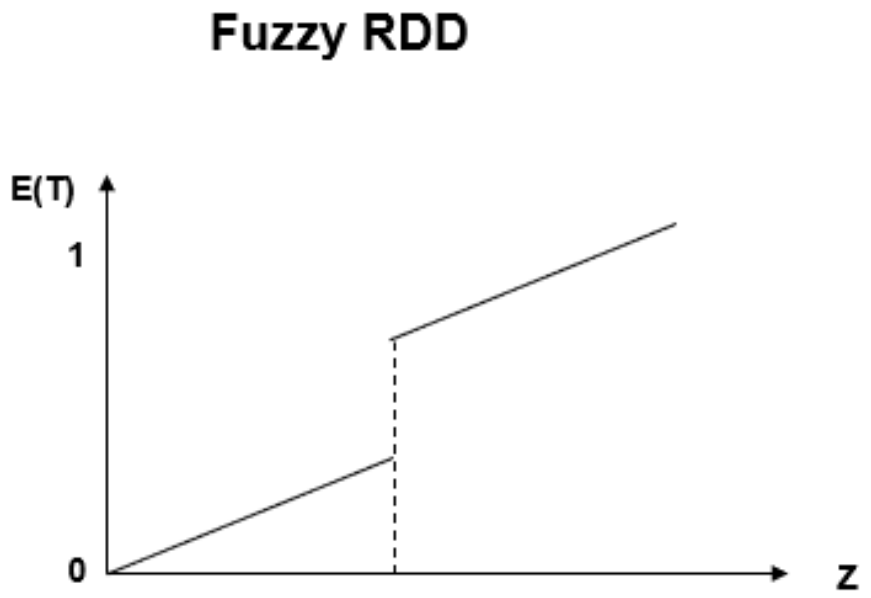
$$\lim_{Z \downarrow c} \Pr(T_i = 1) \neq \lim_{Z \uparrow C} \Pr(T_i = 1)$$



Sharp discontinuities vs. fuzzy discontinuities

- With sharp discontinuities, the probability of receiving treatment at the cutoff goes from 0 to 1
- With fuzzy discontinuities, the change in probability of receiving treatment at the cutoff is less than 1





Assessment of Validity of RDD

- We want the distribution of the running variable to be smooth around the threshold
  - No bunching (i.e. no signs of manipulation)

- We want to see a jump in the probability of receiving treatment at the threshold
  - For sharp RD this jump goes from 0 to 1
  - For fuzzy RDD this jump is somewhere in the range $0<p<1$

- We want the characteristics of those just above and below the threshold to be similar

Practical Tips: Assessing the Validity of RDD

1) Show characteristics (SES, age...) are balanced around cutoff.
   - One way to do this is to estimate the same main regression as RD but your "outcomes" are the covariates. Look for jumps near the cutoff.
   - Similar to balance tables in RCTs
2) Look for "bunching" in the running variable around cutoff
   - Should not be extra mass in distribution near cutoff
   - Can use McCrary 2008 density test to test this formally
3) Estimate causal effect using different bandwidths and check stability of estimates — instability suggests wrong functional form
   - can use formal procedures to choose "optimal" bandwidth (Imbens and Kalyanaraman 2009)
4) Estimate "false threshold"
   - Should not find significant effects at the wrong threshold

Application – home visiting

- *Research question*: What is impact of home visiting on infant mortality?
- *Setting*: Suppose there was policy that allowed families with an infant born weighing <1600 grams to receive nurse home based services
- What is the running variable?
  - Birth weight
- What is the threshold?
  - 1600 grams
- How plausible are the requirements for internal validity in this case?
  - What kind of evidence would you want to see to be more confident about internal validity?
- Who would the treatment effects be relevant for?
  - Those who participate in home visiting because they are near the threshold

Interrupted Time Series

Logic of interrupted time series analysis

- Time series data allows us to carefully model the general trends in our outcome of interest

- Most programs/treatments are introduced at a very specific time point: if the program worked, we should observe that the outcome shifts onto a different trajectory after the policy change happened

- The main assumption of this interrupted time series (ITS) approach is that the pre-intervention trend can be continued after the intervention as the counterfactual

ITS counterfactual

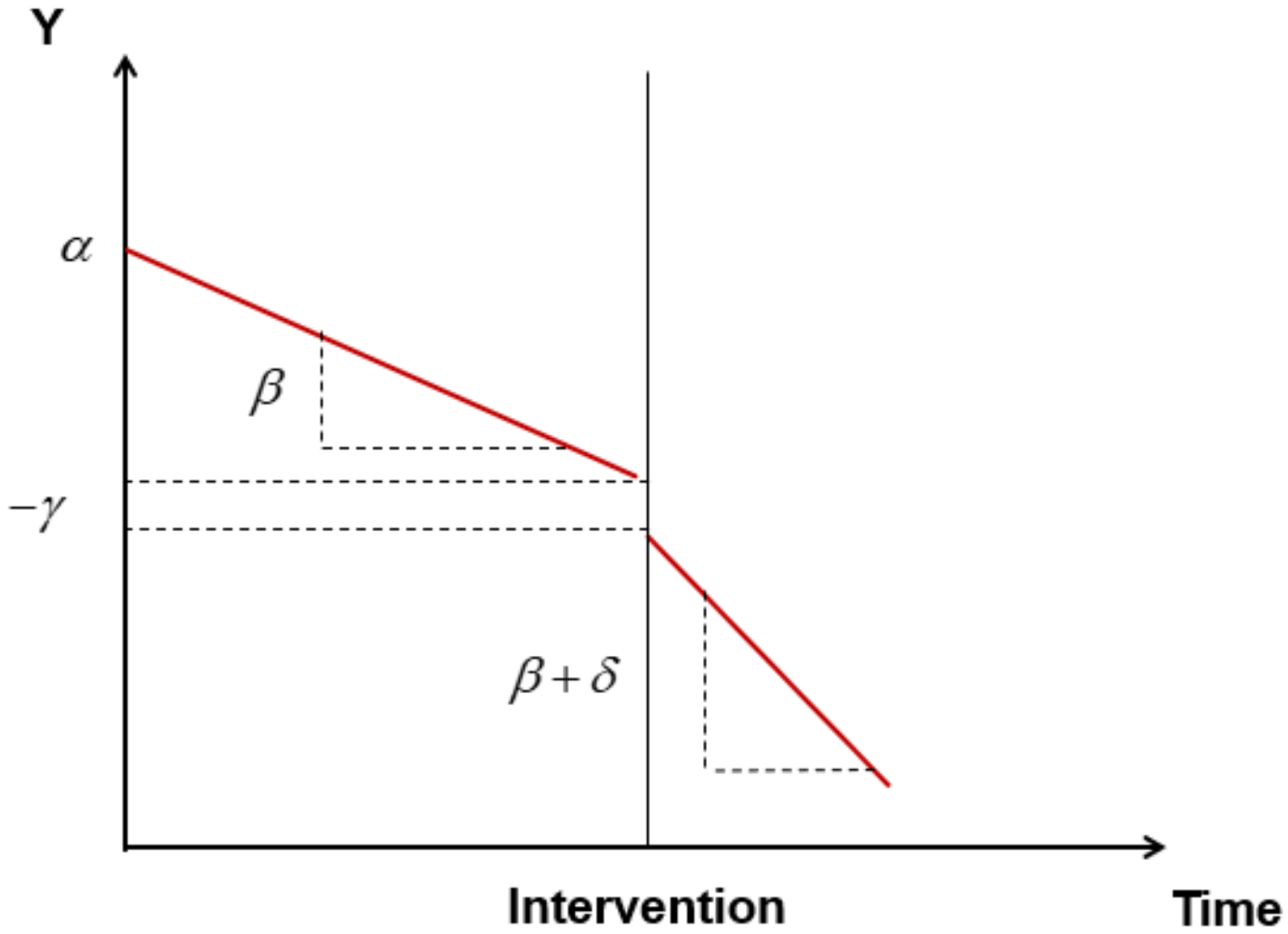
- In a randomized trial the counterfactual is what happens in the control group – we think this is what would have happened without the treatment
- In ITS we fit a trend to the data in the pre intervention period – the counterfactual is that the data would have follows this trend in the post intervention period without the trend
- Requires that we have adequately fit the trend in the data
- Requires that no other new interventions happening in the post intervention period that could affect the outcome
  - This is a big assumption – not likely to be true when we know the policy landscape changing at the same time

A simple model for intercept and slope

$$Y_t = \alpha + \beta time + \gamma Post + \delta Post * time + \epsilon_t$$

- Y_t: Outcome of interest at time t
- $Post$: an indicator for the "post-reform" period
- α: level of the outcome Y at time $t = 0$
- β: linear time trend
- γ: level shift post-reform
- δ: gradient shift post-reform

Interpretation of coefficients



Challenge: Getting the timing right

- In some instances, the impact of the policy change should be immediate conditional on the policy being effectively implemented
  - For example: handwashing on diarrhea, or speeding laws on traffic accidents

- In many other cases, it is not clear how soon we should expect an impact on outcomes
  - Particularly challenging for outcomes like cancer or children's height which are the result of longer-term processes

- There may possibly be anticipation effects as well
  - For example, if there's an announced change in billing policies around the insertion of long-acting contraceptive methods, hospital practices may start to alter before the billing change goes into effect

- We need to have a clear conceptual framework for when we would expect impact relative to the timing of the reform

Specifying the impact model

- Before running actual regressions, we need to be clear regarding our hypotheses on how the policy change affected the outcomes of interest

- **Intercept shift:** are we expecting an immediate shift in the outcome?

- **Slope shift:** are we expecting outcome trajectories to change?

- **Timing and duration change:** are we expecting changes to start immediately or to start with a lag? To last, or to fade out?

Application – home visiting

- *Research question*: What is impact of home visiting on infant mortality?

- *Setting*: Suppose we had only a time series of infant mortality and we knew that starting in January 2016 all Medicaid eligible families were provided with home visiting services
  - Imagine that before the policy-change very few families received services and afterwards participation was >80%

- When would we expect to see the treatment effects come in?

- How plausible are the requirements for internal validity in this case?
  - What kind of evidence would you want to see to be more confident about internal validity?

Difference in Differences

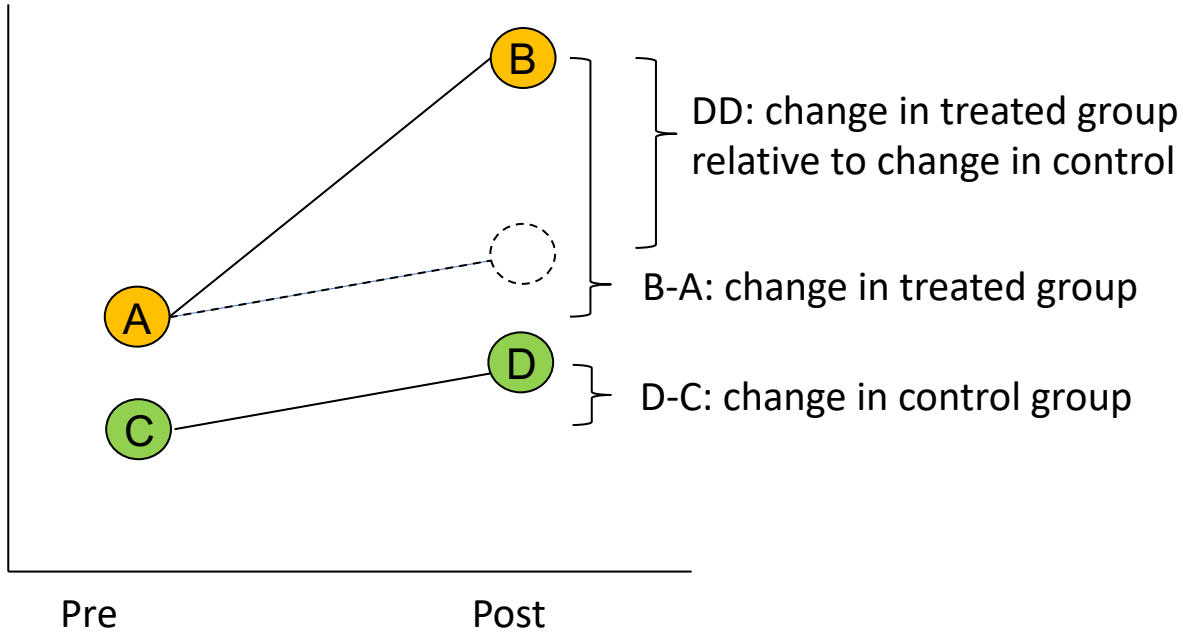
Difference-in-Differences logic

- The real causal question is: How would the “treated” have done without the treatment?
- Differences in differences assumes the control group can tell us *what would have* happened in the post intervention period in the treatment group with no intervention

- The counterfactual is that if the treatment group had not got the treatment, the *change* we would have seen is the *change* observed in the control group over the same period

Difference in difference

	Treatment
Before	A
After	B
Change	B-A



Did the outcomes in the treated areas improve by more than in the control areas?

DD = (change in treatment) – (change in control) = ($Treat^{post}$ - $Treat^{pre}$) – ($Control^{post}$ - $Control^{pre}$)

Key assumptions: Counterfactual

- Key assumption in Difference-in-Differences
  - Treatment group would have followed the same trajectory as the control group in the post-intervention period if the intervention had not occurred

- This counterfactual is **not** testable
  - This counterfactual may feel more believable if
    - Control and treatment areas are qualitatively similar before the intervention
    - The composition of control and treatment groups are not changing over time in different ways
    - The control and treatment group are on a similar trajectory before the intervention

Multiple groups and time periods

- We need at least two groups – treatment and control – and two time periods – pre and post – to get a difference-in-differences estimator
  - Essentially at least four observations
  - However -- need more data for inference
  - With more data we can strengthen our belief in the assumptions that make DiD valid
    - Hard for one district to make a convincing control group → averaging across many districts together means the counterfactual assumption is more plausible
    - With just one observation from the pre-period, harder to know that we haven't gotten noise → more pre-period data allows better modeling of trends and changes over time
- DID is a flexible methodology we can extend in many ways
  - For example, we can include two post periods – an initial period of "transition" and a period where policy is fully implemented

Multiple groups and time periods

- Typical DiD specification
  - We have multiple groups ($g = \{1 \dots M\}$) with some getting the treatment and some not getting it (group fixed effects)
    - ”Treated” binary variable gets subsumed by these group fixed effects
  - We have multiple time periods ($t = \{1 \dots K\}$) both pre- and post-intervention (time fixed effects)
    - “Post” binary variable gets subsumed by these time fixed effects

$$Y_{igt} = \sum_{g=1}^{M} \alpha_g + \sum_{t=1}^{K} \gamma_t + \beta\left(I_{post} * I_{g \in T}\right) + \epsilon_{igt}$$

  - β tells us the average effect of treatment on the treated
    - The difference in the outcome in the post-intervention period in the treated areas after accounting for state and time level differences
  - Models can be more complex, i.e. modeling time trends, covariates, phased in implementation

Challenge: Seasonality of outcomes

- Most time series data has daily, monthly, or quarterly records with rather pronounced seasonality in the outcome

- If seasonal effects are not controlled for, impact estimates may be biased because programs are likely to start in either high or low incidence months (with biases in the opposite direction)

- In regression models, we can:
  - Model the seasonality directly by using “season dummies”
  - Create a “de-trended” time series by first regressing outcome on trend model and season effects, and then just analyzing deviations from trend

Valid standard errors: Autocorrelation

- With autocorrelation, observations for the same group will be correlated over time
- Observations of multiple time points pre- and post-intervention will not be independent and do not add much additional information
- With a large number of groups, clustering the standard errors across time within the same group can correct for autocorrelation
- We can also estimate the autocorrelation structure, but very biased estimates with group fixed effects and small number of periods

Autocorrelation and standard errors

- Even when we de-trend the data, it seems somewhat likely that two adjacent observations are non-independent

- Autocorrelated errors still give unbiased coefficient estimates, but will likely result in incorrect standard errors
  - Autocorrelation means less information in each observation than if the data are independent

- While autocorrelation seems less common in public health time series data, it is always good to test for it and easy to do so

Application – home visiting

- *Research question*: What is impact of home visiting on infant mortality?

- *Setting*: Suppose had data on infant mortality in all states and we know that universal access to home visiting services delivered by Medicaid rolled out at different times in different states

- How plausible are the requirements for internal validity in this case?
  - What kind of evidence would you want to see to be more confident about internal validity?

Final thoughts

- Econometrics provides great tools for thinking about causal questions

- Using them often means narrowing your window of inquiry
  - Not – what is going on with the relationship between housing and health
  - But – what is the impact of specific housing policy x on health outcome Y
    - Pros: we get very specific about our assumptions, we can feel somewhat confident about the internal validity of our results
    - Cons: we aren't always talking about the things that matter most to equity/outcomes