Reading the "Diff-in-diff Practitioner's Guide" together

Decrypting the theory and seeing its application in Stata

Igor Francetic

Scuola universitaria professionale
della Svizzera italiana
SUPSI

The University of Manchester

HOPE training workshop, 3rd October 2025

This workshop is based on two papers

The content of this workshop draws abundantly from the following papers:

- "Difference-in-Differences Designs: A Practitioner's Guide", henceforth "The Guide" by Baker et al (2025). https://arxiv.org/pdf/2503.13323
- "Difference-in-differences with variation in treatment timing", by Goodman-Bacon (2021). https://doi.org/10.1016/j.jeconom.2021.03.014
- "Difference-in-Differences with multiple time periods", by Callaway and Sant'Anna (2021). https://doi.org/10.1016/j.jeconom.2020.12.001

Difference-in-Differences Designs: A Practitioner's Guide

Andrew Baker

Brantly Callaway

Scott Cunningham

Andrew Goodman-Bacon

Pedro H. C. Sant'Anna

JOURNAL OF ECONOMIC LITERATURE (FORTHCOMING)

Abstract

Difference-in-differences (DiD) is arguably the most popular quasi-experimental research design. Its canonical form, with two groups and two periods, is well-understood. However, empirical practices can be ad hoc when researchers go beyond that simple case. This article provides an organizing framework for discussing different types of DiD designs and their associated DiD estimators. It discusses covariates, weights, handling multiple periods, and staggered treatments. The organizational framework, however, applies to other extensions of DiD methods as well.

Journal of Econometrics 225 (2021) 254–277

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Difference-in-differences with variation in treatment timing☆

Andrew Goodman-Bacon*

Opportunity and Inclusive Growth Institute, Federal Reserve Bank of Minneapolis, 90 Hennepin Ave, Minneapolis, MN 55401, USA
National Bureau of Economic Research, USA

ARTICLE INFO

Article history:
Received 19 January 2021
Received in revised form 19 January 2021
Accepted 17 March 2021
Available online 12 June 2021

Keywords:
Difference-in-differences
Variation in treatment timing
Two-way fixed effects
Treatment effect heterogeneity

ABSTRACT

The canonical difference-in-differences (DD) estimator contains two time periods, "pre" and "post", and two groups, "treatment" and "control". Most DD applications, however, exploit variation across groups of units that receive treatment at different times. This paper shows that the two-way fixed effects estimator equals a weighted average of all possible two-group/two-period DD estimators in the data. A causal interpretation of two-way fixed effects DD estimates requires both a parallel trends assumption and treatment effects that are constant over time. I show how to decompose the difference between two specifications, and provide a new analysis of models that include time-varying controls.

Published by Elsevier B.V.

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Difference-in-Differences with multiple time periods☆

Brantly Callaway [a], Pedro H.C. Sant'Anna [b,*]

[a] Department of Economics, University of Georgia, United States of America
[b] Department of Economics, Vanderbilt University, United States of America

ARTICLE INFO

Article history:
Received 1 March 2019
Received in revised form 18 August 2020
Accepted 1 December 2020
Available online 17 December 2020

JEL classification:
C14
C21
C23
J23
J38

Keywords:
Difference-in-Differences
Dynamic treatment effects
Doubly robust
Event study
Variation in treatment timing
Treatment effect heterogeneity
Semi-parametric

ABSTRACT

In this article, we consider identification, estimation, and inference procedures for treatment effect parameters using Difference-in-Differences (DiD) with (i) multiple time periods, (ii) variation in treatment timing, and (iii) when the "parallel trends assumption" holds potentially only after conditioning on observed covariates. We show that a family of causal effect parameters are identified in staggered DiD setups, even if differences in observed characteristics create non-parallel outcome dynamics between groups. Our identification results allow one to use outcome regression, inverse probability weighting, or doubly-robust estimands. We also propose different aggregation schemes that can be used to highlight treatment effect heterogeneity across different dimensions as well as to summarize the overall effect of participating in the treatment. We establish the asymptotic properties of the proposed estimators and prove the validity of a computationally convenient bootstrap procedure to conduct asymptotically valid simultaneous (instead of pointwise) inference. Finally, we illustrate the relevance of our proposed tools by analyzing the effect of the minimum wage on teen employment from 2001–2007. Open-source software is available for implementing the proposed methods.

Objectives

The objectives of this workshop are to:

- Refresh the basics of Difference in Differences designs (i.e. setting, how the design recovers causal estimates, conditions for identification, estimation, inference)
- Understand the key messages of the approach to conduct Difference in Differences proposed by "The Guide"
- See how these concepts are applied in Stata (and collaterally see the HOPE GitHub)

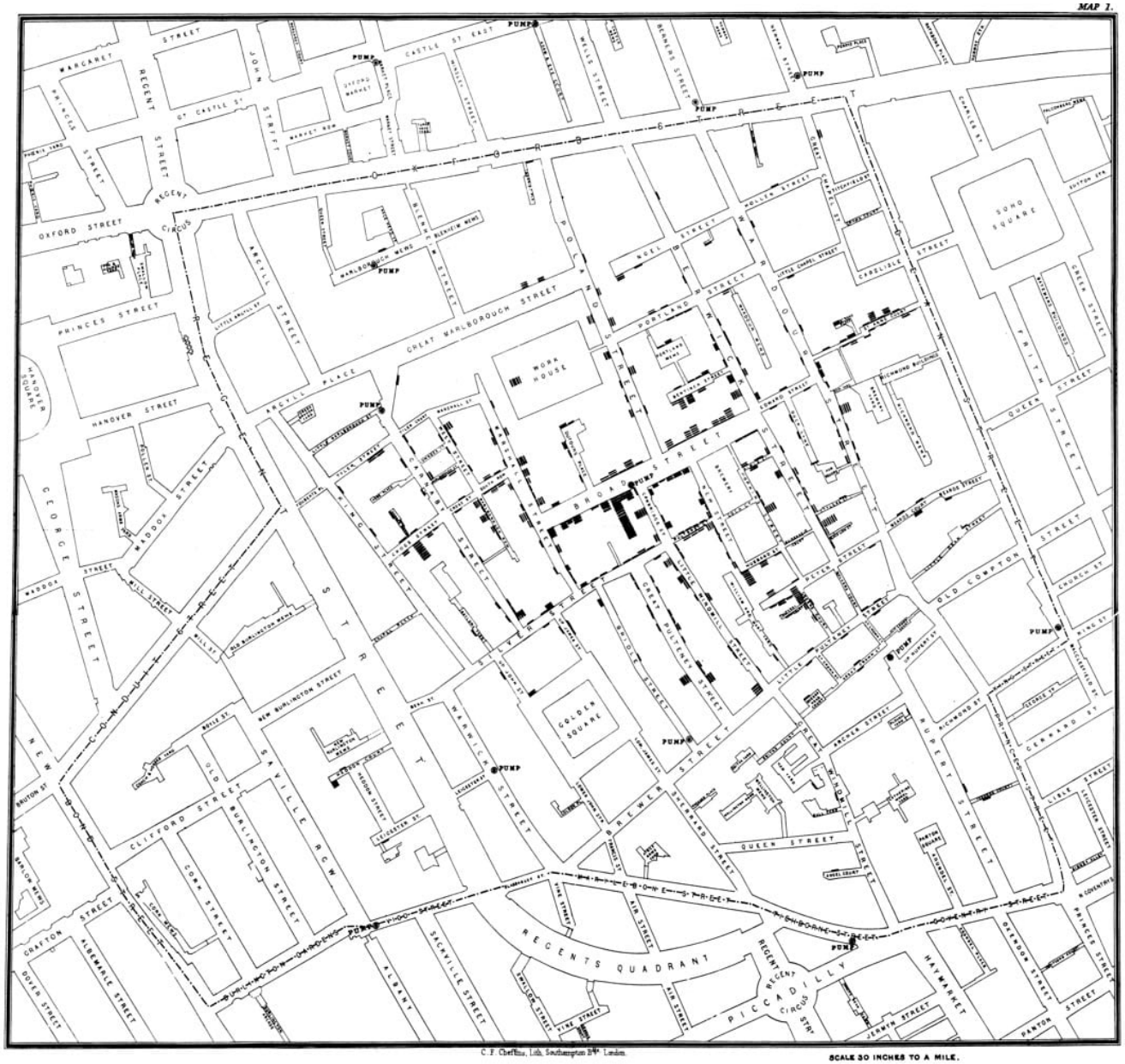
The workshop does **not**:

- Substitute a thorough read of "The Guide"
- Cover "The Guide" in every detail, for example we won't cover inference, survey/population weights, treatments switching on and off, and a few other details, heterogeneity across covariates
- There is interesting elements to cover in a second workshop

Let's settle on some terminology before we start

Shall we?

- Target parameter: the type of "treatment effect" we want to measure. E.g. RCTs target the average treatment effect (ATE)
- Identification: the assumptions required to ensure that the research design allows to measure the target parameter. E.g. RCTs rely on randomisation to ensure that treatment is independent of potential outcomes.
- Estimation: the statistical process of estimating an empirical equivalent of the target parameter. E.g. the sample average of Y, $\bar{Y}$ is an estimate of its expected value $\mathbb{E}[Y]$
- Inference: the process of estimating the uncertainty around the estimates to allow for hypothesis testing (typically $H_0 : \beta = 0$)


MAP 1.
BROAD STREET
GOLDEN SQUARE
REGENTS QUADRANT
PICCADILLY
SCALE 30 INCHES TO A MILE.

The canonical 2×2 DID

- Difference in Differences (DID) is an approach to policy evaluation that dates back at least to the 1840's, which was developed (methodologically) by economists
- In its most basic (canonical) form (2×2), DID requires two groups and two periods
  - Both groups are not exposed to the policy in the first period (pre)
  - One (treatment) group is exposed to the policy in the second (post) period
  - The DID estimates is the difference in average changes in outcomes (pre vs. post) between the two groups: the difference in differences, that is.
- **Only if** changes in outcomes would have been the same in the two groups - had treatment not occurred (i.e. the parallel trends assumption) - DID estimates the average treatment effect among treated units (ATT).

DID among practitioners

- We practitioners rarely focus on canonical 2×2 settings. The most common deviations include
  - Observing more than 2 periods
  - Units may enter (or exit) treatment at different times
  - Treatment may vary in its intensity
  - Control variables are included to make control and treatment group more comparable and increase precision
- Until 2019, we (practitioners) carelessly estimated complex DID designs using so-called two-way fixed effects models (TWFE) assuming
- TWFE = linear regression models with unit and time fixed effects
- In canonical 2×2 settings, TWFE gives exactly the same estimates as a difference in sample means
- We all relied on the previous point to estimate all sorts of complex designs

Then came the Bacon Decomposition...

Journal of Econometrics 225 (2021) 254–277

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Difference-in-differences with variation in treatment timing☆

Andrew Goodman-Bacon *

Opportunity and Inclusive Growth Institute, Federal Reserve Bank of Minneapolis, 90 Hennepin Ave, Minneapolis, MN 55401, USA
National Bureau of Economic Research, USA

ARTICLE INFO

Article history:
Received 19 January 2021
Received in revised form 19 January 2021
Accepted 17 March 2021
Available online 12 June 2021

Keywords:
Difference-in-differences
Variation in treatment timing
Two-way fixed effects
Treatment effect heterogeneity

ABSTRACT

The canonical difference-in-differences (DD) estimator contains two time periods, "pre" and "post", and two groups, "treatment" and "control". Most DD applications, however, exploit variation across groups of units that receive treatment at different times. This paper shows that the two-way fixed effects estimator equals a weighted average of all possible two-group/two-period DD estimators in the data. A causal interpretation of two-way fixed effects DD estimates requires both a parallel trends assumption and treatment effects that are constant over time. I show how to decompose the difference between two specifications, and provide a new analysis of models that include time-varying controls.

Published by Elsevier B.V.

https://doi.org/10.1016/j.jeconom.2021.03.014

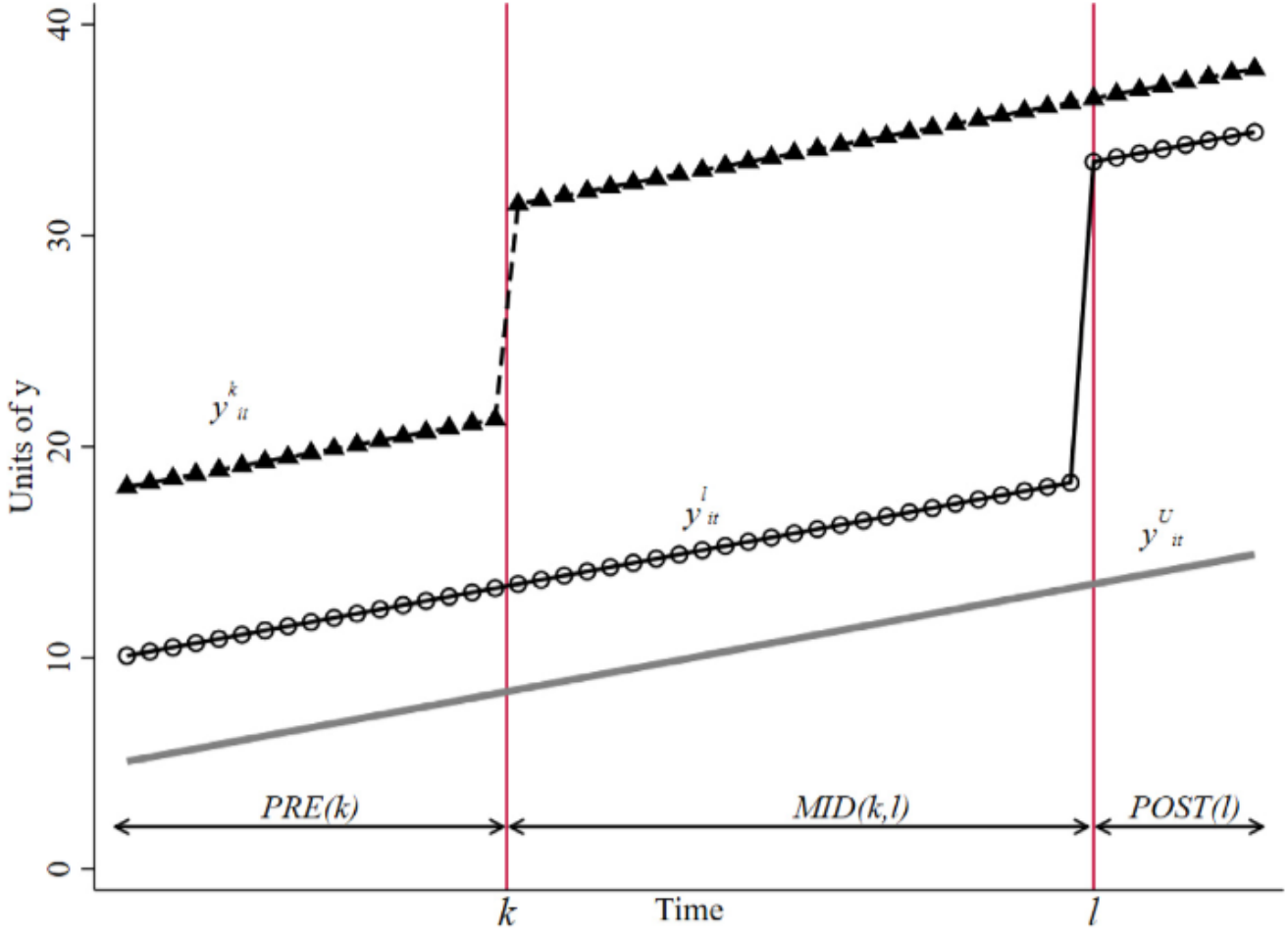
Implication of Goodman-Bacon (2021)

- Goodman-Bacon (2021) proposes a theorem to decompose the TWFE estimator
- The decomposition is not too hard to follow (it uses the Frisch-Waugh-Lovell theorem) but we'll not go over it here
- The core finding is that

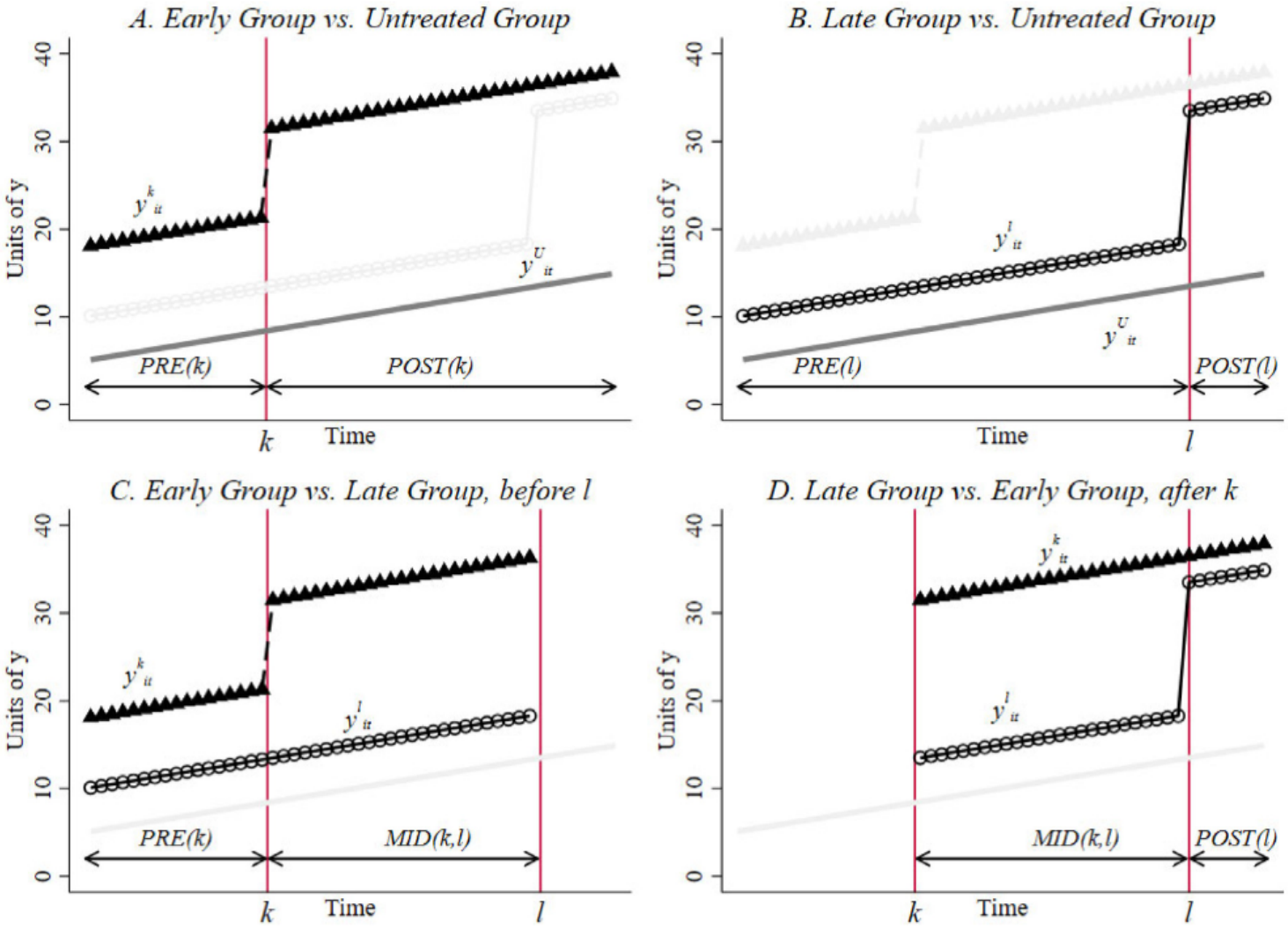
  > *"... (in non-canonical settings) the TWFE estimator equals a weighted average of all possible two-group/two-period DD estimators in the data. A causal interpretation of two-way fixed effects DD estimates requires both a parallel trends assumption and treatment effects that are constant over time"*

- Even with *parallel trends*, with staggered timing and treatment heterogeneity TWFE can be biased **to the point of reversing the effect sign**!
- One of the biggest problems comes from "forbidden comparisons" of late and early treated groups

A graphic decomposition



A graphic decomposition



So, what should applied researchers do instead?

The approach proposed by Baker et al (2025)

Viewing DiD studies through the lens of 2×2 "building blocks" aids in interpretability by clarifying that they yield causal quantities that aggregate the treatment effects identified by each 2×2 component. It also means that identification comes from the simple parallel trends assumptions required for each 2×2 building block. Practically, the building block framework suggests first estimating each 2×2 and then aggregating them. As long as the effective sample size is large, this approach allows for asymptotically valid inference using standard technique

Baker et al (2025)

Step 1: the 2×2 setting

- The Guide starts by considering the conditions and methods to identify a causal effect in the canonical 2×2 case
- That is, we define
  - a treatment dummy D_i equal to 1 for treated units, and 0 for those that do not
  - a time variable (e.g. a dummy for the Post period)

Step 1: the 2×2 case

Target

- Any causal analysis should starts with the definition of the target parameter, in our case we use DID to target the Average Treatment effect on the Treated units (ATT)
- In the language of the potential outcomes framework, at some post-treatment period (relative to Pre), the ATT can be written as

$$ATT(Post) = \mathbb{E}[Y_{i,Post}|D_i = 1] - \mathbb{E}[Y_{i,Post}(0)|D_i = 1]$$

  - LHS: the average outcome for treated units after the treatment
  - RHS: the average untreated outcome for the same units (this is NOT observed, it's the counterfactual, the "what if")

Step 1: the 2×2 case

Identifying assumptions: No anticipation

For DID to identify the ATT we need to make a few assumptions, known in econometrics as the **identifying assumptions**. Let's see what these are and what they achieve, starting with the No anticipation assumption:

- "treatment units do not respond to the treatment before they are exposed; treatment begins to have effects from the first treatment period"
- Formally $Y_{i,t}(1) = Y_{it}(0)$ $\forall$ units i and pre-treatment periods t
- Allows to impose that ATT is zero for all pre treated periods

Step 1: the 2×2 case

Identifying assumptions: Parallel trends

The most famous assumption in the context of DID is Parallel trends (PT). But what is it?

- "in the absence of treatment, the average outcome evolution is the same among treated and comparison groups".
- Formally we're saying that

$$\mathbb{E}[Y_{i,Post}(0)|D_i = 1] - \mathbb{E}[Y_{i,Pre}(0)|D_i = 1] =$$

$$\mathbb{E}[Y_{i,Post}(0)|D_i = 0] - \mathbb{E}[Y_{i,Pre}(0)|D_i = 0]$$

- If that holds, it's easy to construct the counterfactual that we need by projecting the OBSERVED change for the untreated group

$$\mathbb{E}[Y_{i,Post}(0)|D_i = 1] =$$

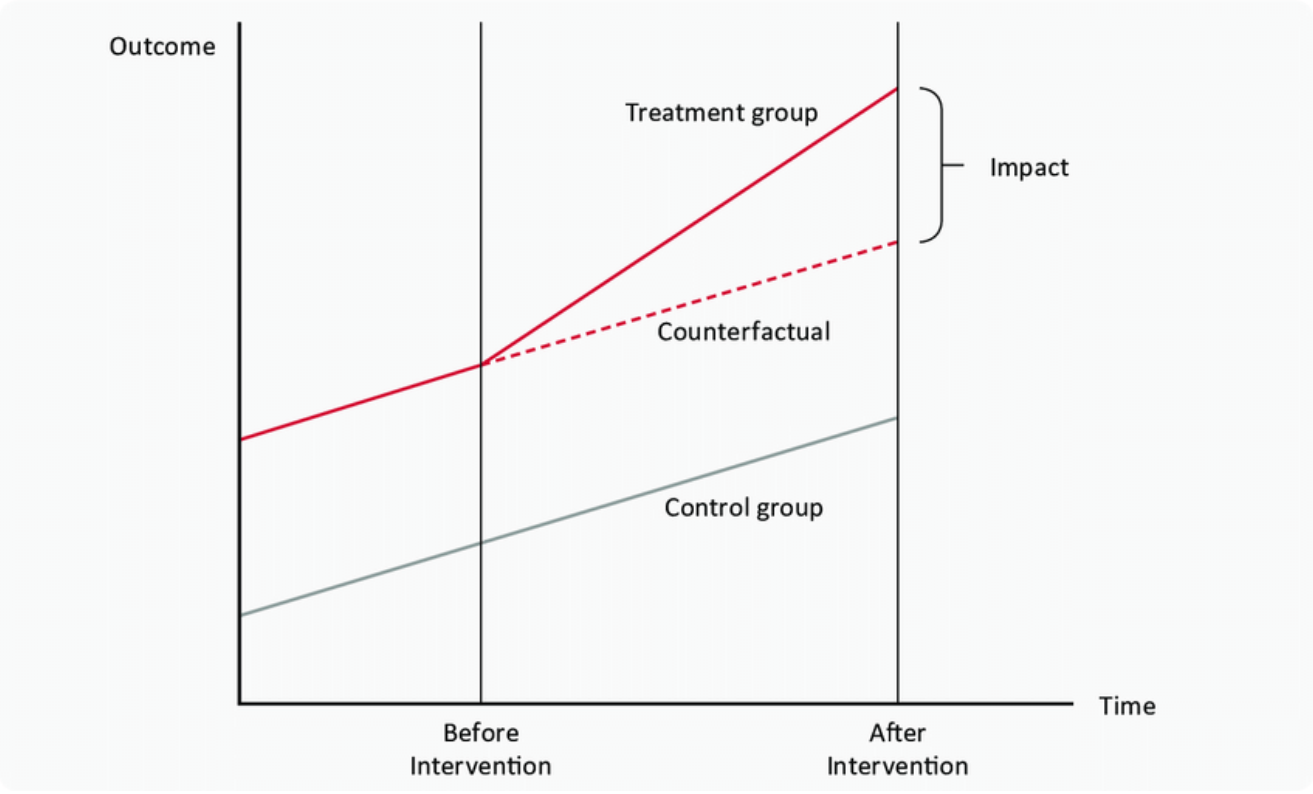
$$\mathbb{E}[Y_{i,Pre}(0)|D_i = 1] + (\mathbb{E}[Y_{i,Post}(0)|D_i = 0] - \mathbb{E}[Y_{i,Pre}(0)|D_i = 0])$$

Step 1: the 2×2 case

Counterfactual and PT

$$\mathbb{E}[Y_{i,Post}(0)|D_i = 1] =$$

$$\mathbb{E}[Y_{i,Pre}(0)|D_i = 1] + (\mathbb{E}[Y_{i,Post}(0)|D_i = 0] - \mathbb{E}[Y_{i,Pre}(0)|D_i = 0])$$



Step 1: the 2×2 case

DID $\rightarrow$ ATT?

How does that help to estimate the ATT? Let's see it with the target $ATT(Post)$

$$
\begin{aligned}
&= \overbrace{\mathbb{E}[Y_{i,Post}|D_i=1]}^{\mathbb{E}[Y_{i,Post}|D_i=1]} - \overbrace{(\mathbb{E}[Y_{i,Pre}|D_i=1] + (\mathbb{E}[Y_{i,Post}|D_i=0] - \mathbb{E}[Y_{i,Pre}|D_i=0]))}^{\mathbb{E}[Y_{i,Post}(0)|D_i=1]} \\
&= \underbrace{\underbrace{\mathbb{E}[Y_{i,Post}|D_i=1] - \mathbb{E}[Y_{i,Pre}|D_i=1]}_{\mathbb{E}[\Delta Y_i|D_i=1]} - \underbrace{((\mathbb{E}[Y_{i,Post}|D_i=0] - \mathbb{E}[Y_{i,Pre}|D_i=0]))}_{\mathbb{E}[\Delta Y_i|D_i=0]}}_{ATT(Post)\star}
\end{aligned}
$$

Notice that:

- LHS: average Pre/Post change in outcome among treated
- RHS: average Pre/Post change in outcome among controls
- The 2×2 DID is estimated by taking the difference between these two differences (the difference in differences)

Step 1: the 2×2 case

Why is DID used so frequently?

- It has very mild data requirements: in simple cases without covariates, $\mathbb{E}[\Delta Y_i | D_i = 1]$ and $\mathbb{E}[\Delta Y_i | D_i = 0]$ can be simply obtained by computing 4 means
- DID is intuitive, whether you look at it graphically or algebraically
- Its identifying assumptions can be stated precisely

Step 1: the 2×2 case

Caveat on PT #1

- PT makes DID distinct from causal designs that are based on statistical independence between treatment and potential outcomes (e.g. randomized studies, instrumental variables)
- Because treatment adoption is often chosen by economic actors or policymakers "inside the model", PT need not hold automatically
- DID analyses relies on the plausibility of the parallel trends assumption in the specific application!
- There are various empirical approaches to support the validity of the PT assumption (see later), but these CANNOT substitute a a rigorous knowledge and discussion of context, institutions, and agents' choices

Step 1: the 2×2 case

Caveat on PT #2

- In realistic scenarios, PT can only hold under some restrictions on the way untreated outcomes enter the treatment selection mechanism
- Example 1: Treatment selection depends on the permanent component of $Y_{it}(0)$ (fixed effects) but not on shorter-term fluctuations ("shocks")
  - Differences between treated and controls are differenced away
- Example 2: Treatment selection also depended on information about a specific pre-treatment value of the outcome
  - In this case, PT would hold only if one imposes stronger time-series restrictions on $Y_{it}(0)$

Step 1: the 2×2 case

Estimation

Estimating a canonical 2×2 DID is very simple!

$$\widehat{ATT(Post)} = (\bar{Y}_{Post,D=1} - \bar{Y}_{Pre,D=1}) - (\bar{Y}_{Post,D=0} - \bar{Y}_{Pre,D=0})$$

- Method 1: $\bar{Y}_{t,D}$ is a simple sample average, so computing four sample averages we can get our estimate
- Method 2: alternatively, we can write a linear regression formulation and get all the coefficients we need from one model:

$$Y_{it} = \beta_0 + \beta_1 D_i + \beta_2 Post_t + \beta^{DID} D_i \times Post_t + \epsilon_{it}$$

Step 1: the 2×2 case

OLS formulation

$$Y_{it} = \beta_0 + \beta_1 D_i + \beta_2 Post_t + \beta^{DID} D_i \times Post_t + \epsilon_{it}$$

OLS models the conditional average of Y_{it}, that is $\mathbb{E}[Y_{it}|X_{it}]$ (though we have no X_{it} yet). This means that

- $\bar{Y}_{Post,D=1}$ (in Post and for treated)? $= \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}^{DID}$
- $\bar{Y}_{Pre,D=1}$ (in Pre and for treated)? $= \hat{\beta}_0 + \hat{\beta}_1$
- $\bar{Y}_{Post,D=0}$ (in Post and for controls)? $= \hat{\beta}_0 + \hat{\beta}_2$
- $\bar{Y}_{Pre,D=0}$ (in Pre and for controls)? $= \hat{\beta}_0$

Plugging them all in, we realise that

$$\widehat{ATT(Post)}) = \Big[(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}^{DID}) - (\hat{\beta}_0 + \hat{\beta}_1)\Big] - \Big[(\hat{\beta}_0 + \hat{\beta}_2) - \hat{\beta}_0\Big]$$

$$= \hat{\beta}^{DID}{}_\star$$

Step 1: the 2×2 case

Inference

- Inference refers to how we compute the uncertainty around our estimated parameters; it is COMPLICATED and we cannot cover it in full today
- The Guide and myself refer to the following fundamental resources for empiricists
  - "Sampling-Based versus Design- Based Uncertainty in Regression Analysis" by Abadie, Athey, Imbens, Wooldridge (2020). `https://doi.org/10.3982/ECTA12675`
  - "When Should You Adjust Standard Errors for Clustering?" by Abadie, Athey, Imbens, Wooldridge (2020). `https://doi.org/10.1093/qje/qjac038`
  - "What's trending in difference-in-differences? A synthesis of the recent econometrics literature" by Roth, Sant'Anna, Bilinski, Poe (2023). `https://doi.org/10.1016/j.jeconom.2023.03.008`
  - "Causal Inference: The Mixtape" by Cunningham (2021). `https://mixtape.scunning.com/`
- In our analyses, we cluster standard errors at the level of GP practices

First Break + Stata application # 1

Let's take a 10min break

When we come back we'll do see how
these work out in Stata

`https://github.com/HOPE-UoM/training_didguide`

Simulation case study

- You may well have good real data examples
- However, a simulation case study allows us to have full control on treatment effect and confounding channels
- This means we can precisely assess the performance our our empirical approach in terms of bias
- Let's try to imagine that the simulated data represents GP practice level data over time
- We are interested in studying cancer diagnostic accuracy of GPs
- A new technology (say a clinical Chat bot that supports cancer referral decisions) becomes available and is adopted by GPs over time

Empirical case study

Setting

- We have data for 10 years
- Our outcome measure is the **conversion rate** (y_{it}), that is the percentage of urgent suspected cancer referrals that result in a cancer diagnosis available and is adopted by GP practices over time
- We observe covariates, in our case Case-mix (casemix_{it})
- We want to understand the **effect the adoption of the new chat bot** on conversion rates
- The adoption of the chat bot is linked to unobservable ability and potentially effort (we'll see later)

Step 2: Incorporating covariates

Three reasons

- Now we've got the basics of the canonical 2×2 design
- In The Guide, this is the fundamental building block of DID analysis
- We'll stay on 2×2 but jump into Step 2 and discuss the role of covariates (yes, so far we haven't really mentioned X_{it})
- Similar to most regression analyses, in DID covariates have various roles that we should think about when planning an analysis
  1. Checking for balance in variables thought to influence the outcome
  2. "Controlling for" those variables in the main estimates
  3. Estimating treatment effect heterogeneity
- Here we follow The Guide and briefly cover how to think about using covariates to
  - Evaluate the plausibility of PT (conditionally vs. unconditionally)
  - Identify ATT parameters under potentially weaker assumptions
  - Study heterogeneity

Step 2: Incorporating covariates

Evaluating the PT assumption: balance in covariates

In DID analyses it is customary to check for balance across groups in

- Baseline (pre) covariate levels

$$\mathbb{E}[X_{i,Pre}|D_i = 1] - \mathbb{E}[X_{i,Pre}|D_i = 0]$$

- Covariate trends from pre- to post-treatment

$$\mathbb{E}[\Delta X_{i,Post}|D_i = 1] - \mathbb{E}[\Delta X_{i,Post}|D_i = 0]$$

- We can obtain these quantities by computing simple sample averages
- It is often useful to also report normalised differences, with the rule of thumb that they are acceptable up to 0.25
- We'll see it later in Stata

Step 2: Incorporating covariates

What about balance?

- PT is fundamentally untestable
- We indirectly test it relying on observable information that we assume to be related to untreated potential outcomes
- The PT assumption is related to changes over time; because of this it is sometimes argued that pre-treatment differences in levels are acceptable as they're differenced out
- This logic does not hold, though, if baseline covariates are related to untreated potential outcome trends themselves
- Balance is a good indication that PT holds, but it is not sufficient
- Similarly, imbalances may reveal that some variables are mechanisms/outcomes rather than covariates

Step 2: Incorporating covariates

Which covariates?

- Whether something is a covariate or a mechanism is not a data question per se: It requires context-specific knowledge (or assumptions) about how treatment works
- Include in X_{it} all the determinants of the change in untreated potential outcome or of the treatment assignment
- Beware of "bad controls"

Step 2: Incorporating covariates

Identifying ATT under conditional PT

- Having detected covariate imbalance that casts doubt on Assumption PT, how should we proceed to estimate ATT? One option is assuming Conditional PT (CPT)
- CPT: conditional on a same level of covariates X_{it}, the evolution of outcomes between treated and control groups, in absence of the treatment
- CPT is the same as PT but holds within each covariate-specific stratum rather than across the whole population. Formally:

$$\mathbb{E}[Y_{i,Post}(0)|X_{it}, D_i = 1] - \mathbb{E}[Y_{i,Pre}(0)|X_{it}, D_i = 1] =$$

$$\mathbb{E}[Y_{i,Post}(0)|X_{it}, D_i = 0] - \mathbb{E}[Y_{i,Pre}(0)|X_{it}, D_i = 0]$$

- At what cost? One additional assumption, that is X_{it} need to be well defined for both treated and controls (strong overlap assumption)

Step 2: Incorporating covariates

ATT under CPT

If CPT and strong overlap hold, we the ATT can be rewritten as follows (see p. 18 of The Guide for more):

$$ATT(Post) =$$

$$\mathbb{E}[\Delta Y_{i,Post}|D_i = 1] - \mathbb{E}[\mathbb{E}[\Delta Y_{i,Pre}|X_{it}, D_i = 0|D_i = 1]]$$

Intuition: the ATT is equal to the path of outcomes experienced by the treated group (LHS) minus the average path of outcomes in the comparison group for each value of the covariates, averaged over the treated group's distribution of covariates (RHS).

Step 2: Incorporating covariates

TWFE and DID estimation under CPT

- Practitioners have traditionally used TWFE including X_{it}
- Recent work on DID showed that this is not an innocent choice
- The resulting TWFE coefficient can miss the ATT because
  - Biased due to model misspecification
  - Biased due to inclusion of covariates that are "bad controls" (i.e. affected by the treatment)
  - Missing crucial effect heterogeneity if not interacted with the treatment
- The Guide once again offers a "forward engineering" approach that starts from 2×2 building blocks and uses more appropriate estimation techniques that avoid these issues

Step 2: Incorporating covariates

Reliable DID estimation under CPT

The goal is to find a suitable way to estimate the following expectations

$$ATT(Post) =$$

$$\mathbb{E}[\Delta Y_{i,Post}|D_i = 1] - \mathbb{E}[\mathbb{E}[\Delta Y_{i,Pre}|X_{it}, D_i = 0|D_i = 1]]$$

- The first term (LHS) can be easily estimated with sample averages, whilst the RHS is a bit more complicated
- Without getting lost in the details, for the second term The Guide proposes three competing estimation approaches
  - Regression adjustment (RA)
  - Inverse Probability Weighting (IPW)
  - Doubly Robust (DR) estimator (that is RA combined with IPW)

Step 2: Incorporating covariates

DID estimation under CPT: In short

- Imbalance in covariates is suggestive evidence that unconditional PT is not realistic; CPT is warranted
- Covariates must satisfy the overlap condition (i.e. must span across the entire distribution for both treatment and controls)
- Old school TWFE including covariates should be avoided unless the setting (and specifically the mechanism guiding selection into treatment) is extremely clean
- For canonical 2×2 settings, The Guide proposes alternative estimators with relative advantages/disadvantages
  - RA: biased if model not specified correctly, but good with small N
  - IPW: good with large imbalances, but needs strong overlap and large N
  - DR: combines RA and IPW, best option if N large enough
- If feasible, the Doubly Robust approach should be preferred as it offers greater protection against model misspecification

Stata application # 2

Back to your do-files in Stata!

Step 3: Multiple time periods, same treatment time

Motivation

In our day-to-day research...

- We often analyse settings where we observe more than 2 periods
- We like to estimate event study plots to (a) give a sense of how ATT evolves over time, and (b) support the PT assumption
- We rarely have treatment happening at the same time for all treated units; we're most faced with a staggered roll-out of treatments/policies
- Once again, recent work highlighted how TWFE likely returns biased estimates for both individual event studies coefficients and aggregate ATT estimates (a summary ATT of effects across all post-treatment periods)
- In the next (and final) part of the workshop we'll discuss the modern approach proposed by The Guide

Step 3: Multiple time periods, same treatment time

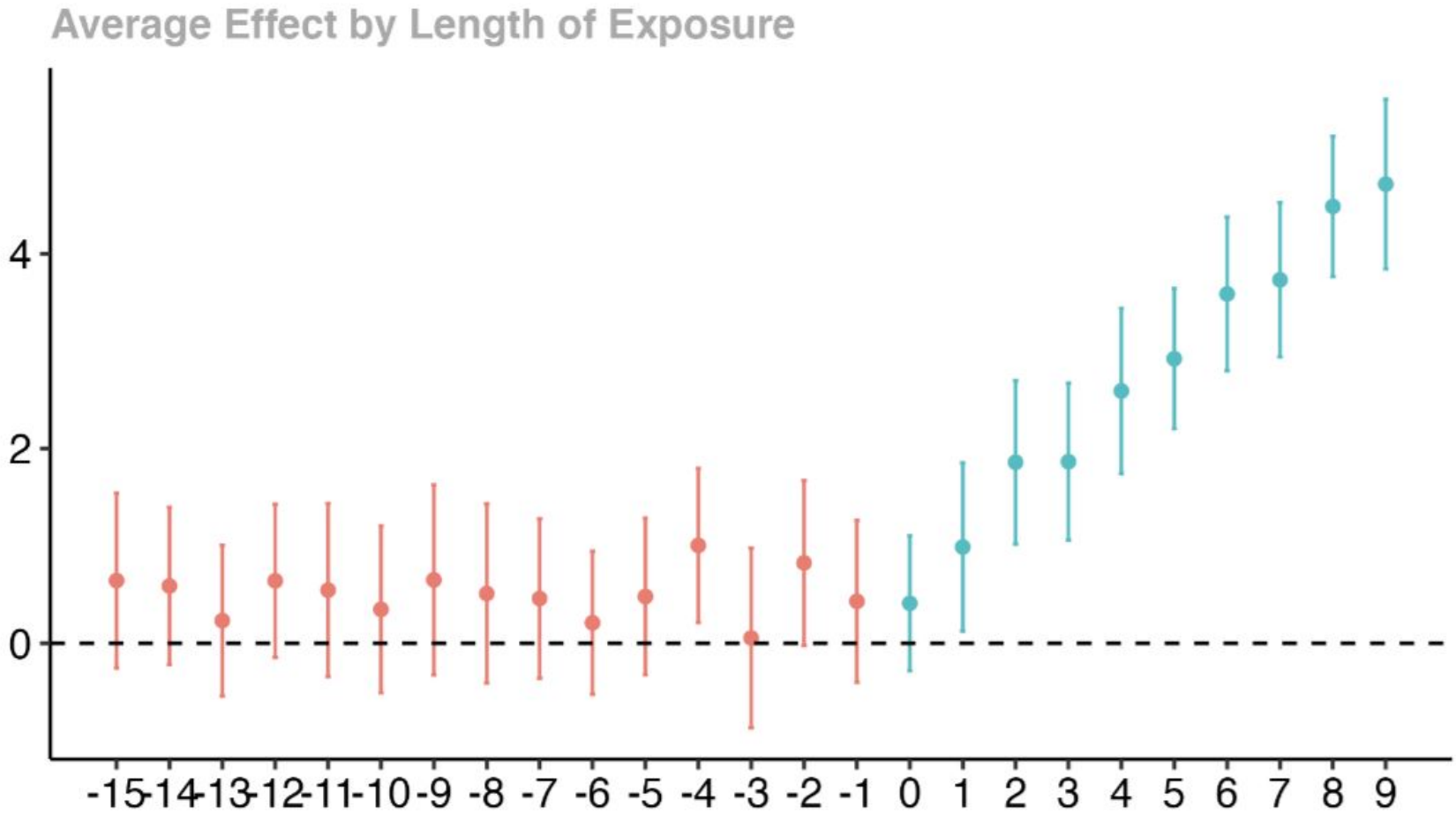
Event study, single treatment timing, post-treatment

Let's start by the simplest case:

- We observe i units over T periods
- For all treated units ($D_i = 1$), treatment happens at the same $t = g$
- Let's start without covariates
- In this setting, we can define an $ATT(t)$ for all post-treatment periods $t \geq g$
- Why would we want to study the ATT over time?
  - People and institutions adjust their behaviour to policies over time
  - If looking at health outcomes, economic models that view health as a stock imply that ATT may grow if the policy stimulates health investment health investments.
- The next slide shows what a typical event study plot looks like

Step 3: Multiple time periods, same treatment time

A typical event study plot: the focus is on the blue dots, **post treatment** $ATT(t)$



Source: Roth (2024)

Step 3: Multiple time periods, same treatment time

Identification of $ATT(t)$ in event studies

- Identification of individual $ATT(t)$'s requires the same assumptions as the 2×2 case (no anticipation and PT), but PT is required for every post-treatment period!
- E.g. say we observe 4 post-treatment periods. If we want to identify $ATT(t = 4)$ then we need PT for $t = 1, 2, 3$ to hold
- Note: that we tend naturally to estimate effects relative to the time period immediately before treatment, $g - 1$
- If no anticipation and (simple, unconditional) PT hold we the $ATT(t)$ at a given time point is given by

$$\widehat{ATT(t)} =$$

$$(\mathbb{E}[Y_{i,t}|D_i = 1] - \mathbb{E}[Y_{i,g-1}|D_i = 1]) - (\mathbb{E}[Y_t|D_i = 0] - \mathbb{E}[Y_{g-1}|D_i = 0])$$

Step 3: Multiple time periods, same treatment time

Estimation of $ATT(t)$ in event studies

- Surprise surprise, THE GUIDE approaches estimation of $ATT(t)$ treating it as a 2×2 building block
- This "forward engineering" approach is natural given that - as we've seen - the individual $ATT(t)$ is a 2×2 comparison with the $t = g - 1$
- In this simple setting without covariates we can simply estimate

$$\widehat{ATT(t)} = (\bar{Y}_{D=1,t} - \bar{Y}_{D=1,g-1}) - (\bar{Y}_{D=0,t} - \bar{Y}_{D=0,g-1})$$

- Method 1: compute 4 sample averages
- Method 2: use a TWFE regression formulation

$$Y_{it} = \theta_t + \eta_i + \sum_{k=1}^{g-2} \beta_k \mathbf{1}\{G_i = g\}.\mathbf{1}\{t = k\} + \sum_{k=g}^{T} \beta_k \mathbf{1}\{G_i = g\}.\mathbf{1}\{t = k\} + \epsilon_{it}$$

Step 3: Multiple time periods, same treatment time

Inference in estimating $ATT(t)$'s in event studies

- Computing 4 sample averages separately or estimating them all at once with TWFE gives point-wise robust standard errors (in our case study, cluster-robust standard errors at the level of GP practices)
- BUT, we are estimating many treatment effect parameters to study the evolution of $ATT(t)$ over time $\rightarrow$ Multiple hypothesis testing!
- The most recent packages (e.g. Callaway and Sant'Anna 2021) adjustment standard errors automatically when estimating coefficients for event study plots

Step 3: Multiple time periods, same treatment time

Aggregating $ATT(t)$'s hard with TWFE

- Even when we observe multiple time periods, we often want to have a single summary measure capturing the overall $ATT(Tot)$ of a given treatment
- A common shortcut was to run a second TWFE regression (instead of the event study)

$$Y_{it} = \theta_t + \eta_i + {}^{\beta\ OLS} D_{i,t} + \epsilon_{it}$$

where $D_{i,t}$ is 1 for treated unit in post-treatment periods

- Unfortunately, generally $\hat{\beta}^{OLS} \neq \widehat{ATT(Tot)}$
- What to do then?

Step 3: Multiple time periods, same treatment time

Aggregation according to The Guide

- The Guide suggest to once again start form individual 2×2 building blocks
- A convenient scalar summary is the average of $ATT(t)'s$

$$ATT^{Avg} = \frac{1}{T-(g-1)} \sum_{t=g}^{T} ATT(t)$$

- Callaway and Sant'Anna (2021) discuss more options

Step 3: Multiple time periods, same treatment time

Pre-treatment event study estimates to test PT?

- We are used to estimated event study coefficients also for pre-treatment periods
- We tend to interpret these "pre-trends" as evidence for plausibility (or not) of the PT assumption
- Although we prefer parallel pre-trends, how informative pre-trends are for PT is case-specific
- Caveats
  - PT is not testable and only makes restrictions on untreated potential outcomes in post periods
  - Under no anticipation, pre-trends measure something similar as parallel trends but in the "wrong" periods!
  - Sometimes, looking too far back is not informative about PT in post-treatment periods, as the economic environment may be too different
- See work by Jonathan Roth for more details

Step 3: Multiple time periods, same treatment time

Conditioning on covariates?

- Given that we work on 2×2 building blocks, the ways to incorporate covariates discussed there immediately apply to each event study estimate
- The only difference is that instead of using "short-differences" $\Delta Y_{i,Post}$ we now use "long-differences", $Y_{i,t} - Y_{i,g-1}$
- The same estimation approaches (RA, IPW, and DR, with preference for the latter) can be employed, under the same conditions

Step 4: Staggered treatment adoption

Motivation

- In Step 3 we've discussed cases with multiple time periods but a same treatment timing
- In practice we often deal with policies characterised by staggered adoption: some units are treated today, some in a year from now, and others in two years
- Before the recent innovation in DID, in staggered cases we would simply estimate event study coefficients relative to each unit's treatment timing using TWFE
- This is very problematic, for all the reasons discussed above (especially when including covariates) AND because TWFE implicitly uses already-treated comparison groups, which results in negatively weighting some ATTs

Step 4: Staggered treatment adoption

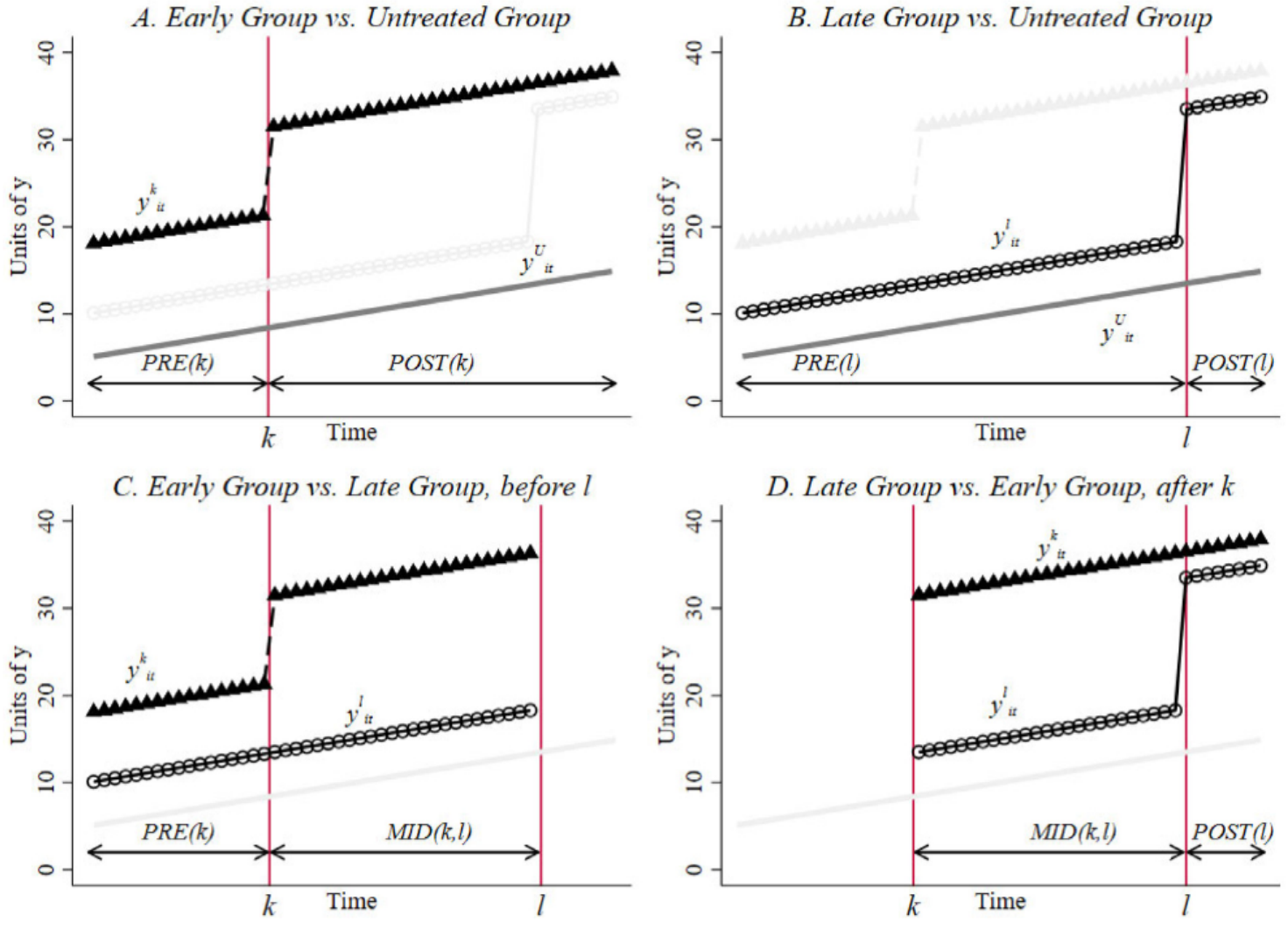
Approach suggested by The Guide

Everything we discussed in Step 3 extends to designs with staggered treatment adoption. The key differences are related to

- The natural grouping $g \in G$ of units in treatment cohorts
  - each has a specific set of $ATT^g(t)$
  - we need to think about how to aggregate them, typically weighting larger treatment groups more than smaller ones
- The type of PT assumption that we impose
  - PT based on never-treated units: this is the same as the one in Step 3. In staggered designs can be undesirable if never-treated units are too different
  - PT based on not-yet-treated units: middle step that uses all not-yet-treated units as comparison. It uses more information which can lead to gains in precision and helps to incorporate covariates.
  - PT across all periods and groups: uses all data but can be restrictive as PT are unlikely to hold across all periods.

Step 4: Staggered treatment adoption

From Goodman-Bacon (2021)



Stata application # 4

Back to your do-files in Stata!

Step 5: The Guide's conclusion

The Concluding section of The Guide offers a condensed check-list style memo of all these steps to follow in your own analysis!

Step 5: The Guide's conclusion

6 Conclusion

The starting point of this paper was a 2×2 DiD design that researchers have been using for almost 200 years. The end point was a design with five treatment groups, 11 years of data, six covariates, three types of parallel trends assumptions, and four estimation techniques. Our fundamental message is that without understanding how complex designs are built up from simpler ones, it is exceedingly difficult to navigate all the empirical tools now available for DiD designs. This lesson applies not only to the design details we considered here—weighting, covariates, and staggered designs—but to any DiD design.

The forward-engineering philosophy we followed in this paper suggests a set of steps that researchers can follow in any DiD study:

Step 1. *Define target parameters.* Adopt a potential outcomes notation that fits the study's specific setting and use it to define causal target parameters that answer the study's motivating question. Building block causal parameters usually aggregate across units using (conditional) weighted averages, and summary target parameters aggregate across the building blocks. This step fixes the study's goals in terms of causal quantities and facilitates comparisons with related studies.

Step 2. *State (formally) the identification assumptions.* DiD studies leverage parallel trends assumptions, but they also rely on no-anticipation and, in some cases, overlap conditions, or more. Be explicit about which form of these assumptions is required for identification in the study. Engage with the theoretical arguments necessary for them to hold and generate appropriate

Conclusion

That's all folks!
Thank you