

Credible Answers to Hard Questions: Differences-in-Differences for Natural Experiments

Clément de Chaisemartin¹

Xavier D'Haultfœuille²

October 25, 2025

¹Economics Department, Sciences Po, clement.dechaisemartin@sciencespo.fr

²CREST-ENSAE, xavier.dhaultfoeuille@ensae.fr

Textbook in progress, comments welcome.

Under contract with Princeton University Press.

Copyright, 2023, Clément de Chaisemartin, Xavier D'Haultfœuille.

Contents

I	Introduction and setup	7
1	Introduction	9
1.1	Overview	11
1.2	Pedagogy	20
2	Data, notation, and assumptions	23
2.1	Data requirements: group-level panel data	23
2.2	Treatment and potential outcomes	24
2.3	Assumptions	24
2.4	Discussion of the book's perspective on statistical inference*	28
II	The classical design	33
3	The classical DID design	35
3.1	Target parameters	37
3.2	Two-Way Fixed Effects estimators	39
3.3	Inference on the ATT and on the event-study effects	53
3.4	Limitations of pre-trend tests	67
3.5	An alternative estimator of the event-study effects	74
3.6	Estimating heterogeneous treatment effects	79
3.7	Non-linear DID	84
3.8	Instrumental-variable DID estimators*	94
3.9	Further topics*	97

3.10 Next steps	103
3.11 Appendix*	104
4 Alternatives to parallel trends	107
4.1 TWFE and DID estimators with control variables	107
4.2 Interactive FEs, synthetic controls, and synthetic DID	129
4.3 Bounded differential trends	153
4.4 Appendix*	158
III Beyond the classical design	163
5 TWFE estimators outside of the classical design	165
5.1 A decomposition of $\hat{\beta}^{\text{fe}}$	168
5.2 $\hat{\beta}^{\text{fe}}$ may be biased for the ATT	170
5.3 $\hat{\beta}^{\text{fe}}$ may not estimate a convex combination of treatment effects	173
5.4 Decompositions of related estimators	175
5.5 Stata and R commands to compute the weights	176
5.6 Application	177
5.7 Next steps	180
6 Designs with variation in treatment timing	181
6.1 Target parameters	183
6.2 Two-Way Fixed Effects estimators	185
6.3 Heterogeneity-robust estimators	209
6.4 Estimating heterogeneous treatment effects	232
6.5 Non-linear DID	235
6.6 Further topics*	244
6.7 Appendix*	247
7 Designs with variation in treatment dose	251
7.1 Identifying assumptions	255
7.2 Target parameters	256

CONTENTS	5
7.3 TWFE estimator in heterogeneous-adoption designs	262
7.4 Heterogeneity-robust estimators	280
7.5 Appendix*	294
8 General designs	297
8.1 Static Two-Way Fixed Effects estimator	300
8.2 Distributed-Lag Two-Way Fixed Effects estimators	315
8.3 Heterogeneity-robust estimators	318
8.4 Heterogeneity-robust estimators, ruling out dynamic effects	346
8.5 Heterogeneity-robust estimators in designs without stayers?	358
8.6 Appendix*	361
9 Conclusion: practitioners' checklist	365

Part I

Introduction and setup

Chapter 1

Introduction

Many scientific questions are causal inference questions. Much of science is concerned with causal inference, namely estimating the effect of a “treatment” on an “outcome”. For that purpose, a gold-standard method is to run a randomized experiment, where units’ exposure to treatment is determined randomly. Then, one can compare the average outcome of treated and untreated units, to unbiasedly estimate the so-called average treatment effect (ATE), namely the average effect of the treatment in the population of interest.

In the social sciences, many important causal-inference questions are hard or impossible to study using a randomized experiment. For instance, one cannot randomly increase imports from China to some countries and not to other countries, so as to measure the effect of imports from China on destination-countries’ employment. Similarly, randomly assigning firms to high- and low-minimum-wage groups to study the minimum wage’s effect on employment has never been done so far. Even when they are feasible, randomized experiments sometimes lack “external validity”: their findings may not be extrapolated from the experimental sample to the population whose ATE the researcher would like to learn. One reason is that research ethics requires enlightened consent to include subjects in an experiment. Then, the hypothetical minimum wage experiment would unbiasedly estimate the minimum wage’s effect among firms that enrolled in the experiment, but that effect would probably differ from the minimum wage’s effect among all firms.

Instead, social scientists often resort to natural experiments. To answer hard causal-inference questions for which randomized experiments are unfeasible or would lack external validity, researchers usually rely on so-called natural experiments, where a treatment and a control group arise naturally, without any intervention from the researcher.¹ Typically, natural experiments used in the social sciences are policy changes. For instance, a US state increases its minimum wage while the neighboring state does not, thus giving researchers a treatment group facing a high minimum wage, and a control group facing a lower minimum wage. Natural experiments often affect an entire region or province, so the findings from studies leveraging natural experiments typically apply to large and unselected populations, unlike the findings from randomized experiments.

Two crucial differences between randomized and natural experiments. In a natural experiment, assignment to the treatment is not randomized by a researcher, it is decided by a policy maker. In a sharp reversal of Keynes' famous quote, modern applied researchers, who believe themselves to be quite exempt from any practical influences, have to hope that practical men will give them good natural experiments they can work with. This fact has two important consequences. First, as policy makers do not randomly choose where to implement a policy change, treated and control groups may not be comparable, and a simple comparison of their mean outcome may not yield an unbiased estimator of the ATE. In the minimum wage example, the treated and control states may for instance have different employment levels even before the treated state increased its minimum wage. Then, comparing their employment levels after that increase will measure the sum of the minimum wage's effect, and of pre-existing differences between the two states. Therefore, this comparison will not isolate the minimum wage's effect. A second crucial difference is that while a researcher running a randomized experiment would just randomly assign some states to the treatment and others to the control, and then observe what happens in the two groups for as long as their research question dictates it, legislative changes are not made to help researchers estimate causal effects, and they are often full of twists and turns. In the minimum wage example, policy makers in states that raised their minimum wage may then decide to decrease it, or to increase it again. And in control states that initially

¹We adopt here a broad definition of “natural experiments”, not restricting it to cases where nature acts as a randomization device.

did not change their minimum wage, policy makers may decide to change it at a later point. Also, some states may implement large minimum wage increases, while other states implement smaller increases. This creates lots of treatment variation, which complicates the analysis, and may preclude researchers from estimating the clean-and-simple effects they were initially hoping to estimate. The complexity of the majority of natural experiments leveraged by social scientists to learn causal effects is a central theme of this book.

Statement of purpose. The purpose of this book is to introduce applied researchers to modern Differences-in-Differences (DID) estimators, tailored to potentially complicated natural experiments, that they can use to obtain credible answers to hard causal inference questions, for which randomized experiments are unfeasible.

1.1 Overview

1.1.1 The classical design

The classical DID design. After presenting the book’s set-up and notation in Chapter 2, Chapter 3 reviews the classical difference-in-differences (DID) design. In that design, the treatment is a binary variable, the treatment is absorbing, meaning that one cannot switch out of treatment once treated, and there is no variation in treatment timing: all treated groups are treated at the same time. In this introduction, we consider a simplified version of this design, with two groups and two periods: group s switches from untreated to treated from period 1 to 2, and group n is untreated at both dates. Throughout the book, a “cell” refers to a pair (g, t) , namely a given group at a given period of time. In the simplified example we consider for now, we have four cells: $(s, 1)$, $(s, 2)$, $(n, 1)$, and $(n, 2)$.

Potential and observed outcomes. Let us also introduce simplified potential outcome notation.² For $g \in \{s, n\}$ and $t \in \{1, 2\}$, let $Y_{g,t}(0)$ and $Y_{g,t}(1)$ denote the potential outcomes in g at t without and with treatment, respectively (Neyman, Dabrowska and Speed, 1990; Rubin,

²The next chapter introduces the general potential outcome notation used throughout the book.

1974). In the minimum wage example, $Y_{g,t}(0)$ is the employment level that g will have at t if its minimum wage is low, and $Y_{g,t}(1)$ is the employment level that g will have at t if its minimum wage is high. Let $Y_{g,t}$ denote the observed outcome in g at t . If cell (g,t) is untreated, its observed outcome is its potential outcome without treatment: $Y_{g,t} = Y_{g,t}(0)$. If (g,t) is treated, its observed outcome is its potential outcome with treatment: $Y_{g,t} = Y_{g,t}(1)$. Letting $D_{g,t}$ be an indicator equal to one if (g,t) is treated, we have the following relationship between the observed and potential outcomes:

$$Y_{g,t} = (1 - D_{g,t})Y_{g,t}(0) + D_{g,t}Y_{g,t}(1).$$

Justify the previous equality, by showing that it holds for any value that $D_{g,t}$ can take.

If $D_{g,t} = 1$, the observed outcome $Y_{g,t}$ is equal to $Y_{g,t}(1)$, and the right-hand-side of the previous display is also equal to $Y_{g,t}(1)$ so the equality holds. If $D_{g,t} = 0$, the observed outcome $Y_{g,t}$ is equal to $Y_{g,t}(0)$, and the right-hand-side of the previous display is also equal to $Y_{g,t}(0)$ so the equality holds. The unobserved potential outcome of (g,t) is referred to as its counterfactual outcome.

Target parameter, and three possible estimators. We would like to estimate

$$E(Y_{s,2}(1) - Y_{s,2}(0)),$$

the average effect of the treatment in group s at period 2. $(s,2)$ is the only treated (g,t) cell, so $E(Y_{s,2}(1) - Y_{s,2}(0))$ is the average treatment effect on the treated (ATT). As group s is treated in period 2, $Y_{s,2}(1)$ is just $Y_{s,2}$, the observed outcome in s at period 2. Therefore, $Y_{s,2}(1)$ does not need to be estimated. On the other hand, $Y_{s,2}(0)$, the counterfactual outcome that s would have had at period 2 if it had been untreated, is unobserved and has to be estimated. To estimate $E(Y_{s,2}(1) - Y_{s,2}(0))$, we could use

$$Y_{s,2} - Y_{n,2}, \tag{1.1}$$

a *treated-versus-control* comparison of s 's and n 's outcomes at period 2. However, which group got treated was not determined randomly, so $Y_{n,2}$ may not be a good estimator of $Y_{s,2}(0)$, the

outcome that s would have had at period 2 without treatment. Alternatively, we could use

$$Y_{s,2} - Y_{s,1}, \quad (1.2)$$

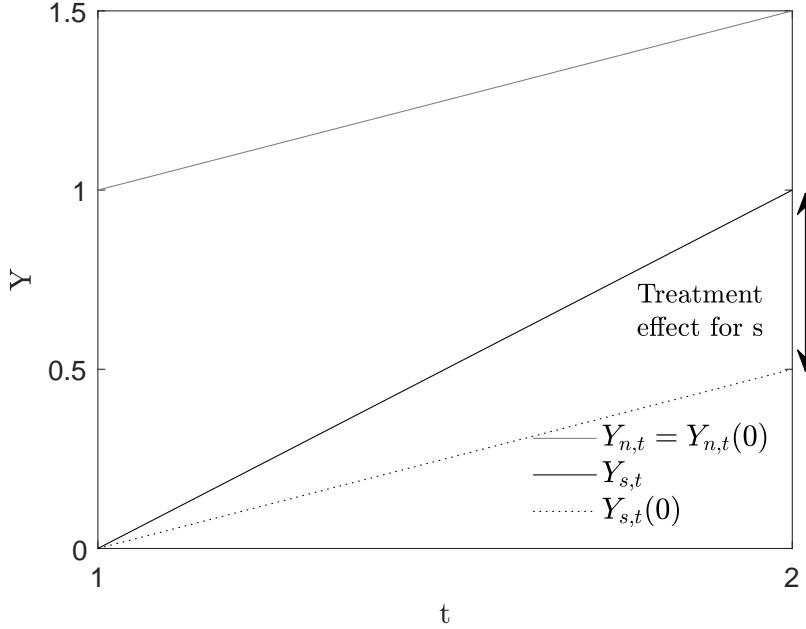
a *before-after* comparison of s 's outcome at periods 1 and 2. However, perhaps that s 's outcome would have changed from period 1 to 2 even without treatment, so $Y_{s,1}$ may again not be a good estimator of $Y_{s,2}(0)$. Instead, we could use

$$\text{DID} := Y_{s,2} - Y_{s,1} - (Y_{n,2} - Y_{n,1}), \quad (1.3)$$

the comparison of the period-1-to-2 outcome evolutions of s and n . DID is called a Difference-in-Differences estimator: the before-after longitudinal difference in (1.2) is combined with the treated-versus-control cross-sectional difference in (1.1). **For DID to be a reliable estimator of the treatment effect, one assumption has to hold, which one?**

Parallel-trends: graphical intuition. DID compares groups' outcome evolutions. If, even without the treatment, the outcome would have, say, increased more in group s than in group n , DID could be strictly positive even if the treatment has no effect at all. Thus, for DID to be a reliable estimator of the treatment's effect, a so-called “parallel-trends” assumption has to hold: in the absence of the treatment, both groups would have experienced the same outcome evolutions. This assumption is illustrated graphically in Figure 1.1 below. The solid grey line represents the outcome evolution from period 1 to 2 of group n , without the treatment. The dashed grey line represents the counterfactual outcome evolution that group s would have had without the treatment. We see that the two lines are parallel: the parallel-trends assumption holds. Accordingly, DID is a reliable estimator of the treatment effect. The outcome of s increases from 0 to 1, while that of n increases from 1 to 1.5. Therefore $\text{DID} = 1 - 0 - (1.5 - 1) = 0.5$. And 0.5 is also the value of the treatment effect in group s at period 2: with the treatment the outcome of s at period 2 is equal to 1, while the dashed line shows that without the treatment its outcome would have been equal to 0.5: $1 - 0.5 = 0.5$.

Figure 1.1: Illustration of the parallel-trends assumption



Parallel-trends: formal statement. Mathematically, the parallel-trends assumption requires that

$$E [Y_{s,2}(0) - Y_{s,1}(0)] = E [Y_{n,2}(0) - Y_{n,1}(0)] : \quad (1.4)$$

s and n have the same average outcome evolutions without treatment. Under that assumption, DID is unbiased for our target parameter:

$$\begin{aligned} E [\text{DID}] &= E [Y_{s,2} - Y_{s,1} - (Y_{n,2} - Y_{n,1})] \\ &= E [Y_{s,2}(1) - Y_{s,1}(0) - (Y_{n,2}(0) - Y_{n,1}(0))] \\ &= E [Y_{s,2}(1) - Y_{s,2}(0)] + E [Y_{s,2}(0) - Y_{s,1}(0)] - E [Y_{n,2}(0) - Y_{n,1}(0)] \\ &= E [Y_{s,2}(1) - Y_{s,2}(0)]. \end{aligned} \quad (1.5)$$

Justify each equality.

The first equality follows from the fact that $(s, 2)$ is treated, and all other cells are untreated. The second equality follows from adding and subtracting $Y_{s,2}(0)$. The third equality follows from the parallel-trends assumption. Many unbiasedness proofs in this book have the same structure as that of (1.5). First, one maps observed outcomes into potential outcomes. Second, one adds and subtracts the treated group's missing counterfactual outcome. Third, one invokes the parallel-trends assumption.

DID relies on a weaker assumption than treated-versus-control and before-after comparisons. DID is unbiased for the ATT under the parallel-trends assumption. Instead, a sufficient condition for the treated-versus-control cross-sectional comparison in (1.1) to be unbiased for the ATT is

$$E [Y_{s,t}(0)] = E [Y_{n,t}(0)] \text{ for all } t. \quad (1.6)$$

(1.6) requires that s and n have the same untreated-outcome *levels*, while (1.4) only requires that they have the same untreated-outcome *evolutions*. (1.6) is a stronger assumption than (1.4): if (1.6) holds then (1.4) holds, but the converse is not true. For instance, in Figure 1.1 (1.4) holds but (1.6) fails: s and n have different untreated-outcome levels.³ Similarly, a sufficient condition for the before-after comparison in (1.2) to be unbiased for the ATT is

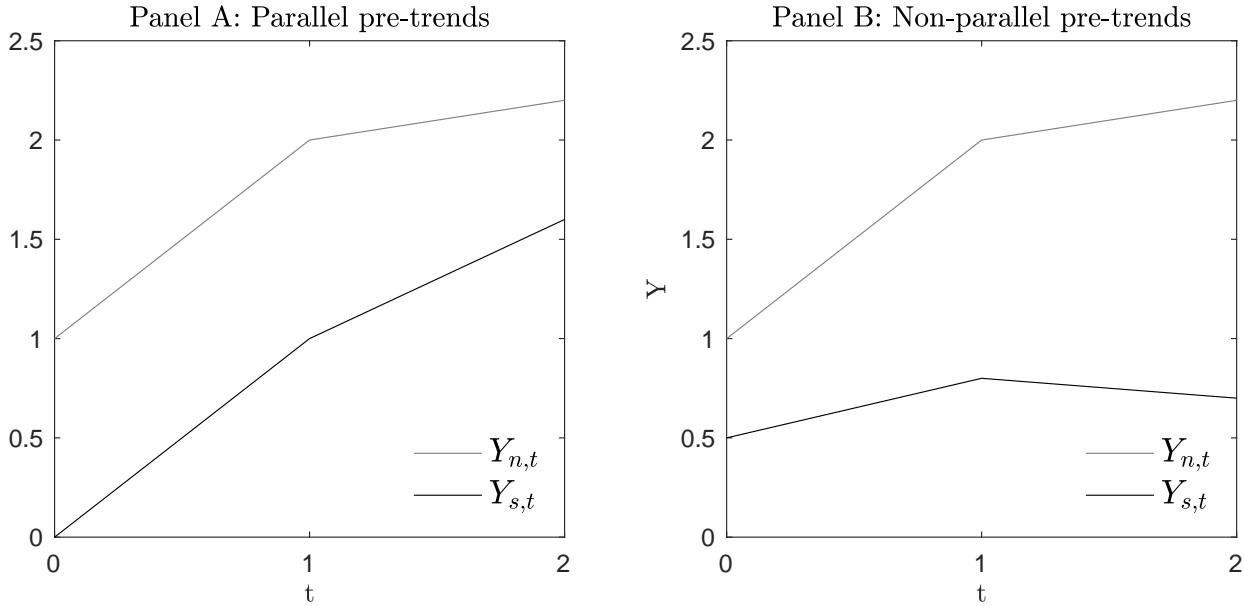
$$E [Y_{g,2}(0)] = E [Y_{g,1}(0)] \text{ for all } g. \quad (1.7)$$

(1.7) requires that groups' untreated outcome does not evolve over time: that variable should not have a trend. Instead, (1.4) requires that the trend affecting that variable be the same in the two groups. (1.7) is a stronger assumption than (1.4): if (1.7) holds then (1.4) holds, but the converse is not true. For instance, in Figure 1.1 (1.4) holds but (1.7) fails: the untreated outcomes of s and n evolve over time.

³(1.6) is sufficient but not necessary for the estimator in (1.1) to be unbiased for the ATT. The necessary and sufficient condition for that estimator to be unbiased is that (1.6) holds at $t = 2$, a condition that is neither stronger nor weaker than (1.4). However, we do not believe that (1.6) at $t = 2$ only is a substantially weaker assumption than (1.6) at $t \in \{1, 2\}$. (1.6) is testable at $t = 1$ because s and n are untreated at that date, and we believe that when s and n have different outcome levels at $t = 1$, few researchers will be willing to assume that their outcome levels would have been the same at $t = 2$ without treatment.

“Pre-trends” or “placebo” tests of the parallel-trends assumption. Still, parallel trends is not a weak assumption. To demonstrate its plausibility, researchers conduct “pre-trends” or “placebo” tests,⁴ where they assess if s and n were indeed experiencing parallel outcome trends before s got treated. In Figure 1.2 below, we consider two examples with three periods ($t = 0, 1, 2$) and our two groups s and n . A graph like the one in the left panel lends support to the parallel-trends assumption: the outcome evolutions of s and n are perfectly parallel from period 0 to 1, before s gets treated. On the other end, a graph like the one in the right panel cast doubt on the plausibility of this assumption: the outcome evolutions of s and n are clearly not parallel from period 0 to 1. (1.6) is also placebo-testable, by comparing the outcome levels of s and n before s received the treatment. However, in the empirical applications reviewed in this book, placebo tests of (1.6) are rejected much more often than those of (1.4): there are many examples where treated and control groups have different outcome levels but similar trends.

Figure 1.2: Illustrations of pre-trends



While useful, pre-trend tests have two limitations. First, while we can test for parallel trends before treatment, we cannot do so after treatment, because the untreated outcome of s

⁴In this book, we use the expressions “pre-trend tests” and “placebo tests” interchangeably

is no longer observed once s becomes treated. Even if s and n were on parallel trends before the date when the treatment started, that does not necessarily mean that without treatment, they would have remained on parallel trends after that date. Accordingly, parallel-trends tests remain suggestive. Second, a recent literature has shown that tests of parallel trends are sometimes underpowered. Those tests may fail to detect differential trends between treated and control groups that are large enough to significantly bias DID estimators.

Relaxations of the parallel-trends assumption. There are at least three instances where one may want to relax the parallel-trends assumption. First, pre-trend tests may be rejected. Second, even if pre-trend tests are not rejected, one may worry that they lack power. Third, even if pre-trend tests are not rejected and one does not worry that they lack power, one may still worry that in the absence of treatment, treated and control groups could have experienced differential trends after the date when the treatment started. Thus, an important literature, reviewed in Chapter 4, has proposed estimators relying on relaxations of the parallel-trends assumption. First, in Section 4.1 we consider DID estimators with control variables, which rely on a conditional parallel-trends assumption. Second, in Section 4.2 we consider interactive-fixed-effects, synthetic-control, and synthetic-DID estimators. Third, in Section 4.3 we consider estimators relying on a bounded-differential-trends assumption.

Historical notes. Baker, Callaway, Cunningham, Goodman-Bacon and Sant'Anna (2025) exhibit what constitutes, as of now, the earliest example where DID was used. In the 1840's, Ignaz Semmelweis was an assistant physician at the Vienna maternity clinic. At the time, maternal mortality from childbed fever was very high, especially in one of Vienna's clinics that was exclusively staffed with physicians. In Vienna's other clinic, exclusively staffed with midwives, mortality was lower. Semmelweis conjectured that the gap may be due to hygiene: doctors routinely performed autopsies before seeing to laboring mothers, without always washing "cadaverous particles" from their hands. Midwives, on the other hand, performed no autopsies. To assess whether contaminated hands caused childbed fever, Semmelweis mandated a simple protocol of handwashing using chlorinated lime. After the implementation of the hand-washing protocol, mortality sharply dropped in the doctor-staffed clinic, while mortality did not change in the midwives-staffed clinic. This DID analysis demonstrated that hand-washing was a simple yet

powerful way of preventing childbed fever (Semmelweis, 1983). Semmelweis's theory was mocked by his peers, who refused to admit that their actions were the cause of women's mortality. In 1865, the increasingly outspoken Semmelweis allegedly suffered a nervous breakdown and was committed to an asylum by his colleagues, where he died 14 days later in unclear circumstances. Another famous early DID example is John Snow's study of whether cholera is transmitted by air or water, the two leading theories in the 1850s. Snow used a change in the water supply in one district of London, namely the switch from polluted water taken from the Thames in the centre of London to a supply of cleaner water taken upriver. He showed that following that change, cholera outbreaks diminished in that district, while they did not change in neighboring districts which were still getting their water from central London. As the treated and control districts had the same air quality, his DID analysis showed that cholera is transmitted by water (Snow, 1856; Lechner, 2011).

1.1.2 Beyond the classical design

TWFE estimators are ubiquitous, and the majority of TWFE estimators are computed outside of the classical design. In a classical design, the DID estimator is equal to the treatment coefficient in a two-way fixed effects (TWFE) linear regression of $Y_{g,t}$, on group fixed effects, period fixed effects, and $D_{g,t}$, the treatment of group g at period t . Motivated by this fact, researchers have also estimated TWFE regressions in more complicated designs, with treatments that may be non-absorbing and/or non-binary, and where groups may experience several treatment changes, at different points in time. Their hope was that there as well, TWFE was giving them an estimator that only relied on a placebo-testable parallel-trends assumption. Accordingly, TWFE regressions have become a very commonly-used technique in economics. de Chaisemartin and D'Haultfœuille (2025) conducted a survey of the 100 papers with the most Google Scholar citations published by the American Economic Review from 2015 to 2019. Of those, 26 use a TWFE regression to estimate the effect of a treatment on an outcome. By comparison, 11 of those 100 papers use a randomized experiment (six use a field experiment, three use a survey experiment, and two use a laboratory experiment), and three use a regression

discontinuity design.⁵ Of the 26 papers estimating a TWFE regression, only two have a classical design with an absorbing and binary treatment, and no variation in treatment timing. TWFE regressions are also very common in political sciences: Chiu, Lan, Liu and Xu (2023) find that of all papers published from 2017 to 2022 by the American Political Science Review, the American Journal of Political Science, and the Journal of Politics, 93 estimate a TWFE regression, and only nine have a classical design. TWFE regressions are also commonly used in sociology, environmental sciences, and epidemiology.

Outside of the classical design, TWFE estimators can be misleading if the treatment effect varies across groups and over time. In Chapter 5, we show that outside of the classical design, the parallel-trends assumption is not sufficient to ensure that the TWFE estimator is unbiased for the ATT. Under parallel trends, the TWFE estimator is unbiased for a weighted sum of group-and-period specific treatment effects across all treated (g, t) cells. The weight assigned to the treatment effect of cell (g, t) is not proportional to the population of cell (g, t) , so the TWFE estimator does not estimate the ATT. Perhaps more worryingly, the TWFE estimator may weight negatively the treatment effects of some (g, t) cells. In the minimum wage example, the TWFE estimator could for instance be unbiased for three times the effect of the minimum wage in California, minus twice the effect in Michigan. If increasing the minimum wage reduces employment by 5% in California and by 10% in Michigan, then the TWFE estimator is unbiased for $3 \times -5\% - (2 \times -10\%) = +5\%$: increasing the minimum wage reduces employment in both states, but the TWFE estimator leads us to conclude that its effect is positive.

Heterogeneity-robust DID estimators. In the two following chapters, we focus on two seemingly small departures from the classical design: designs with an absorbing, binary treatment and variation in treatment timing (Chapter 6), and designs with two periods and variability in the treatment dose received by treated groups at period two (Chapter 7). In both designs, we will describe alternatives to TWFE estimators, the so-called heterogeneity-robust DID estimators. Like TWFE estimators, those estimators rely on a partly-testable parallel-trends

⁵Comparing the number of papers using an estimation method (TWFE regressions) to the number using a research design (randomized experiments or regression discontinuity) is in line with the long-held belief that the TWFE regression is the uncontroversial treatment-effect estimator in a DID research design.

assumption. But unlike TWFE estimators, if treatment effects vary across groups and/or over time, they remain unbiased for an average of (g, t) -specific effects. Finally, in Chapter 8, we will combine the insights from the two preceding chapters to propose heterogeneity-robust DID estimators in general designs, with non-absorbing and/or non-binary treatments.

Parallel-trends assumptions whose plausibility can be assessed via pre-trend tests.

Throughout the book, we try to work under parallel-trends assumptions whose plausibility can be assessed via pre-trend tests. Despite the two aforementioned limitations of those tests, we still view them as a very important part of the DID research process. While in simple designs, natural target parameters like the ATT can be identified under such parallel-trends assumptions, this is no longer true in more complex designs, in particular in those considered in Chapters 7 and 8. There, the parameters one can identify under placebo-testable parallel-trends assumptions may be quite different from those one would ideally like to learn. Accordingly, while the book’s first chapters follow the traditional approach of defining first a target parameter before introducing an identifying assumption under which that target can be estimated, in Chapters 7 and 8 we introduce our identifying assumption first, before defining the target parameters that can be estimated under that assumption. Thus, those two chapters belong more to the so-called “reverse engineering” approach to causal inference, where the researcher estimates what they can, rather than what they want. This approach is of course not the only possible one. Proposing alternative assumptions to identify other target parameters is an interesting avenue for future research, though we would caution researchers against imposing assumptions whose plausibility cannot be assessed, at least suggestively, via some kind of placebo test.

1.2 Pedagogy

Theory. As we realize that the book’s level of technicality may be intimidating to many applied readers, we have included, throughout the book, questions in blue, to help the reader make sense of notation, understand concepts, and follow derivations. Those questions do not bear on topics covered earlier in the book. Rather, they bear on what comes next, just after the question. A large body of research in cognitive sciences shows that quizzing learners on topics

the instructor has not covered yet, in the spirit of Socrates' maieutics, improves their learning more than quizzing them on topics the instructor has just covered. The “blue questions” are an essential element of the book's pedagogy. An instructor may ask those questions to students in class, giving them a couple of minutes to discuss the question with their classroom neighbors. Students reading the book together in a reading group may discuss the question for a couple of minutes, before reading the answer. Readers studying the book on their own may stop their reading for a couple of minutes to think about those questions when they encounter one.

Applications. The methods are illustrated by revisiting several empirical articles. Chapters 3 to 8 each have an empirical running example, used throughout the chapter. Stata datasets and dofiles to perform the book's empirical exercises can be obtained by typing `ssc desc cc_xd_didtextbook` and then `net get cc_xd_didtextbook` from within Stata. Each chapter includes “green questions”, asking the reader to perform an estimation on the chapter's dataset. Reproducing the estimations outlined in the green questions will facilitate the transition from theoretical understanding to practical application. Soon, R datasets and codes will also be made available. Many green questions leverage user-written Stata packages. To install them, you need to run `ssc install packagename` within Stata. Some of those packages have dependencies (other user-written packages that need to be installed to run the package), which you will also need to install.

Computation. To facilitate the adoption of the estimators we review, we mostly focus on estimators that are computed by a pre-canned software tool in Stata and/or R, and we introduce the reader to the software's syntax. Those tools are provided freely by the researchers who create them, a crucial yet under-appreciated part of the research process. As package developers, we cannot stress how time-consuming it is to create and maintain a package, and answer users' questions. As there is not always a one-to-one mapping between the authors of a paper that proposes an estimator and the authors of a user-written package, we encourage the readers of this book, whenever they use a user-written package, to cite the package, together with the paper that has proposed the estimator.

Core and non-core material. This book is admittedly much longer than what one may expect for a book covering “only” DIDs. This reflects the fact that many seemingly unrelated methods actually have a connection to DIDs (the book for instance discusses interactive-fixed effects and synthetic control methods, as well as Bartik regressions). Moreover, DID-like methods are ubiquitous, and are used in a very broad set of applications, so covering comprehensively all common use-cases is, in and of itself, a big undertaking. Still, to make the book more approachable, we have divided its material into two categories. Starred material correspond to non-core and/or technically more difficult material, which readers may skip when reading the book for the first time.

Conflict of interest. Both of us received free Stata licenses from StataCorp.

Acknowledgements. This book has been used as a reference in courses that we have taught at the OECD, the Ruhr Graduate School of Economics, INED, Sciences Po, the Berlin School of Economics, KU Leuven, Université libre de Bruxelles, UC Louvain, Université de Clermont-Ferrand, the World Bank, Bocconi University, LISER, ENSEA Abidjan, and Université de Poitiers. We are grateful to participants in these courses: their comments and questions have greatly improved the book. We are also grateful to Andrea Albanese, Bocar Ba, Damian Clarke, David Drukker, Bruno Ferman, Deepti Goel, Yagan Hazard, Germain Lankoande, David McKenzie, Matteo Pinna Pintor, and Ahmed Tritah, for their helpful comments. We also thank Romain Angotti, David Arboleda Cárcamo, Diego Ciccia, Felix Knau, Bingxue Li, Mélitine Malézieux, and Doulo Sow for exceptional research assistance. Finally, this book has been nourished by the hundreds of DID practitioners that have reached out to us over the years, with relevant questions that did not seem to be answered anywhere. We hope that this book contains the answers they were looking for.

Chapter 2

Data, notation, and assumptions

2.1 Data requirements: group-level panel data

We seek to estimate the effect of a treatment on an outcome. For that purpose, we use a panel of G groups observed at T periods, respectively indexed by g and t . Typically, groups are locations, like states, counties, or municipalities, but a group could also just be a single individual or firm. The group-level panel data may be constructed by aggregating an individual-level repeated cross-section data set at the (g, t) level, defining groups, say, as individuals' county of birth. The group-level panel data may also be constructed from a single cross-section dataset, with cohort of birth playing the role of the time variable. Thus, the data requirements to construct a group-level panel data set are fairly minimal, which probably explains the pervasiveness of the estimators described in this book. In most of this textbook, we consider estimators that are not weighted by $N_{g,t}$, the population of cell (g, t) , and we also assume that the group-level panel dataset is balanced, meaning that the outcome and treatment of each group is observed at every period. This is mostly to reduce notational complexity, but we discuss the consequences of imbalancedness and weighting at the end of Chapters 3 and 6.

2.2 Treatment and potential outcomes

Treatment. Let $D_{g,t}$ denote the treatment of group g at period t , let \mathcal{D}_t be the set of values $D_{g,t}$ can take at period t (i.e.: its support), let $\mathbf{D}_g = (D_{g,1}, \dots, D_{g,T})$ be a $1 \times T$ vector stacking the treatments of group g from period 1 to T , and let $\mathbf{D} = (\mathbf{D}_1, \dots, \mathbf{D}_G)$ be a vector stacking the treatments of all groups at every period. \mathbf{D} is referred to as the design of a study.

Potential outcomes. For all $(d_1, d_2, \dots, d_T) \in \mathcal{D}_1 \times \mathcal{D}_2 \times \dots \times \mathcal{D}_T$, let $Y_{g,t}(d_1, d_2, \dots, d_T)$ be the potential outcome of cell (g, t) if $(D_{g,1}, D_{g,2}, \dots, D_{g,T}) = (d_1, d_2, \dots, d_T)$. For instance, with a binary treatment, and if $T = 4$, $Y_{1,3}(0, 1, 1, 0)$ is the outcome of group one at period three if that group is untreated at period one, treated at periods two and three, and untreated at period four. This potential outcome model, proposed by Robins (1986), allows for *dynamic* effects of lagged treatments on the current outcome, and for *anticipation* effects of future treatments on the current outcome.

2.3 Assumptions

2.3.1 Identifying assumptions

2.3.1.1 Exclusion restrictions

We start by considering assumptions that exclude some arguments from the potential outcome function introduced above.

SUTVA The notation above implicitly assumes that g 's potential outcomes only depend on g 's treatments, not on the treatments of other groups, the so-called Stable Unit Treatment Value Assumption (SUTVA, Rubin, 1978). SUTVA may fail in the presence of spatial spillovers: a policy is targeted to some locations, but the effect of treatment spills over onto “nearby” locations. For example, individuals in control areas can travel to the treated areas and receive treatment. Few papers have attempted to relax the SUTVA assumption in DID designs, though this is a topic that has started receiving more attention recently. We will review some of this

burgeoning literature in the next chapters.

No anticipation. In most of this textbook, we also maintain a no-anticipation assumption.

Assumption NA (*No Anticipation*) *For all g and $(d_1, \dots, d_T) \in \mathcal{D}_1 \times \dots \times \mathcal{D}_T$, $Y_{g,t}(d_1, \dots, d_T) = Y_{g,t}(d_1, \dots, d_t)$.*

Assumption NA requires that a group's current outcome does not depend on its future treatments. It is plausible when treatment's introduction is hard to anticipate. It is less plausible when treatment's introduction is announced saliently ahead of time. Then, researchers sometimes redefine a (g, t) cell as treated if at period t , it has been announced that group g will get treated in the future.

Initial conditions. The potential outcome notation $Y_{g,t}(d_1, \dots, d_t)$ implicitly assumes that groups' treatment prior to period one, the first time period in the data, does not affect their outcome, the so-called “initial conditions” assumption. In many of the applications we will consider in this textbook, groups cannot be exposed to treatment before period one. Then, the “initial conditions” assumption is innocuous. When groups might have been exposed to treatment before period one, this assumption is not innocuous, though very few papers have attempted to relax it in the DID literature.

No dynamic effects. Some of our results rely on the following assumption:

Assumption ND (*No Dynamic Effects*) *For all g and $(d_1, \dots, d_t) \in \mathcal{D}_1 \times \dots \times \mathcal{D}_t$, $Y_{g,t}(d_1, \dots, d_t) = Y_{g,t}(d_t)$.*

Assumption ND requires that a group's current outcome do not depend on its past treatments, the so-called no-dynamic-effects or no-carry-over-effects hypothesis. When one studies the effect of a tax on prices, Assumption ND is plausible if prices adjust quickly, while Assumption ND is implausible if prices are sticky. Under Assumptions NA and ND and with a binary treatment, each cell (g, t) has two potential outcomes: $Y_{g,t}(0)$ if g is untreated at t , and $Y_{g,t}(1)$ if g is treated at t . Then, we are back to the standard Neyman-Rubin model of potential outcomes (Neyman et al., 1990; Rubin, 1974).

2.3.1.2 Parallel trends

In most of this textbook we will assume that groups are on parallel trends. Our baseline parallel-trends assumption is as follows. For all t , let $\mathbf{0}_t$ denote a vector of t zeros. $Y_{g,t}(\mathbf{0}_t)$ is the outcome of group g at t if it remains untreated from period 1 to t , hereafter referred to as the never-treated outcome of g at t .

Assumption PT (*Parallel trends for the never-treated outcome*) For all $t \geq 2$, $E[Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})]$ does not vary across g .

Assumption PT requires that every group experiences the same expected evolution of its never-treated potential outcome. Under Assumption ND, Assumption PT reduces to:

$$\text{For all } t \geq 2, E[Y_{g,t}(0) - Y_{g,t-1}(0)] \text{ does not vary across } g. \quad (2.1)$$

In Chapter 4, we consider various relaxations of Assumption PT.

Assumption PT is equivalent to assuming that the never-treated outcome follows a TWFE model. It follows from Assumption PT that for $t \geq 2$,

$$E[Y_{g,t}(\mathbf{0}_t) - Y_{g,1}(0)] = \sum_{t'=2}^t E[Y_{g,t'}(\mathbf{0}_{t'}) - Y_{g,t'-1}(\mathbf{0}_{t'-1})]$$

is constant across g . Then, let $\gamma_t = E[Y_{g,t}(\mathbf{0}_t) - Y_{g,1}(0)]$, and let $\alpha_g = E[Y_{g,1}(0)]$. One has that

$$E[Y_{g,t}(\mathbf{0}_t)] = E[Y_{g,1}(0)] + E[Y_{g,t}(\mathbf{0}_t) - Y_{g,1}(0)] = \alpha_g + \gamma_t. \quad (2.2)$$

Conversely, it is easy to verify that if $E[Y_{g,t}(\mathbf{0}_t)] = \alpha_g + \gamma_t$, then Assumption PT holds. Therefore, letting

$$\varepsilon_{g,t} = Y_{g,t}(\mathbf{0}_t) - E[Y_{g,t}(\mathbf{0}_t)] \quad (2.3)$$

denote the deviation of g 's never treated outcome at period t from its expectation, one has that Assumption PT holds if and only if the never-treated outcome follows a TWFE model:

$$Y_{g,t}(\mathbf{0}_t) = \alpha_g + \gamma_t + \varepsilon_{g,t}, \quad E[\varepsilon_{g,t}] = 0. \quad (2.4)$$

Parallel trends for all groups, or parallel trends on average? The parallel-trends condition in Assumption PT is strong, as it requires that all groups experience parallel trends. Actually, many of the results in this book rely on the weaker condition that on average, treated and control groups experience parallel trends. For instance, in the classical design we study in Chapter 3, with G_1 treated ($D_g = 1$) and G_0 control groups ($D_g = 0$), most results hold if

$$E \left[\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})) \right] = E \left[\frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})) \right], \quad (2.5)$$

meaning that the average expected evolution of the never-treated outcome is the same across treated and control groups.

Pre-trend tests. As we will discuss in details later, Assumptions NA and PT are partly testable, by checking that on average, treated and control groups follow parallel outcome evolutions before the treated get treated, a so-called pre-trends test.

2.3.2 Assumption for statistical inference

We now introduce the main statistical assumption underlying the inferential procedures (confidence intervals, tests...) introduced in this book.

Assumption IND (*Independent groups*) *The vectors $(Y_{g,t}(d_1, \dots, d_t))_{(d_1, \dots, d_t) \in \mathcal{D}_1 \times \dots \times \mathcal{D}_t, 1 \leq t \leq T}$ are mutually independent across g .*

Assumption IND allows groups' potential outcomes to be serially correlated over time, an important feature to account for in DID studies (Bertrand et al., 2004). Assumption IND does not make any assumption on groups' treatments, because groups' treatments are implicitly conditioned upon and treated as non-stochastic in most of this book. The next section contains a thorough discussion of Assumption IND, of the fact that treatments are conditioned upon, and of the book's perspective on statistical inference. Readers that plan to skip the book's starred sections may skip this section when reading the book for the first time.

2.4 Discussion of the book’s perspective on statistical inference*

Three common perspectives on statistical inference. In treatment-effect estimation, the goal of statistical inference is, loosely speaking, to assess if an estimator could just be the product of chance, rather than the product of the treatment’s effect. Thus, drawing inference requires introducing a random component in our modelling, and there are three common ways of doing so. First, one may take a design-based perspective, where potential outcomes are non-stochastic or conditioned upon, and randomness comes from the treatment: one assumes that the treatment, or at least some element of it like its timing, is randomly assigned. Second, one may take a model-based perspective, where the study design \mathbf{D} is non-stochastic or conditioned upon, and randomness comes from potential outcomes: one assumes that stochastic shocks affect groups’ potential outcomes. For instance, in a study where the outcome of interest is agricultural yield, stochastic weather shocks can affect the outcome. Third, one may take a sampling-based perspective, where randomness comes from the random selection of the G groups we observe from a larger population. As different samples lead to different study designs and potential outcomes, both the study design \mathbf{D} and the potential outcomes are random under that perspective.

Unavoidable thought experiments. An appeal of the design-based and sampling-based perspectives is that they lead to uncontroversial inferential procedures when the researcher effectively randomizes treatment, or when the researcher effectively draws the study sample from a larger population. Then, it is possible to compute estimators’ variances with respect to the randomization and/or sampling distribution (see, e.g., Neyman et al., 1990). By contrast, the model-based perspective always relies on a thought experiment, where one imagines that nature draws some shocks affecting potential outcomes. Then, to compute estimators’ variances and draw inference, the researcher needs to make a judgment call on the joint distribution of the shocks, and in particular on their correlations, a process that may be unpalatable to researchers who would prefer that their personal opinions do not interfere with their statistical analyses. Yet, in the context of this book, we do not view this as an important disadvantage of the model-based approach. In the natural experiments we consider, researchers do not effectively randomize treatment, so that the design-based perspective also relies on a thought experiment,

involving judgement calls by the researcher. For instance, in a thought experiment where one imagines that treatments were randomly assigned, would it make more sense to imagine that the assignment was independent across g , across t , across both dimensions, across none of these two dimensions? Similarly, in the applications we consider, researchers rarely draw their sample from a larger population. In fact their study sample often includes all the states or municipalities of a country. Then, the sampling-based perspective also relies on a thought experiment, where one imagines that the sample is drawn from an hypothetical infinite super-population.

We primarily adopt a “model-based” perspective on statistical inference. Random treatment assignment or random treatment timing have many testable implications, and we will show that, in all the empirical applications revisited in this book, these testable implications are rejected. Accordingly, we do not adopt the design-based perspective, as it seems unrealistic in many natural experiments. The two remaining perspectives on statistical inference do not greatly differ, but we favor the model-based one, for pedagogical reasons. Specifically, under that perspective groups’ treatments are non-random/conditioned upon, thus making it easy to derive the finite-sample properties of the estimators we study. Thus, to study and compare estimators, we can use simple concepts, like unbiasedness and variance, that most researchers are familiar with, and we do not necessarily need to resort to more abstract identification or asymptotic arguments. Accordingly, the design \mathbf{D} is conditioned upon in most of this book, though this conditioning is left implicit to alleviate notation. Concretely, $E[X]$ or $V[X]$ should actually be understood as $E[X|\mathbf{D}]$ or $V[X|\mathbf{D}]$. Conditioning on the design does not affect much results concerning estimators’ expectation: if $\hat{\theta}$ is conditionally unbiased for a design-dependent parameter $\theta_0(\mathbf{D})$, then by the law of iterated expectations $\hat{\theta}$ is unconditionally unbiased for $E[\theta_0(\mathbf{D})]$:

$$E[\hat{\theta}|\mathbf{D}] = \theta_0(\mathbf{D}) \Rightarrow E[\hat{\theta}] = E[\theta_0(\mathbf{D})].$$

On the other hand, conditioning on the design affects the unbiasedness of the variance estimators $\hat{V}(\hat{\theta})$ we consider: by the law of total variance, if a variance estimator is unbiased for an estimator’s conditional variance, it is downward biased for the estimator’s unconditional variance:

$$E[\hat{V}(\hat{\theta})|\mathbf{D}] = V[\hat{\theta}|\mathbf{D}] \Rightarrow E[\hat{V}(\hat{\theta})] = E[V[\hat{\theta}|\mathbf{D}]] \leq V[\hat{\theta}]. \quad (2.6)$$

The conditional-variance estimators we consider remain valid measures of estimators' variability: they just do not account for the variability that would arise from modifying the design.

Statistically independent groups. As mentioned earlier, statistical inference in natural experiments, and in observational research more generally, needs to rely on an assumption, namely a judgement call, on the model's random components. The main statistical assumption underlying our inferences is Assumption IND. While it allows groups' potential outcomes to be serially correlated over time, it requires that potential outcomes of different groups be independent. This may sound like a strong assumption, that rules out common statistical shocks affecting several groups. Actually, common shocks are not necessarily ruled out by Assumption IND, they are just implicitly conditioned upon: groups' potential outcomes are assumed independent conditional on common shocks. For instance, with US county-level panel data, one may have that

$$Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1}) = M_{s(g),t} + m_{g,t}, \quad (2.7)$$

where $M_{s(g),t}$ is a “macro” stochastic shock common to all counties in the same state as g , while $m_{g,t}$ is a “micro” stochastic shock specific to county g , assumed to be mean zero and mean-independent of $M_{s(g),t}$. Then, Assumption IND cannot hold unconditionally, as potential outcomes of different counties in the same state are necessarily correlated. But it can hold conditional on the macro-shocks $M_{s(g),t}$. The weaker version of the parallel-trends assumption in (2.5) also holds conditional on the state-level shocks if

$$\frac{1}{G_1} \sum_{g:D_g=1} M_{s(g),t} = \frac{1}{G_0} \sum_{g:D_g=0} M_{s(g),t}, \quad (2.8)$$

meaning that the state-level shocks are on average the same across treated and control counties.¹

At what level should one cluster standard errors? Since the work of Bertrand, Duflo and Mullainathan (2004), it has become common practice to make the independence assumption

¹If one supposes that Assumption IND and (2.5) hold conditional on the macro shocks, all the analysis is conditional on those shocks. In particular, the treatment effects defined below depend on those shocks: as the two previous displays show, different state-level shocks may lead to different potential outcomes, and to different treatment effects. Those conditional effects remain valid measures of the treatment's effect, they just depend on the specific shocks that took place over the study period (Deeb and de Chaisemartin, 2019).

at the level at which the treatment is assigned, and use standard errors clustered at that level. For instance, with county-level data but a treatment assigned at the state level, one clusters at the state level. This approach is uncontroversial in the design-based approach to inference (Abadie, Athey, Imbens and Wooldridge, 2023). It is also uncontroversial in the sampling-based approach, which is often used in the DID literature (Abadie, 2005; Callaway and Sant'Anna, 2021). Counties' treatments are not conditioned upon in this approach and counties in the same state mechanically have the same treatment, so that correlation has to be accounted for by clustering at the state level. But in the model-based approach we adopt, where treatments are conditioned upon, a state-level treatment assignment does not warrant clustering at the state level. Similarly, correlated shocks across counties belonging to the same state do not warrant clustering at the state level, if the parallel-trends assumption holds conditional on those shocks, as in (2.8). Clustering standard errors at the state level leads to a slightly weaker parallel-trends assumption than clustering at the county level. With standard errors clustered at the state level, (2.5) has to hold unconditionally. Under (2.7), this means that the following condition should hold:

$$\frac{1}{G_1} \sum_{g:D_g=1} E(M_{s(g),t}) = \frac{1}{G_0} \sum_{g:D_g=0} E(M_{s(g),t}), \quad (2.9)$$

which is slightly weaker than (2.8). On the other hand, clustering at the state level can lead to non-trivial power losses,² both to estimate the treatment's effect, but also to conduct pre-trend tests of the parallel-trends assumption. Then, there will be situations where both (2.8) and (2.9) fail, but a pre-trends test of (2.9), with standard errors clustered at the state level, is not rejected due to a lack of power, while a pre-trends test of (2.8) with standard errors clustered at the county level is rejected. In such situations, a researcher who clusters at the state level will mistakenly assume that they have parallel trends and proceed with their DID analysis. Instead, a researcher who clusters at the county level will know that they do not have parallel trends, which will lead them to adjust their model, say by controlling for some covariates or allowing for bounded differential trends, as discussed in Chapter 4 below. To conclude, we believe that due

²One can show that variance estimators clustered at the county level are unbiased for estimators' variances conditional on the state-level shocks, while variance estimators clustered at the state level are unbiased for their unconditional variances. Conditional variances are always weakly smaller, in expectation, than unconditional ones (see, e.g., (2.6)).

to the panel structure of the data, allowing for serial correlation across potential outcomes over time is important, so using clustered standard errors is in order. Then, we recommend clustering either at the level at which the treatment is assigned, following Bertrand et al. (2004), or at the most disaggregated level at which one can construct a panel dataset, following the arguments above.

We sometimes adopt the sampling-based perspective, when doing so simplifies the exposition. Sometimes, assuming that the G groups we observe are an independent and identically distributed (i.i.d.) sample from an infinite super population greatly simplifies the exposition. For instance, in Chapter 7 the treatment might be a continuously distributed variable, taking as many values as there are treated groups. Then, to help us think about a continuous distribution of the treatment, it is useful to introduce an infinite super-population of groups. Therefore, while we favor the model-based perspective in most of the book, we sometimes adopt the sampling-based one. Going back and forth between these two conceptual frameworks imposes a cognitive cost on the reader. At the same time, strengthening one's translation skills between those two languages can be useful, because both are used in the methodological papers on DID. When we adopt the sampling-based perspective, we replace Assumption IND by the slightly stronger requirement that groups' potential outcomes and treatments are i.i.d.:

Assumption IID (i.i.d. groups) *The vectors $((Y_{g,t}(d_1, \dots, d_t))_{(d_1, \dots, d_t) \in \mathcal{D}_1 \times \dots \times \mathcal{D}_t}, D_{g,t})_{1 \leq t \leq T}$ are i.i.d.*

As groups are i.i.d., the g subscript can be dropped. In the sampling-based perspective, expectations are taken with respect to both the distribution of groups' potential outcomes and treatments. To highlight the difference with the conditional expectations introduced above, we let $E_u[\cdot]$ denote such unconditional expectations. Finally, in the sampling-based perspective, and thus under Assumption IID, one can show that (2.5) implies

$$E_u [Y_t(\mathbf{0}_t) - Y_{t-1}(\mathbf{0}_{t-1}) | D = 1] = E_u [Y_t(\mathbf{0}_t) - Y_{t-1}(\mathbf{0}_{t-1}) | D = 0], \quad (2.10)$$

a common way of stating the parallel-trends assumption (see, e.g., Abadie, 2005).

Part II

The classical design

Chapter 3

The classical design

Designs with an absorbing and binary treatment, and no variation in treatment timing. Throughout this chapter, we assume that the design is a classical DID design: the treatment is binary, absorbing, meaning that groups cannot switch out of treatment, and there is no variation in treatment timing.

Design CLA (*Classical DID design*) $D_{g,t} = 1\{t > T_0\}D_g$, with $T_0 \geq 1$, $D_g \in \{0, 1\}$ for all g , and $\min_g D_g = 0$ and $\max_g D_g = 1$.

D_g is an indicator equal to 1 for treatment groups, that all become treated at period $T_0 + 1$ and remain treated thereafter, while D_g is equal to 0 for control groups that never become treated. We require that $T_0 \geq 1$, meaning that there is at least one pre-treatment period, and that there is at least one treatment and one control group. Unlike the conditions in say, Assumption NA, those in Design CLA only involve observed quantities, and one can directly verify from the data whether those conditions are satisfied or not. Let $G_1 = \sum_{g=1}^G D_g$ and $G_0 = G - G_1$ respectively denote the number of treated and control groups, let $T_1 = T - T_0$ denote the number of treated periods, and let $N_1 = G_1 T_1$ denote the number of treated (g, t) cells.

Chapter's running example: the effect of compulsory patent licensing on US innovation. On October 6, 1917, US Congress passed the Trading with the Enemy Act (TWEA), whereby US firms were allowed to violate enemy-owned patents if they contributed to the war effort. Over the next couple of years, the TWEA became more and more punitive, and in February 1919, German-owned patents were systematically licensed to US firms. At the time of the

TWEA, the US organic chemical industry was lagging behind Germany's. Moser and Voena (2012) study the effect of the compulsory licensing of German chemical patents on US domestic invention. For that purpose, they use a balanced panel of the number of patents granted by the US Patent and Trademark Office per year, for each of the 7 248 organic-chemistry subclasses. D_g is an indicator for 336 subclasses that received at least one German patent under the TWEA. The treatment is $D_{g,t} = 1\{t > T_0\}D_g$, with $T_0 + 1 = 1919$.

Dataset used in this chapter. To answer the green questions in this chapter, you need to use the `moser_voena_didtextbook` dataset, which contains the following variables:

- `subclass`: a subclass identifier;
- `year`: a year identifier;
- `treatmentgroup`: an indicator for the 336 treated subclasses;
- `post`: an indicator for all years after 1919;
- `yearpost`: the interaction of `year` and `post`;
- `twea`: the treatment indicator $D_{g,t}$;
- `reftimeplus1`, `reftimeplus2`, ..., `reftimeplus21`: indicators for having been exposed to treatment for 1, 2, ..., 21 years ($D_g = 1, t = T_0 + \ell$ for $\ell \in \{1, \dots, 21\}$);
- `retimeminus1`, `retimeminus2`, ..., `retimeminus18`: indicators for being 1, 2, ..., 18 years before 1918, the year before treatment adoption ($D_g = 1, t = T_0 + \ell$ for $\ell \in \{-1, \dots, -18\}$);
- `patents`: the number of patents in subclass g and year t ;
- `diffpatentswrt1918`: the change in the number of patents in subclass g from 1918 to t ;
- `patents1900`: the number of patents in subclass g in 1900 (that variable will be used in Chapter 4).

While the original data of Moser and Voena (2012) is a 1875 to 1939 panel, `moser_voena_didtextbook` starts in 1900, thus yielding a lighter dataset on which some of the commands below take less time to run.

3.1 Target parameters

(g, t) -specific effects. For all t , let $\mathbf{1}_t$ denote a vector of t ones. For all (g, t) such that $D_g = 1$ and $t > T_0$, let

$$\text{TE}_{g,t} = E [Y_{g,t} - Y_{g,t}(\mathbf{0}_t)] = E [Y_{g,t}(\mathbf{0}_{T_0}, \mathbf{1}_{t-T_0}) - Y_{g,t}(\mathbf{0}_t)]$$

denote the expected effect in cell (g, t) of having been treated rather than untreated for $t - T_0$ periods, from $T_0 + 1$ to t . If we impose Assumption ND, thus ruling out dynamic effects, $\text{TE}_{g,t}$ reduces to

$$\text{TE}_{g,t} = E [Y_{g,t}(1) - Y_{g,t}(0)].$$

In the set-up we consider, $\text{TE}_{g,t}$ applies to only one group, and can therefore not be consistently estimated, an issue we return to in Section 3.6 below. Instead, researchers often consider effects aggregated across all treated groups, that can be consistently estimated.

Average treatment effect on the treated. A first commonly-studied aggregated parameter is

$$\text{ATT} = \frac{1}{G_1 T_1} \sum_{(g,t): D_{g,t}=1} \text{TE}_{g,t},$$

the average effect of having been treated rather than untreated for $t - T_0$ periods, across all treated (g, t) cells. If we rule out dynamic effects, ATT reduces to

$$\text{ATT} = \frac{1}{G_1 T_1} \sum_{(g,t): D_{g,t}=1} E [Y_{g,t}(1) - Y_{g,t}(0)],$$

the standard average treatment effect on the treated (ATT).

Average effect of having been treated for ℓ periods. For any $\ell \in \{1, \dots, T_1\}$, let

$$\text{ATT}_\ell = \frac{1}{G_1} \sum_{g: D_g=1} E [Y_{g,T_0+\ell}(\mathbf{0}_{T_0}, \mathbf{1}_\ell) - Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell})].$$

ATT_ℓ is the average effect of having been treated for ℓ periods, across all treated groups and at period $T_0 + \ell$. One has

$$\text{ATT} = \frac{1}{T_1} \sum_{\ell=1}^{T_1} \text{ATT}_\ell,$$

so the ATT is just the average of the average effects of having been treated for one, two, ..., and T_1 periods.

Can we use $\ell \mapsto \text{ATT}_\ell$ to learn how treatment effects vary with length of exposure? In many contexts, it would be interesting to learn if treatment effects vary with length of exposure. Assume that the treatment is costly, and the treatment cost has to be paid at each period of exposure. Then, if treatment effects do not increase beyond, say, three periods of exposure, it may be optimal to stop the treatment after three periods. Unfortunately, $\ell \mapsto \text{ATT}_\ell$ cannot be used to learn how the treatment effect varies with length of exposure, unless one is willing to assume that the treatment effect does not vary with calendar time. For instance, if $\text{ATT}_2 > \text{ATT}_1 > 0$, can we conclude that two periods of exposure to treatment have a larger effect than one period of exposure?

$\text{ATT}_2 = \frac{1}{G_1} \sum_{g:D_g=1} E [Y_{g,T_0+2}(\mathbf{0}_{T_0}, 1, 1) - Y_{g,T_0+2}(\mathbf{0}_{T_0+2})]$ is the effect of two periods of exposure at period $T_0 + 2$, while $\text{ATT}_1 = \frac{1}{G_1} \sum_{g:D_g=1} E [Y_{g,T_0+1}(\mathbf{0}_{T_0}, 1) - Y_{g,T_0+1}(\mathbf{0}_{T_0+1})]$ is the effect of one period of exposure at period $T_0 + 1$. Then, if $\text{ATT}_2 > \text{ATT}_1 > 0$, this difference could be due to the fact that being treated for two periods has a larger effect than being treated for one period, but it could also be due to the fact that the effect of one period of exposure is larger at period $T_0 + 2$ than at period $T_0 + 1$. It is only if one is ready to assume that the effect of one period of exposure is the same at periods $T_0 + 1$ and $T_0 + 2$ that $\text{ATT}_2 > \text{ATT}_1 > 0$ implies that two periods of exposure has a larger effect than one period of exposure. In this chapter's appendix, we show this point more formally.

3.2 Two-Way Fixed Effects estimators

3.2.1 Static Two-Way Fixed Effects estimator

Let $\hat{\beta}^{\text{fe}}$ denote the sample coefficient on $D_{g,t}$, the treatment in group g at period t , in an OLS regression of $Y_{g,t}$, the outcome of group g at period t , on group fixed effects (FEs), period FEs, and $D_{g,t}$:

$$Y_{g,t} = \sum_{g'=1}^G \hat{\alpha}_{g'} 1\{g = g'\} + \sum_{t'=1}^T \hat{\gamma}_{t'} 1\{t = t'\} + \hat{\beta}^{\text{fe}} D_{g,t} + \hat{\epsilon}_{g,t}, \quad (3.1)$$

where $\hat{\epsilon}_{g,t}$ denotes the regression residual.¹ Hereafter, $\hat{\beta}^{\text{fe}}$ is referred to as the Two-Way Fixed Effects (TWFE) estimator.

Application to the compulsory licensing example. Using the `moser_voena_didtextbook` dataset, run the TWFE regression of the number of patents in subclass g and year t on subclass and year FEs and the `twea` treatment, clustering standard errors at the subclass level.² You can either use the `xtreg` or the `reghdfe` command. According to this regression, does compulsory licensing have an effect on US innovation?

```
xtreg patents tweas i.year, fe i(subclass) robust cluster(subclass)
reghdfe patents tweas, absorb(subclass year) cluster(subclass)
```

The coefficient on the `twea` treatment is equal to 0.288, and it is statistically significant, so according to this regression compulsory licensing has a positive effect on US innovation. In the treatment group and in the post period, the average number of patents per subclass is

¹Formally, (3.1) is a definition of the regression residual $\hat{\epsilon}_{g,t}$. Throughout this textbook, regression equations such as (3.1) should be understood as definitions of regression residuals. With a slight abuse of notation, all residuals are denoted $\hat{\epsilon}_{g,t}$.

²In this application, we do not use the degrees-of-freedom-adjusted standard errors discussed in Section 3.3.1 below. As the number of treated and control groups is large in this application, these adjustments leave standard errors almost unchanged.

equal to 0.724. According to the TWFE regression, this average would have been equal to $0.724 - 0.288 = 0.436$ without treatment. Thus, the treatment increased innovation in the treated subclasses by more than two thirds. Our 0.288 TWFE coefficient is very close to the 0.255 coefficient in Table 2 Column (2) Moser and Voena (2012). The difference comes from the fact they use a 1875 to 1939 panel, while we only use years 1900 to 1939.

In Design CLA, $\hat{\beta}^{\text{fe}}$ is a simple DID estimator. In designs with an absorbing and binary treatment and no variation in treatment timing, one can show that $\hat{\beta}^{\text{fe}}$ is equal to the coefficient on $D_{g,t}$ in a simpler regression of $Y_{g,t}$ on an intercept, the treatment-group indicator D_g , the post-treatment indicator $1\{t > T_0\}$, and $D_{g,t}$, the interaction of the two indicators. Therefore, $\hat{\beta}^{\text{fe}}$ is a simple DID estimator comparing the average outcome's evolution, before and strictly after T_0 , in treatment and control groups:

$$\begin{aligned}\hat{\beta}^{\text{fe}} = & \frac{1}{G_1 T_1} \sum_{g:D_g=1, t>T_0} Y_{g,t} - \frac{1}{G_1 T_0} \sum_{g:D_g=1, t\leq T_0} Y_{g,t} \\ & - \left(\frac{1}{G_0 T_1} \sum_{g:D_g=0, t>T_0} Y_{g,t} - \frac{1}{G_0 T_0} \sum_{g:D_g=0, t\leq T_0} Y_{g,t} \right).\end{aligned}\quad (3.2)$$

Proof of (3.2).* By Frisch-Waugh-Lovell's theorem (see the appendix of this chapter for a restatement of this theorem), $\hat{\beta}^{\text{fe}}$ is the coefficient on the regression of $Y_{g,t}$ on the residual from the regression of $D_{g,t}$ on group and time FEs. Since $D_{g,t}$ only depends on D_g and $1\{t > T_0\}$, this latter regression yields the same residuals as that of $D_{g,t}$ on D_g and $1\{t > T_0\}$. Thus, by applying Frisch-Waugh-Lovell's theorem in the other direction, $\hat{\beta}^{\text{fe}}$ is also the coefficient on $D_{g,t}$ in a simpler regression of $Y_{g,t}$ on an intercept, the treatment-group indicator D_g , the post-treatment indicator $1\{t > T_0\}$, and $D_{g,t}$, the interaction of the two indicators. This regression is saturated in $(D_g, 1\{t > T_0\})$ (see Section 3.1.4 of Angrist and Pischke, 2009, for a definition of a saturated regression), so the regression function is equal to the conditional mean function of $Y_{g,t}$ given $(D_g, 1\{t > T_0\})$ (see Theorem 3.1.4 of Angrist and Pischke, 2009), and all coefficients are functions of that conditional mean function. In particular, it follows from the ninth unnumbered equation on page 50 of Angrist and Pischke (2009) that $\hat{\beta}^{\text{fe}}$ satisfies (3.2) **QED**.

Application to the compulsory licensing example. Using the `moser_voena_didtextbook` dataset, verify that $\hat{\beta}^{\text{fe}}$ is equal to the coefficient on the `twea` treatment in a regression of the

number of patents in subclass g and year t on the treatment group indicator, the post-treatment indicator, and the `twea` treatment.

```
reg patents treatmentgroup post tweas, cluster(subclass)
```

The coefficient on the `twea` treatment is identical to that in the previous regressions. The standard error is very slightly different, due to differences in degrees-of-freedom adjustment, as the regression above has much fewer explanatory variables than the TWFE regression. As a further exercise, one may also verify that (3.2) holds, by computing in Stata or R the DID in the right hand side of (3.2), to verify that it is equal to $\hat{\beta}^{\text{fe}}$.

$\hat{\beta}^{\text{fe}}$ is unbiased for the ATT.

Theorem 1 *In Design CLA, if Assumptions NA and PT hold,*

$$E[\hat{\beta}^{\text{fe}}] = \text{ATT}. \quad (3.3)$$

Theorem 1 shows that in Design CLA, $\hat{\beta}^{\text{fe}}$ is unbiased for ATT.

Proof of Theorem 1

$$E[Y_{g,t}] = E[Y_{g,t}(\mathbf{0}_t)] + D_{g,t}E[Y_{g,t} - Y_{g,t}(\mathbf{0}_t)] = E[Y_{g,t}(\mathbf{0}_t)] + D_{g,t}\text{TE}_{g,t}. \quad (3.4)$$

The second equality simply follows from the definition of $\text{TE}_{g,t}$. The first equality is an example of definitional equalities that relate observed and potential outcomes, that are commonly used in the causal inference literature. Justify it, by showing that it holds for any value that $D_{g,t}$ can take.

If $D_{g,t} = 1$, the right hand side is equal to $E[Y_{g,t}(\mathbf{0}_t)] + E[Y_{g,t} - Y_{g,t}(\mathbf{0}_t)] = E[Y_{g,t}]$ so the equality holds. If $D_{g,t} = 0$, the left hand side is equal to $E[Y_{g,t}(\mathbf{0}_t)]$, because $Y_{g,t} = Y_{g,t}(\mathbf{0}_t)$ if

$D_{g,t} = 0$: in Design CLA, if a (g, t) cell is untreated its observed outcome is its never-treated outcome because it has never been treated yet. If $D_{g,t} = 0$, the right hand side is also equal to $E[Y_{g,t}(\mathbf{0}_t)]$ so the equality holds. The equality is true when $D_{g,t} = 1$ and when $D_{g,t} = 0$, and $D_{g,t}$ cannot take another value. Therefore, the equality is always true.

Then,

$$\begin{aligned}
& E[\widehat{\beta}^{\text{fe}}] \\
&= \frac{1}{G_1 T_1} \sum_{g:D_g=1, t>T_0} E[Y_{g,t}] - \frac{1}{G_1 T_0} \sum_{g:D_g=1, t\leq T_0} E[Y_{g,t}] - \frac{1}{G_0 T_1} \sum_{g:D_g=0, t>T_0} E[Y_{g,t}] + \frac{1}{G_0 T_0} \sum_{g:D_g=0, t\leq T_0} E[Y_{g,t}] \\
&= \frac{1}{G_1 T_1} \sum_{g:D_g=1, t>T_0} (E[Y_{g,t}(\mathbf{0}_t)] + \text{TE}_{g,t}) - \frac{1}{G_1 T_0} \sum_{g:D_g=1, t\leq T_0} E[Y_{g,t}(\mathbf{0}_t)] \\
&\quad - \frac{1}{G_0 T_1} \sum_{g:D_g=0, t>T_0} E[Y_{g,t}(\mathbf{0}_t)] + \frac{1}{G_0 T_0} \sum_{g:D_g=0, t\leq T_0} E[Y_{g,t}(\mathbf{0}_t)] \\
&= \text{ATT} \\
&\quad + \frac{1}{G_1 T_1} \sum_{g:D_g=1, t>T_0} (\alpha_g + \gamma_t) - \frac{1}{G_1 T_0} \sum_{g:D_g=1, t\leq T_0} (\alpha_g + \gamma_t) - \frac{1}{G_0 T_1} \sum_{g:D_g=0, t>T_0} (\alpha_g + \gamma_t) + \frac{1}{G_0 T_0} \sum_{g:D_g=0, t\leq T_0} (\alpha_g + \gamma_t) \\
&= \text{ATT}.
\end{aligned}$$

Justify each step of this derivation.

The first equality follows from (3.2), the fact the design is conditioned upon, and the linearity of the expectation operator. The second equality follows from (3.4) and the fact that in Design CLA $D_{g,t} = 0$ if $D_g = 0$ or $t \leq T_0$. The third equality follows from the definition of ATT and (2.2). The last equality follows from the following four facts. First, the average of α_g in the first and second summations cancel each other. Second, the average of α_g in the third and fourth summations cancel each other. Third, the average of γ_t in the first and third summations cancel each other. Fourth, the average of γ_t in the second and fourth summations cancel each other. **QED.**

3.2.1.1 Assuming randomized treatment rather than parallel trends?

If treatment is randomly assigned, there are more efficient treatment-effect estimators than $\hat{\beta}^{\text{fe}}$. Instead of assuming parallel trends, one could make the stronger assumption that D_g is randomly assigned, so that treated and control units are representative of each other. Under this assumption, $\hat{\beta}^{\text{fe}}$ is still unbiased for ATT, and in fact for the average treatment effect, but there are more efficient estimators. If $T = 2$, all units are untreated at period 1, and some are randomly assigned to treatment at period 2, results in Frison and Pocock (1992) and McKenzie (2012) imply that $\hat{\beta}^{\text{fe}}$ can have a higher variance than the treatment coefficient in a regression of $Y_{g,2}$ on an intercept and $D_{g,2}$, controlling for $Y_{g,1}$.

Randomly-assigned treatment is a strong assumption, which should be thoroughly tested. In natural experiments where researchers do not effectively randomize treatment, a randomization claim should be substantiated with thorough and, ideally, pre-specified balancing checks showing that groups' treatment status does not predict covariates correlated with the outcome. For instance, if D_g is randomly assigned, one should have

$$D_g \perp\!\!\!\perp (Y_{g,1}(\mathbf{0}_t), \dots, Y_{g,T}(\mathbf{0}_T)),$$

where “ $\perp\!\!\!\perp$ ” denotes independence of random variables. For all $t \leq T_0$, $Y_{g,t} = Y_{g,t}(\mathbf{0}_t)$: groups' outcome is equal to their untreated outcome till period T_0 . Then, the previous display implies that

$$D_g \perp\!\!\!\perp (Y_{g,1}, \dots, Y_{g,T_0}), \quad (3.5)$$

an equation that only involves observed variables, and can therefore be tested. To test (3.5), one can run a pooled regression of $Y_{g,t}$ on D_g , for all $t \leq T_0$. However, this test is not consistent against all violations of the independence condition in (3.5). For instance, the test will fail to reject if $Y_{g,t}$ is positively correlated to D_g for some t s, negatively correlated to D_g for other t s, and those positive and negative correlations offset each other. Alternatively, one can run T_0 regressions of $Y_{g,t}$ on D_g , for all $t \leq T_0$. However, one needs to account for multiple testing when computing the T_0 p-values of the coefficients on D_g , which may lead to a low-power test. To avoid this issue, one can regress D_g on $(Y_{g,1}, \dots, Y_{g,T_0})$, and run an F-test that all coefficients are equal to zero.

A test consistent against all violations of (3.5).* Even a regression of D_g on $(Y_{g,1}, \dots, Y_{g,T_0})$ is still not consistent against all violations of (3.5). For instance, this test could fail to reject if $Y_{g,1} \times Y_{g,2}$ increases the probability that $D_g = 1$. To obtain a test consistent against all violations of (3.5), one can use a permutation test based on the Kolmogorov-Smirnov test-statistic:

$$KS_G = \sup_{(d,y_1,\dots,y_{T_0}) \in \mathcal{S}} \sqrt{G} \left| \frac{1}{G} \sum_{g=1}^G \mathbb{1} \{D_g \leq d, Y_{g,1} \leq y_1, \dots, Y_{g,T_0} \leq y_{T_0}\} \right. \\ \left. - \left(\frac{1}{G} \sum_{g=1}^G \mathbb{1} \{D_g \leq d\} \right) \left(\frac{1}{G} \sum_{g=1}^G \mathbb{1} \{Y_{g,1} \leq y_1, \dots, Y_{g,T_0} \leq y_{T_0}\} \right) \right|,$$

where \mathcal{S} denotes the values taken by $(D_g, Y_{g,1}, \dots, Y_{g,T_0})$. The test compares KS_G to the $(1 - \alpha)$ -quantile of its permuted version:

$$KS_G^{\Pi} = \sup_{(d,y_1,\dots,y_{T_0}) \in \mathcal{S}} \sqrt{G} \left| \frac{1}{G} \sum_{g=1}^G \mathbb{1} \{D_{\Pi(g)} \leq d, Y_{g,1} \leq y_1, \dots, Y_{g,T_0} \leq y_{T_0}\} \right. \\ \left. - \left(\frac{1}{G} \sum_{g=1}^G \mathbb{1} \{D_{\Pi(g)} \leq d\} \right) \left(\frac{1}{G} \sum_{g=1}^G \mathbb{1} \{Y_{g,1} \leq y_1, \dots, Y_{g,T_0} \leq y_{T_0}\} \right) \right|,$$

where Π is a random permutation over $\{1, \dots, G\}$, with uniform distribution over the $G!$ possible permutations. This test has exact size, and unlike the three aforementioned tests it is consistent against all alternatives (see, e.g. van der Vaart and Wellner, 2023). On the other hand, it is less straightforward and computationally more costly to implement than those three tests.

Application to the compulsory licensing example. Using the `moser_voena_didtextbook` dataset, regress the number of patents in subclass g and year t on the treatment group indicator, restricting the sample to years strictly before 1919.

```
reg patents treatmentgroup if year<=1918, cluster(subclass)
```

Before the treatment, subclasses in the treatment group produced 0.136 patents per year, while subclasses in the control group produced 0.314 patents, namely more than twice as many. Thus, the coefficient on the `treatmentgroup` indicator in the regression is equal to -0.178 , and it is highly significant. Treatment is not as good as randomly assigned in this application, even using the simplest and presumably least powerful test of random assignment discussed above.

3.2.2 Event-study TWFE estimators

In Design CLA, alongside (3.1) researchers often also estimate the following regression:

$$Y_{g,t} = \hat{\alpha}_0 + \hat{\alpha}_1 D_g + \sum_{t'=1, t' \neq T_0}^T \hat{\gamma}_{t'} 1\{t = t'\} + \sum_{\ell=-(T_0-1), \ell \neq 0}^{T_1} \hat{\beta}_\ell^{\text{fe}} 1\{t = T_0 + \ell\} D_g + \hat{\epsilon}_{g,t}. \quad (3.6)$$

If $\ell > 0$, $1\{t = T_0 + \ell\} D_g$ is an indicator equal to 1 if at t , group g has been treated for ℓ periods. If $\ell < 0$, $1\{t = T_0 + \ell\} D_g$ is an indicator equal to 1 if at t , group g will be treated in $-\ell+1$ periods. (3.6) is often referred to as a TWFE event-study (ES) regression, and researchers often plot its coefficients $(\hat{\beta}_\ell^{\text{fe}})_{\ell \neq 0}$ on a so-called ES graph, with $\ell = t - T_0$, the relative time to treatment onset for the treated groups,³ on the x -axis. The regression has $2T$ explanatory variables, which is the number of values that the vector of its explanatory variables $(D_g, (1\{t = t'\})_{t' \in \{1, \dots, T\}})$ can take. Therefore, the regression is saturated in $(D_g, (1\{t = t'\})_{t' \in \{1, \dots, T\}})$ (see Section 3.1.4 of Angrist and Pischke, 2009, for a refresher on saturated regressions). Then, its predicted values are just equal to the conditional mean function of $Y_{g,t}$ given $(D_g, (1\{t = t'\})_{t' \in \{1, \dots, T\}})$ (see Theorem 3.1.4 of Angrist and Pischke, 2009), and all its coefficients are functions of that conditional mean function. In particular, one can show that for all $\ell \neq 0$,

$$\hat{\beta}_\ell^{\text{fe}} = \frac{1}{G_1} \sum_{g: D_g=1} (Y_{g,T_0+\ell} - Y_{g,T_0}) - \frac{1}{G_0} \sum_{g: D_g=0} (Y_{g,T_0+\ell} - Y_{g,T_0}), \quad (3.7)$$

a simple DID comparing the T_0 to $T_0 + \ell$ outcome evolution in treatment and control groups. As period T_0 is the omitted time period in (3.6), all DIDs are relative to that period. Accordingly, for $\ell \geq 1$, the DIDs in (3.7) consider evolutions from the past to the future, whereas for $\ell \leq -1$, those DIDs consider evolutions from the future to the past.

3.2.2.1 Estimating event-study effects

Theorem 2 *In Design CLA, if Assumptions NA and PT hold, then for all $\ell \in \{1, \dots, T_1\}$,*

$$E \left[\hat{\beta}_\ell^{\text{fe}} \right] = ATT_\ell. \quad (3.8)$$

³Researchers usually rather define relative time as $(t - (T_0 + 1))$, with treatment onset corresponding to relative time 0 rather than to relative time 1. We prefer to define relative time as $t - T_0$, to ensure that the ES graph is normalized at 0, and that estimated effects and pre-trends are shown symmetrically around 0.

Proof of Theorem 2. For $\ell \in \{1, \dots, T_1\}$,

$$\begin{aligned}
& E \left[\widehat{\beta}_\ell^{\text{fe}} \right] \\
&= \frac{1}{G_1} \sum_{g: D_g=1} E [Y_{g,T_0+\ell} - Y_{g,T_0}] - \frac{1}{G_0} \sum_{g: D_g=0} E [Y_{g,T_0+\ell} - Y_{g,T_0}] \\
&= \frac{1}{G_1} \sum_{g: D_g=1} E [Y_{g,T_0+\ell}(\mathbf{0}_{T_0}, \mathbf{1}_\ell) - Y_{g,T_0}(\mathbf{0}_{T_0})] - \frac{1}{G_0} \sum_{g: D_g=0} E [Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0}(\mathbf{0}_{T_0})] \\
&= \frac{1}{G_1} \sum_{g: D_g=1} E [Y_{g,T_0+\ell}(\mathbf{0}_{T_0}, \mathbf{1}_\ell) - Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell})] \\
&\quad + \frac{1}{G_1} \sum_{g: D_g=1} E [Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0}(\mathbf{0}_{T_0})] \\
&\quad - \frac{1}{G_0} \sum_{g: D_g=0} E [Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0}(\mathbf{0}_{T_0})] \\
&= \text{ATT}_\ell. \tag{3.9}
\end{aligned}$$

Justify each step of this derivation.

The first equality follows from (3.7) and the fact the design is conditioned upon, the second equality follows from Design CLA, the third equality follows adding and subtracting $Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell})$, the fourth equality follows from Assumption PT and the definition of ATT_ℓ . This proves (3.8) **QED.**

Estimation under a weaker parallel-trends assumption. The parallel-trends condition in Assumption PT is strong, as it requires that all groups experience parallel trends. Actually, and as mentioned in Chapter 2, Theorem 1, Theorem 2, and Theorem 3 below still hold under the weaker “average parallel-trends” assumption in (2.5).

3.2.2.2 Pre-trend tests

Theorem 3 In Design CLA, if Assumptions NA and PT hold, and $T_0 \geq 2$, then for all $\ell \in \{-1, \dots, -(T_0 - 1)\}$,

$$E \left[\widehat{\beta}_\ell^{\text{fe}} \right] = 0. \tag{3.10}$$

Proof of Theorem 3. If $T_0 \geq 2$, for $\ell \in \{-1, \dots, -(T_0 - 1)\}$,

$$\begin{aligned} & E \left[\hat{\beta}_\ell^{\text{fe}} \right] \\ &= \frac{1}{G_1} \sum_{g:D_g=1} E [Y_{g,T_0+\ell} - Y_{g,T_0}] - \frac{1}{G_0} \sum_{g:D_g=0} E [Y_{g,T_0+\ell} - Y_{g,T_0}] \\ &= \frac{1}{G_1} \sum_{g:D_g=1} E [Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0}(\mathbf{0}_{T_0})] - \frac{1}{G_0} \sum_{g:D_g=0} E [Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0}(\mathbf{0}_{T_0})] \\ &= 0. \end{aligned}$$

The first equality follows from (3.7) and the fact the design is conditioned upon, the second equality follows from Design CLA, the third equality follows from Assumption PT. This proves (3.10) **QED**.

Based on Theorem 3, how can one test Assumptions NA and PT?

Assumptions NA and PT have a testable implication. (3.10) shows that if $T_0 \geq 2$, Assumptions NA and PT have a testable implication: they imply that $E \left[\hat{\beta}_\ell^{\text{fe}} \right] = 0$ for all $\ell \in \{-1, \dots, -(T_0 - 1)\}$. Therefore, if we can reject the null that $E \left[\hat{\beta}_\ell^{\text{fe}} \right] = 0$ for all $\ell \in \{-1, \dots, -(T_0 - 1)\}$, we can reject the null that Assumptions NA and PT both hold. Pre-trend tests have important limitations, discussed later in this chapter. Yet, we believe that in research making use of observational data to study cause-and-effect relationships, such partial or “placebo” tests are an essential component of the research process, to help establish the credibility of the identifying assumption one has to make to recover the missing counterfactual outcome (Imbens et al., 2001; Imbens and Xu, 2024). Not all identifying assumptions we consider in this book are partly testable. The fact that Assumptions NA and PT are partly testable is an appealing feature of those assumptions.

If $T_0 \geq 3$, $E \left[\hat{\beta}_\ell^{\text{fe}} \right] = 0$ for all $\ell \in \{-1, \dots, -(T_0 - 1)\}$ is a null hypothesis on a vector of dimension $T_0 - 1 > 1$. Then, would testing separately that $E \left[\hat{\beta}_\ell^{\text{fe}} \right] = 0$ for all $\ell \in \{-1, \dots, -(T_0 - 1)\}$ yield a valid test of Assumptions NA and PT? If not, what is the issue with this testing procedure?

Doing pre-trend tests correctly. This testing procedure does not account for multiple hypothesis testing. To account for it, one can run an F-test that $E[\widehat{\beta}_\ell^{\text{fe}}] = 0$ for all $\ell \in \{-1, \dots, -(T_0 - 1)\}$. Alternatively, one can adjust p-values for multiple testing, using, say, a Bonferroni adjustment. Related to, but different from, the Bonferroni adjustment, one can use the “Sup-t” test proposed by Montiel Olea and Plagborg-Møller (2019), where $\max_{\ell \in \{-1, \dots, -(T_0 - 1)\}} |\widehat{\beta}_\ell^{\text{fe}}/\widehat{\sigma}_\ell|$, the largest pre-trend t-statistic, is compared to the quantiles of the max of $T_0 - 1$ normal variables with mean 0 and variance equal to the estimated variance of $(\widehat{\beta}_\ell^{\text{fe}}/\widehat{\sigma}_\ell)_{\ell \in \{-1, \dots, -(T_0 - 1)\}}$. The F- and Sup-t test both have asymptotically nominal size under the null. Neither test is universally more powerful than the other. Schematically, the Sup-t test will be more powerful if $E[\widehat{\beta}_\ell^{\text{fe}}]$ is far from zero for one ℓ but equal to zero for all other ℓ , while the F-test will be more powerful if $E[\widehat{\beta}_\ell^{\text{fe}}]$ is never very far from zero but slightly different from 0 for several ℓ . The “Sup-t” test can be computed using the `sotable` Stata command (Drukker, 2023).

How many pre-trend estimators should one show? When T_0 is large, one can potentially compute many pre-trend estimators, thus assessing if treatment and control groups experienced parallel trends over a long period. In such cases, one may find that while treated and control groups were experiencing parallel trends a few periods before treatment, their trends are not parallel anymore when one moves further back into the past. Should this be a cause of concern? In other words, how many insignificant pre-trend estimators should one show to convincingly demonstrate parallel pre-trends? While very natural, this question has received little attention, so we do not have a good answer, backed by a sound theory. Yet, we would like to suggest the following, simple rule: researchers should not compute much fewer pre-trends than event-study estimators. Consider two researchers. The first one can only compute three pre-trend estimators, because $T_0 = 4$. The second one can compute more than three pre-trend estimators ($T_0 > 4$), but while they find that treated and control groups experience roughly parallel trends for three periods before treatment, their trends become markedly different more than three periods before treatment. In our opinion, both researchers should not report more than three event-study estimators. The rationale for this recommendation is simple: an estimator of ATT_ℓ relies on a parallel-trends assumption from T_0 to $T_0 + \ell$. To support this identifying assumption, the researcher should show that parallel trends holds from T_0 to $T_0 - \ell$. With our recommendation,

large and significant pre-trend estimators just before treatment are more problematic than large and significant pre-trend estimators long before treatment, which makes intuitive sense. When a researcher computes many more event-study than pre-trend estimators, their readers and reviewers should factor the fact that the parallel-trends assumption underlying some of their long-run event-study estimators was not placebo tested.

First-difference pre-trend estimators. For $\ell \in \{-1, \dots, -(T_0 - 1)\}$, let

$$\hat{\beta}_{\ell}^{\text{fd}} = \frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,T_0+\ell} - Y_{g,T_0+\ell+1}) - \frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,T_0+\ell} - Y_{g,T_0+\ell+1}) \quad (3.11)$$

be another pre-trend estimator, that should also not significantly differ from zero under Assumptions NA and PT. $\hat{\beta}_{\ell}^{\text{fe}}$ is a “long-difference” pre-trend estimator, that compares treatment- and control-groups’ outcome evolution over several periods before the treatment onset. $\hat{\beta}_{\ell}^{\text{fd}}$ instead is a “first-difference” pre-trend estimator, that compares treatment- and control-groups’ outcome evolution over consecutive periods before the treatment onset. Accordingly,

$$\hat{\beta}_{\ell}^{\text{fe}} = \sum_{k=\ell}^{-1} \hat{\beta}_k^{\text{fd}}. \quad (3.12)$$

As first- and long-difference pre-trend estimators are linearly dependent, F-tests that $E[\hat{\beta}_{\ell}^{\text{fe}}] = 0$ for all $\ell \in \{-1, \dots, -(T_0 - 1)\}$ and that $E[\hat{\beta}_{\ell}^{\text{fd}}] = 0$ for all $\ell \in \{-1, \dots, -(T_0 - 1)\}$ are equal, so using one or the other will always yield the same results.

How to interpret pre-trend coefficients? $\hat{\beta}_{-\ell}^{\text{fe}}$ is the difference between the average outcome evolutions of treated and control groups from T_0 to $T_0 - \ell$. As this difference goes from the future to the past, $\hat{\beta}_{-\ell}^{\text{fe}} > 0$ means that before the treatment, the outcome was increasing less (or decreasing more) in the treatment than in the control group. Under the assumption that without treatment, the outcome would have kept increasing less in the treatment than in the control group after the treatment date, then $\hat{\beta}_{\ell}^{\text{fe}}$ is a downward biased estimator of the effect of ℓ periods of exposure to treatment. Conversely, if $\hat{\beta}_{-\ell}^{\text{fe}} < 0$ and one is ready to assume that the same differential trends observed prior to the treatment date would have continued after the treatment date, then $\hat{\beta}_{\ell}^{\text{fe}}$ is an upward biased estimator of the treatment’s effect. Applied researchers often use pre-trend estimators to assess if differential trends are likely to bias upward or downward their event-study estimators.

How to interpret commonly-found patterns of pre-trend coefficients? Figure 3.1 below displays three common patterns of non-zero pre-trend coefficients, that lead to different interpretations. First, one may have that $\hat{\beta}_\ell^{\text{fe}}$ is significantly different from zero for all $\ell < 0$, but $\ell \mapsto \hat{\beta}_\ell^{\text{fe}}$ is approximately constant, as in Panel A. Then, it follows from (3.12) that $\hat{\beta}_{-1}^{\text{fd}} \neq 0$ but $\hat{\beta}_\ell^{\text{fd}} = 0$ for all $\ell < -2$: treated and control groups are on parallel-trends throughout the pre-treatment period, except from $T_0 - 1$ to T_0 . This suggests that Assumption PT holds, but Assumption NA fails: at period T_0 , treatment groups are already affected by their upcoming treatment in the next period. Then, one may just recompute the estimators defined above, redefining the date when the treatment started as T_0 or as the date when the treatment was announced. Second, one may have that $\hat{\beta}_\ell^{\text{fe}} \approx \ell\theta$ for some real number θ , as in Panel B. This suggests that Assumption PT fails due to differential linear trends between the treated and control groups. Then, one may allow for linear trends in the estimation, see Section 4.1.3.3 for further details. Third, one may have that $\hat{\beta}_{-1}^{\text{fe}}$ is significantly different from zero but $\hat{\beta}_\ell^{\text{fe}} \approx 0$ for all $\ell \leq -2$, as in Panel C. Then, it follows from (3.12) that $\hat{\beta}_{-1}^{\text{fd}} \approx -\hat{\beta}_{-2}^{\text{fd}} \neq 0$, and $\hat{\beta}_\ell^{\text{fd}} \approx 0$ for all $\ell \leq -3$. Therefore, such pre-trend coefficients cannot be interpreted as evidence of anticipation effects arising at period T_0 , and there is no obvious alternative estimation method one can suggest in the face of such pre-trend estimators.

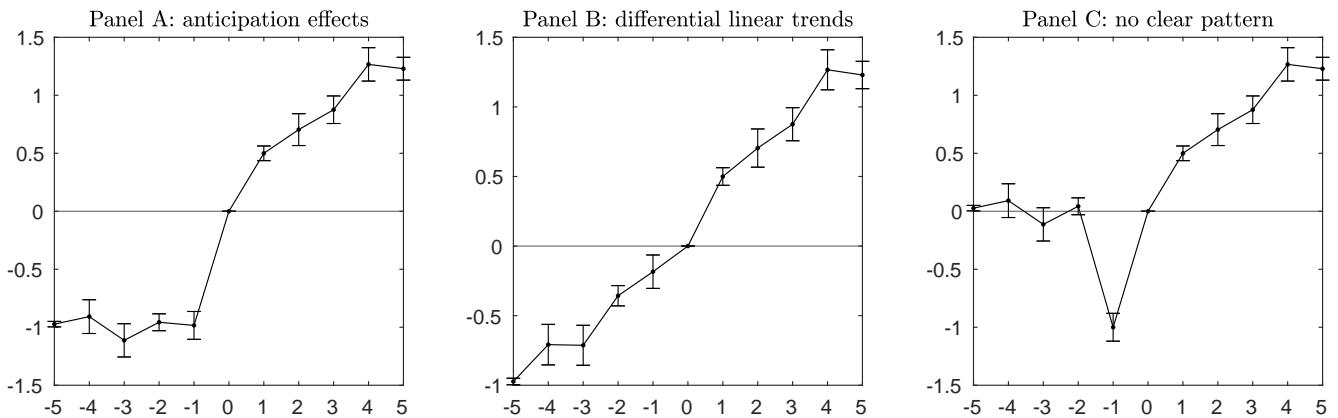


Figure 3.1: Three patterns of pre-trends

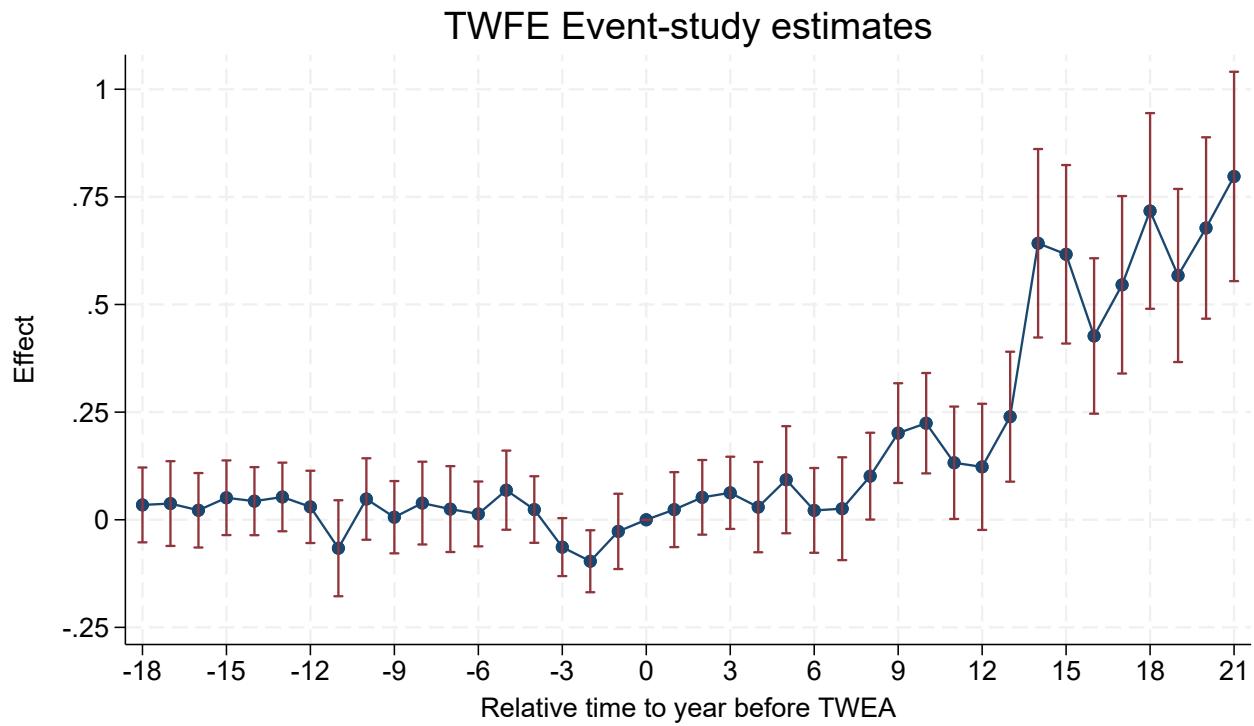
3.2.2.3 Application to the compulsory licensing example

Using the `moser_voena_didtextbook` dataset, estimate an event-study TWFE regression of patents on a treatment group FE, year FEs, indicators for having been exposed to treatment for 1, 2, ..., 21 years, and indicators for being a treatment group subclass 1, 2, ..., 18 years before 1918, clustering standard errors at the subclass level. According to this regression, does compulsory licensing have an effect on US innovation? Do Assumptions NA and PT seem to hold?

```
reg patents i.year treatmentgroup reltimeminus* reltimeplus*, cluster(subclass)
```

Figure 3.2 below shows the regression's coefficients. According to this regression, compulsory licensing does not have short-term effects on US innovation: the coefficients $\hat{\beta}_\ell^{\text{fe}}$ are small and insignificant for $\ell \in \{1, \dots, 7\}$. After eight years, effects become positive and significant, and after 14 years effects become very large. Prior to 1919, patenting was low in the treated subclasses. Then, as argued by Moser and Voena (2012), US firms in those subclasses had to bridge a large gap to the technological frontier before they could patent their own inventions. Moreover, patent grants typically occur two to three years after application. All this could explain why the effects of compulsory licensing take time to emerge. The pre-trends estimates are small, and substantially smaller than the estimated mid- and long-run effects of compulsory licensing. Only one out of 18 pre-trend estimates is individually significant at the 5% level. However, an F-test that all pre-trends coefficients are equal to zero is rejected ($p\text{-value} < 0.001$), thus suggesting modest differential trends between treated and control subclasses.

Figure 3.2: Effects of compulsory licensing on US innovation, according to a TWFE ES regression



Note: This figure shows the estimated effects of compulsory licensing on patents, as well as pre-trends estimates, using years 1900 to 1939 of the data from Moser and Voena (2012), and the TWFE event-study regression in (3.6). Standard errors are clustered at the patent subclass level. 95% confidence intervals are shown in red.

Verify that (3.7) holds for $\ell = 1$, by computing in Stata the DID in the right hand side of that equation.

```
sum patents if year==1919&treatmentgroup==1
scalar m1=r(mean)
sum patents if year==1918&treatmentgroup==1
scalar m2=r(mean)
```

```

sum patents if year==1919&treatmentgroup==0
scalar m3=r(mean)
sum patents if year==1918&treatmentgroup==0
scalar m4=r(mean)
di m1-m2-(m3-m4)

```

The result, 0.0235202 is identical, up to rounding error, to $\hat{\beta}_1^{\text{fe}}$.

3.3 Inference on the ATT and on the event-study effects

Depending on how large the sample is, there are several possible approaches to test hypotheses on, or build confidence intervals for, the event-study effects $(\text{ATT}_\ell)_{\ell \in \{1, \dots, T_1\}}$ and the ATT. With many treated and control groups, standard asymptotic inference is possible. If there are few treated groups but many control groups, we cannot consistently estimate the $(\text{ATT}_\ell)_{\ell \in \{1, \dots, T_1\}}$ and the ATT, but we can still perform valid inference, without making parametric assumptions. With few treated and control groups, inference needs to rely either on parametric assumptions or on homogeneity conditions. Finally, we discuss how to determine which of the three aforementioned approaches may be more appropriate and reliable in a specific application. Note that we mostly focus on confidence intervals (CIs), but hypotheses tests can be obtained similarly. Note also that our review of the literature on inference with few treated and/or few control groups is not exhaustive, though we hope it provides a useful starting point for practitioners.

3.3.1 Many treated and control groups

Estimating the variance of $\hat{\beta}_\ell^{\text{fe}}$. Under Assumption IND, groups are independent, so it directly follows from (3.7) that

$$V(\hat{\beta}_\ell^{\text{fe}}) = \frac{1}{G_1^2} \sum_{g:D_g=1} V(Y_{g,T_0+\ell} - Y_{g,T_0}) + \frac{1}{G_0^2} \sum_{g:D_g=0} V(Y_{g,T_0+\ell} - Y_{g,T_0}). \quad (3.13)$$

Then, we consider the following variance estimator:

$$\hat{\sigma}_\ell^2 = \frac{1}{G_1} \hat{\sigma}_{\ell,1}^2 + \frac{1}{G_0} \hat{\sigma}_{\ell,0}^2,$$

where, for $d \in \{0, 1\}$,

$$\hat{\sigma}_{\ell,d}^2 = \frac{1}{G_d - 1} \sum_{g:D_g=d} \left(Y_{g,T_0+\ell} - Y_{g,T_0} - \frac{1}{G_d} \sum_{g':D_{g'}=d} Y_{g',T_0+\ell} - Y_{g',T_0} \right)^2$$

is the sample variance of the T_0 to $T_0 + \ell$ outcome evolution, among groups with $D_g = d$. We can obtain $\hat{\sigma}_\ell^2$ with the option `vce(hc2 group)` of the Stata command `regress` (starting with version 18 of Stata) and with the R command `dfadjustSE` (Kolesar, 2023).⁴ For all g in the control group, under Assumption PT,

$$E[Y_{g,T_0+\ell} - Y_{g,T_0}] = E[Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0}(\mathbf{0}_{T_0})] = \gamma_{T_0+\ell} - \gamma_{T_0},$$

which does not vary with g . Then, $\hat{\sigma}_{\ell,0}^2$ is just the sample variance of independent variables that all have the same expectation, so it follows from standard arguments that

$$E[\hat{\sigma}_{\ell,0}^2] = \frac{1}{G_0} \sum_{g:D_g=0} V(Y_{g,T_0+\ell} - Y_{g,T_0}). \quad (3.14)$$

On the other hand, for all g in the treatment group,

$$E[Y_{g,T_0+\ell} - Y_{g,T_0}] = \text{TE}_{g,T_0+\ell} + \gamma_{T_0+\ell} - \gamma_{T_0},$$

which varies with g if treatment effects are heterogeneous. Then,

$$E[\hat{\sigma}_{\ell,1}^2] = \frac{1}{G_1} \sum_{g:D_g=1} V(Y_{g,T_0+\ell} - Y_{g,T_0}) + \frac{1}{G_1 - 1} \sum_{g:D_g=1} (\text{TE}_{g,T_0+\ell} - \text{ATT}_\ell)^2 \quad (3.15)$$

$$\geq \frac{1}{G_1} \sum_{g:D_g=1} V(Y_{g,T_0+\ell} - Y_{g,T_0}), \quad (3.16)$$

where the first equality is proven in this chapter's appendix. Combined with (3.14) and (3.13), this implies that $E[\hat{\sigma}_\ell^2] \geq V(\hat{\beta}_\ell^{\text{fe}})$. If treatment effects are homogenous, the inequality in (3.16) becomes an equality.

Asymptotically conservative CIs for ATT_ℓ . When both G_0 and G_1 tend to infinity, (3.7) and the central limit theorem imply that

$$\frac{\hat{\beta}_\ell^{\text{fe}} - \text{ATT}_\ell}{V(\hat{\beta}_\ell^{\text{fe}})^{1/2}} \xrightarrow{d} \mathcal{N}(0, 1).$$

⁴With the `robust` option of the Stata command `regress`, the denominator of $\hat{\sigma}_{\ell,d}^2$ is replaced by G_d . Both estimators are equivalent asymptotically.

Then, as $\widehat{\sigma}_\ell^2$ overestimates $V(\widehat{\beta}_\ell^{\text{fe}})$, one can show that the CI $[\widehat{\beta}_\ell^{\text{fe}} \pm z_{1-\alpha/2} \widehat{\sigma}_\ell]$, where $z_{1-\alpha/2}$ denotes the quantile of order $1 - \alpha/2$ of a standard normal distribution, is asymptotically conservative: it includes ATT_ℓ with probability tending to at least $1 - \alpha$. If treatment effects are homogenous, this CI is exact. In practice, researchers estimate ATT_ℓ for more than one value of ℓ . Then, one can show that a multivariate version of the asymptotic normality result in the previous display holds, and use that result to derive a joint test that $\text{ATT}_\ell = 0$ for all ℓ , or jointly valid confidence intervals for the effects ATT_ℓ .

Inference on the ATT. Comparing (3.2) and (3.7), one can see that $\widehat{\beta}^{\text{fe}}$ has the same structure as $\widehat{\beta}_\ell^{\text{fe}}$, with $Y_{g,T_0+\ell}$ replaced by $(1/T_1) \sum_{t > T_0} Y_{g,t}$ and Y_{g,T_0} replaced by $(1/T_0) \sum_{t \leq T_0} Y_{g,t}$. Hence, one can follow the same steps as above to estimate the variance of $\widehat{\beta}^{\text{fe}}$ and construct an asymptotically conservative CI for the ATT.

Finite-sample adjustment. The CI above uses normal quantiles, neglecting the randomness in the estimation of $V(\widehat{\beta}_\ell^{\text{fe}})$. Bell and McCaffrey (2002) suggest using instead the quantiles of a t -distribution, with degrees-of-freedom (DOF)

$$K := \frac{(G_0 + G_1)^2(G_0 - 1)(G_1 - 1)}{G_0^2(G_0 - 1) + G_1^2(G_1 - 1)}.$$

The rationale behind this choice is that if $\sigma_{\ell,0}^2 = \sigma_{\ell,1}^2$, with $\sigma_{\ell,d}^2 := E[\widehat{\sigma}_{\ell,d}^2]$, then $K\widehat{\sigma}_\ell^2/V(\widehat{\beta}_\ell^{\text{fe}})$ has the same first two moments as a $\chi^2(K)$ distribution. Imbens and Kolesar (2016) show that this modification can yield CIs with better coverage in small samples. The corresponding p-values and CIs can be obtained with the option `vce(hc2 group, dfadjust)` of the Stata command `regress` (starting with Stata 18) and the R command `dfadjustSE`.

3.3.2 When are G_0 and G_1 large enough to rely on the CIs in Section 3.3.1? A simulation study

The question we ask in this section is a difficult one, which we try to answer with simulations. Our simulations are inspired from those in Tables 1 and 2 of Imbens and Kolesar (2016), but with different sample sizes, and different errors' distributions.

Motivating our simulation designs. Remember that

$$\varepsilon_{g,t} = Y_{g,t}(\mathbf{0}_t) - E(Y_{g,t}(\mathbf{0}_t))$$

denotes the deviation of g 's never treated outcome at period t from its expectation. Then, let

$$\begin{aligned} \eta_{g,\ell} &= \varepsilon_{g,T_0+\ell} - \varepsilon_{g,T_0} \\ &= Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0}(\mathbf{0}_{T_0}) - E(Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0}(\mathbf{0}_{T_0})). \end{aligned} \quad (3.17)$$

With this notation, using the same steps as those used to obtain (3.9), and using the fact that under Assumption PT $E(Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0}(\mathbf{0}_{T_0}))$ does not depend on g , one can show that

$$\widehat{\beta}_\ell^{\text{fe}} = \frac{1}{G_1} \sum_{g:D_g=1} Y_{g,T_0+\ell}(\mathbf{0}_{T_0}, \mathbf{1}_\ell) - Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) + \frac{1}{G_1} \sum_{g:D_g=1} \eta_{g,\ell} - \frac{1}{G_0} \sum_{g:D_g=0} \eta_{g,\ell}. \quad (3.18)$$

Under the sharp null of no treatment effect $(Y_{g,T_0+\ell}(\mathbf{0}_{T_0}, \mathbf{1}_\ell) - Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell})) = 0$ for all (g, ℓ)), the previous display simplifies to

$$\widehat{\beta}_\ell^{\text{fe}} = \frac{1}{G_1} \sum_{g:D_g=1} \eta_{g,\ell} - \frac{1}{G_0} \sum_{g:D_g=0} \eta_{g,\ell}.$$

If the treatment has no effect, our TWFE ES estimators just compare the average of the errors $\eta_{g,\ell}$ in the treatment and in the control group, so their distribution depends on G_0 , G_1 , the distribution of the errors, and how that distribution differs in the treatment and in the control groups. For our simulations, it is enough to consider one event-study coefficient, so we drop the ℓ subscript in the remainder of this section.

Data Generating Process. We consider three distributions for the errors $(\eta_g)_{g=1,\dots,G}$:

1. $\eta_g | D_g \sim N(0, \sigma^2(D_g))$.
2. If $D_g = 1$ (resp. $D_g = 0$), $\eta_g / \sigma(1)$ (resp. $-\eta_g / \sigma(0)$) follows a recentered log-normal distribution $LN(0, 1)$.
3. η_g is drawn from the empirical distribution of the $(\widehat{\eta}_{g,14})_{g=1,\dots,G}$ in Moser and Voena (2012), where we let

$$\widehat{\eta}_{g,\ell} = Y_{g,T_0+\ell} - Y_{g,T_0} - \frac{1}{G_0} \sum_{g':D_{g'}=0} (Y_{g',T_0+\ell} - Y_{g',T_0}) \quad (3.19)$$

for control groups, and

$$\hat{\eta}_{g,\ell} = Y_{g,T_0+\ell} - Y_{g,T_0} - \frac{1}{G_1} \sum_{g':D_{g'}=1} (Y_{g',T_0+\ell} - Y_{g',T_0}) \quad (3.20)$$

for treatment groups.

Data Generating Process (DGP) 1 follows that in Table 1 of Imbens and Kolesar (2016). DGP2 is similar to that in their Table 2, where they assume that conditional on D_g , $\eta_g/\sigma(D_g)$ follows a recentered log-normal distribution. However, our DGP should be more adversarial than theirs: assuming that $-\eta_g/\sigma(0)$ rather than $\eta_g/\sigma(0)$ follows a log-normal in the control group preserves the asymmetry of the log-normal distribution in $\hat{\beta}^{\text{fe}} = \bar{\eta}_1 - \bar{\eta}_0$. DGP3 is based on the dataset of Moser and Voena (2012), with errors drawn from the empirical distribution of the estimated errors $(\hat{\eta}_{g,14})_{g=1,\dots,G}$ in that application. We choose $\ell = 14$, because large treatment effects start emerging after 14 years of exposure to the TWEA, but results are similar for other values of ℓ . For each model, we consider both an homoscedastic and an heteroscedastic error distribution. In Models 1 and 2, $\sigma(1) = 1$, and either $\sigma(0) = 1/2$ (heteroscedastic errors) or $\sigma(0) = 1$ (homoscedastic errors).⁵ For Model 3, in the homoscedastic case we randomly allocate groups to the control or treatment, whereas in the heteroscedastic case we draw the control group's errors from the empirical distribution of $(\hat{\eta}_{g,14})_{g:D_g=0}$ in Moser and Voena (2012), and we draw the treatment group's errors from the empirical distribution of $(\hat{\eta}_{g,14})_{g:D_g=1}$. Finally, we consider eight sample sizes: $G_1 \in \{5, 10, 20, 40\}$, and $G_0 = G_1$ or $G_0 = 4G_1$.

CIs. We assess the coverage of the following 95%-level CIs:

1. “HC2-BM” uses the HC2 variance estimator $\hat{\sigma}_\ell^2$, and the critical value from a t-distribution with the degrees-of-freedom adjustement of Bell and McCaffrey (2002);
2. “HC2- ∞ ” uses the HC2 variance estimator $\hat{\sigma}_\ell^2$, and the critical value from a normal distribution;
3. “HC3” uses the HC3 variance estimator recommended for instance by MacKinnon, Nielsen and Webb (2023), and the critical value from a t-distribution with $G_0 + G_1 - 2$ degrees-of-freedom;

⁵With $G_1 \leq G_0$, coverage rates increase with $\sigma(0)$, so coverage rates with $\sigma(0) > 1$ are higher than those obtained in Table 3.1 below.

4. “EHW” uses the standard robust Eicker-Huber-White variance estimator, and the critical value from a normal distribution.

Results. Results in Table 3.1 below show that the HC2-BM CI often has a coverage rate closer to the 95% nominal value than the other CIs. Yet, with the log-normal DGP, it can still exhibit non-trivial size distortions, even with relatively large sample sizes. For instance, with $(G_0, G_1) = (80, 20)$ its coverage rate is still only equal to 88%. The size distortions we find for the HC2-BM CI are larger than those in Table 2 of Imbens and Kolesar (2016). This is partly due to the fact that we consider a more adversarial DGP, and partly due to the fact that for the specific sample size they consider ($(G_0, G_1) = (3, 27)$), the coverage of the HC2-BM CI is particularly good. In less adversarial DGPs, the HC2-BM CI has close-to-nominal coverage with as few as five treated and control groups.

DGP	G_1	G_0	Heteroskedasticity				Homoskedasticity			
			HC2-BM	HC2- ∞	HC3	EHW	HC2-BM	HC2- ∞	HC3	EHW
1	5	5	0.941	0.903	0.957	0.872	0.95	0.914	0.964	0.882
	5	20	0.939	0.89	0.926	0.86	0.955	0.91	0.94	0.886
	10	10	0.946	0.93	0.956	0.915	0.949	0.934	0.959	0.921
	10	40	0.943	0.922	0.938	0.909	0.951	0.931	0.947	0.92
	20	20	0.948	0.94	0.953	0.934	0.948	0.941	0.954	0.935
	20	80	0.947	0.938	0.946	0.932	0.95	0.94	0.948	0.935
	40	40	0.948	0.944	0.951	0.941	0.951	0.947	0.953	0.944
	40	160	0.947	0.943	0.947	0.94	0.95	0.946	0.95	0.943
2	5	5	0.832	0.794	0.834	0.766	0.842	0.805	0.843	0.777
	5	20	0.84	0.789	0.805	0.766	0.875	0.825	0.841	0.806
	10	10	0.859	0.843	0.86	0.829	0.867	0.85	0.868	0.838
	10	40	0.852	0.832	0.839	0.821	0.885	0.864	0.872	0.855
	20	20	0.885	0.878	0.886	0.871	0.892	0.884	0.893	0.878
	20	80	0.88	0.87	0.874	0.865	0.9	0.89	0.894	0.886
	40	40	0.907	0.903	0.908	0.9	0.913	0.909	0.914	0.906
	40	160	0.903	0.898	0.901	0.896	0.915	0.91	0.912	0.908
3	5	5	0.953	0.905	0.959	0.872	0.967	0.943	0.975	0.908
	5	20	0.927	0.874	0.896	0.852	0.979	0.931	0.952	0.91
	10	10	0.954	0.936	0.959	0.923	0.97	0.953	0.978	0.939
	10	40	0.932	0.91	0.921	0.898	0.959	0.937	0.947	0.927
	20	20	0.954	0.946	0.959	0.939	0.961	0.954	0.967	0.947
	20	80	0.941	0.932	0.938	0.926	0.952	0.942	0.948	0.937
	40	40	0.954	0.95	0.957	0.948	0.958	0.955	0.961	0.952
	40	160	0.949	0.944	0.947	0.941	0.95	0.946	0.949	0.943

Table 3.1: Coverage rate for various DGPs and confidence intervals

Recommendations.

1. Like Imbens and Kolesar (2016), we recommend that researchers use HC2 standard errors, with the DOF adjustment of Bell and McCaffrey (2002). The only instance where they may not follow this recommendation is if G_0 and G_1 are very

large: then HC2 standard errors with critical values from a normal distribution will also be reliable, and somewhat surprisingly, the implementation in standard statistical software of the correction of Bell and McCaffrey (2002) can be computationally costly.

2. **When both G_0 and G_1 are larger than 40, researchers may use HC2-BM CIs, without assessing the validity of those CIs via simulations.** This recommendation is based on the fact that when $\min(G_0, G_1) \geq 40$, the coverage of HC2-BM CIs is always larger than 0.90 in our simulations, so those CIs never massively undercover. Of course, we do not claim that their coverage rate could not be lower for other distributions of the $(\eta_g)_{g=1,\dots,G}$. But this log-normal DGP seems adversarial enough to suggest that our rule of thumb should work well in many cases. Importantly, this recommendation only applies to unweighted TWFE regressions: as we will discuss later, with weighting HC2-BM CIs may require larger numbers of treated and control groups to be reliable.

3. **When either G_0 or G_1 is lower than 40, researchers may run simulations tailored to their application to assess if HC2-BM CIs have satisfactory coverage in their data.** Our DGP 1 and 3 show that HC2-BM CIs can still have very good coverage, even with much fewer than 40 treated or control groups. Then, before resorting to the inference methods described in the next section, researchers should conduct simulations, to assess if in their data, the coverage of HC2-BM CIs is closer to that in our Models 1 and 3, or closer to that in our Model 2. To conduct those simulations, they may follow the method we used to conduct simulations based on the data from Moser and Voena (2012) (see also Ferman, 2019, for related proposals).

3.3.3 What can researchers do when HC2-BM CIs seem unreliable?*

Our answer to that question depends on whether G_1 is low but G_0 is large, or G_1 and G_0 are both low. As a word of caution on what follows, note that this area of research is still active and no consensus has emerged on it yet.

3.3.3.1 Few treated groups but many control groups

Inference problem. There are many applications where G_1 , the number of treated groups, is low. For instance, a well-studied US schooling merit aid program is the HOPE scholarship. It was implemented in Georgia only, so in a state-level analysis, $G_1 = 1$. We saw earlier that

$$\widehat{\beta}_\ell^{\text{fe}} = \frac{1}{G_1} \sum_{g:D_g=1} Y_{g,T_0+\ell}(1) - Y_{g,T_0+\ell}(0) + \frac{1}{G_1} \sum_{g:D_g=1} \eta_{g,\ell} - \frac{1}{G_0} \sum_{g:D_g=0} \eta_{g,\ell}.$$

If $G_0 \rightarrow \infty$ but G_1 remains finite, the third average in the previous display converges to zero by the law of large numbers, but the first two averages do not and they remain random. Therefore, $\widehat{\beta}_\ell^{\text{fe}}$ is not consistent anymore.

Approach assuming identically distributed disturbances. Conley and Taber (2011) propose a method to draw valid inference on ATT_ℓ , under two conditions. First, they assume that treatment effects are not random:

$$Y_{g,T_0+\ell}(1) - Y_{g,T_0+\ell}(0) = \text{TE}_{g,T_0+\ell}. \quad (3.21)$$

Does (3.21) still allow for heterogeneous treatment effects?

Yes it does, as $\text{TE}_{g,T_0+\ell}$ may vary with g and $T_0 + \ell$. Plugging (3.21) into the previous equation for $\widehat{\beta}_\ell^{\text{fe}}$,

$$\begin{aligned} \widehat{\beta}_\ell^{\text{fe}} &= \text{ATT}_\ell + \frac{1}{G_1} \sum_{g:D_g=1} \eta_{g,\ell} - \frac{1}{G_0} \sum_{g:D_g=0} \eta_{g,\ell} \\ &= \text{ATT}_\ell + \frac{1}{G_1} \sum_{g:D_g=1} \eta_{g,\ell} + o_P(1), \end{aligned} \quad (3.22)$$

where $o_P(1)$ denotes a random variable tending to 0 in probability. To understand the construction of Conley and Taber (2011), suppose for the moment that the distribution of $\frac{1}{G_1} \sum_{g:D_g=1} \eta_{g,\ell}$ is known and let $q_{\alpha/2}$ (resp. $q_{1-\alpha/2}$) denote its quantile of order $\alpha/2$ (resp. $1 - \alpha/2$). Then,

$$\begin{aligned} P(\text{ATT}_\ell \in [\widehat{\beta}_\ell^{\text{fe}} - q_{1-\alpha/2}, \widehat{\beta}_\ell^{\text{fe}} - q_{\alpha/2}]) &= P(q_{\alpha/2} \leq \widehat{\beta}_\ell^{\text{fe}} - \text{ATT}_\ell \leq q_{1-\alpha/2}) \\ &\rightarrow 1 - \alpha, \end{aligned}$$

which implies that $[\widehat{\beta}_\ell^{\text{fe}} - q_{1-\alpha/2}, \widehat{\beta}_\ell^{\text{fe}} - q_{\alpha/2}]$ is an asymptotically valid CI. Now, obviously, the distribution of $\frac{1}{G_1} \sum_{g:D_g=1} \eta_{g,\ell}$ is unknown. To recover it, the authors impose a second condition, namely:

$$\text{The distribution of } \eta_{g,\ell} \text{ does not vary across } g. \quad (3.23)$$

(3.23) is a strengthening of Assumption PT. Under Assumption PT, $E[\eta_{g,\ell}] = 0$, so the expectation of $\eta_{g,\ell}$ does not vary across g . (3.23) further requires that the distribution of $\eta_{g,\ell}$ does not vary across g . Then, for all g in the control group, one has

$$\begin{aligned} \widehat{\eta}_{g,\ell} &= Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0}(\mathbf{0}_{T_0}) - \frac{1}{G_0} \sum_{g':D_{g'}=0} (Y_{g',T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g',T_0}(\mathbf{0}_{T_0})) \\ &= \gamma_{T_0+\ell} - \gamma_{T_0} + \eta_{g,\ell} - \frac{1}{G_0} \sum_{g':D_{g'}=0} (\gamma_{T_0+\ell} - \gamma_{T_0} + \eta_{g',\ell}) \\ &= \eta_{g,\ell} + o_P(1). \end{aligned}$$

Therefore, $\widehat{\eta}_{g,\ell} \xrightarrow{P} \eta_{g,\ell}$. Then, the empirical distribution of the $(\widehat{\eta}_{g,\ell})_{g:D_g=0}$ is a consistent estimator of the distribution of $\eta_{g,\ell}$, for any treated group g . If $G_1 = 1$, $q_{\alpha/2}$ and $q_{1-\alpha/2}$ are just the quantiles of $\eta_{g,\ell}$, which we can consistently estimate by $\widehat{q}_{\alpha/2}$ and $\widehat{q}_{1-\alpha/2}$, the quantiles of order $\alpha/2$ and $1 - \alpha/2$ of the $(\widehat{\eta}_{g,\ell})_{g:D_g=0}$, and

$$[\widehat{\beta}_\ell^{\text{fe}} - \widehat{q}_{1-\alpha/2}, \widehat{\beta}_\ell^{\text{fe}} - \widehat{q}_{\alpha/2}]$$

is an asymptotically valid CI. If $G_1 > 1$, the following algorithm produces consistent estimators of $q_{\alpha/2}$ and $q_{1-\alpha/2}$:

1. For $s = 1$ to S (large):

(a) Draw (with replacement) a sample of size G_1 from $\{g : D_g = 0\}$. Let $(g_1^s, \dots, g_{G_1}^s)$ denote the labels of the corresponding groups.

(b) Compute $\bar{\eta}^s := \frac{1}{G_1} \sum_{i=1}^{G_1} \widehat{\eta}_{g_i^s}$.

2. Compute $\widehat{q}_{\alpha/2}$ and $\widehat{q}_{1-\alpha/2}$, the empirical quantile of order $\alpha/2$ and $1 - \alpha/2$ of $(\bar{\eta}^1, \dots, \bar{\eta}^S)$.

Approach allowing for heteroscedastic disturbances. As discussed by Ferman and Pinto (2019), the condition in (3.23) may be restrictive. In particular, the variance of $\eta_{g,\ell}$ may vary

with groups' size N_g ,⁶ and the distribution of groups' size may not be the same in control and treated groups. We can still perform valid inference if we replace (3.23) by:

$$\eta_{g,\ell} = \sigma(N_g) \times \zeta_{g,\ell}, \text{ where the distribution of } \zeta_{g,\ell} \text{ does not depend on } g. \quad (3.24)$$

Without loss of generality, we can assume that $V(\zeta_{g,\ell}) = 1$. The function $\sigma(N_g)$ is unknown but we suppose we have a consistent estimator $\hat{\sigma}(N_g)$ of it. For instance, Ferman and Pinto (2019) assume that $\sigma^2(N_g) = A + B/N_g$, as is the case if $\eta_{g,\ell}$ is the sum of a group-specific component and the average of independent individual-specific components. Then, one can consistently estimate A and B by a regression, within the control group, of $\hat{\eta}_{g,\ell}^2$ on a constant and $1/N_g$. The following algorithm modifies that from Conley and Taber (2011), so as to still produce consistent estimators of $q_{\alpha/2}$ and $q_{1-\alpha/2}$ in this set-up (to simplify notation, we suppose here that the treated groups are groups $\{1, \dots, G_1\}$):

1. Compute $\hat{\zeta}_{g,\ell} = \hat{\eta}_{g,\ell}/\hat{\sigma}(N_g)$ for all g in the control group.
2. For $s = 1$ to S (large), do:
 - (a) Draw (with replacement) a sample of size G_1 from $\{g : D_g = 0\}$. Let $(g_1^s, \dots, g_{G_1}^s)$ denote the labels of the corresponding groups.
 - (b) Compute $\bar{\hat{\eta}}^s := (1/G_1) \sum_{i=1}^{G_1} \hat{\sigma}(N_i) \hat{\zeta}_{g_i^s, \ell}$.
3. Compute $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$, the empirical quantile of order $\alpha/2$ and $1 - \alpha/2$ of $(\bar{\hat{\eta}}^1, \dots, \bar{\hat{\eta}}^S)$.

We have considered here that $V(\eta_{g,\ell})$ varies with group size but obviously, the same reasoning would apply with any other variable X_g , provided that we can consistently estimate $\sigma(X_g)$.

3.3.3.2 Few treated and control groups

Approach assuming normally distributed disturbances. If the number of control groups is also small, the previous strategies do not apply anymore, because we cannot consistently estimate the $\eta_{g,\ell}$ of control groups. Following Donald and Lang (2007), a first idea, then, is to reinforce (3.23) by imposing a normality assumption on $\eta_{g,\ell}$:

$$\text{For all } g, \eta_{g,\ell} \sim \mathcal{N}(0, \sigma_{\eta,\ell}^2). \quad (3.25)$$

⁶For simplicity, we assume that groups' size is time-invariant.

This assumption may approximately hold if $Y_{g,t}$ is an average over many individuals in cell (g, t) , all cells have a similar number of individuals, and we have weak dependence within each cell, see Bester, Conley and Hansen (2011). Then, we estimate $\sigma_{\eta,\ell}^2$ by

$$\hat{\sigma}_{\eta,\ell}^2 = \frac{1}{G_0 - 1} \sum_{g:D_g=0} \hat{\eta}_{g,\ell}^2.$$

It follows from (3.22) and standard properties of Gaussian vectors that under (3.21) and (3.25),⁷

$$\begin{aligned} \sqrt{\frac{G_0 G_1}{G_0 + G_1}} \frac{\hat{\beta}_\ell^{\text{fe}} - \text{ATT}_\ell}{\sigma_{\eta,\ell}} &\sim \mathcal{N}(0, 1), \\ (G_0 - 1) \frac{\hat{\sigma}_{\eta,\ell}^2}{\sigma_{\eta,\ell}^2} &\sim \chi^2(G_0 - 1). \end{aligned}$$

Moreover, $\hat{\beta}_\ell^{\text{fe}}$ is independent of $\hat{\sigma}_{\eta,\ell}$. As a result,

$$\sqrt{\frac{G_0 G_1}{G_0 + G_1}} \frac{\hat{\beta}_\ell^{\text{fe}} - \text{ATT}_\ell}{\hat{\sigma}_{\eta,\ell}} \sim t_{G_0 - 1},$$

a t distribution with $G_0 - 1$ degrees of freedom. Thus, if we let $q_{G_0-1}(1-\alpha/2)$ denote the quantile of order $1 - \alpha/2$ of such a distribution,

$$\left[\hat{\beta}_\ell^{\text{fe}} - \frac{q_{G_0-1}(1-\alpha/2)}{\sqrt{\frac{G_0 G_1}{G_0 + G_1}}} \hat{\sigma}_{\eta,\ell}, \hat{\beta}_\ell^{\text{fe}} + \frac{q_{G_0-1}(1-\alpha/2)}{\sqrt{\frac{G_0 G_1}{G_0 + G_1}}} \hat{\sigma}_{\eta,\ell} \right]$$

is a CI for ATT_ℓ with coverage equal to $1 - \alpha$. While similar in spirit to Donald and Lang (2007), this CI slightly differs from theirs: our variance estimator only uses control groups and periods T_0 and $T_0 + \ell$. This allows for heterogeneous treatment effects, and autocorrelation and non-stationarity of $(\varepsilon_{g,1}, \dots, \varepsilon_{g,T})$, since $V(\eta_{g,\ell})$ may vary with ℓ .

Approach assuming homogeneous treatment effects. With few treated and control groups, a second idea is to use a permutation-test approach, following DiCiccio and Romano (2017).⁸ Specifically, we maintain (3.21) and (3.23) and add the following constant treatment

⁷The term $(G_0 G_1 / (G_0 + G_1))^{1/2}$ comes from the fact that under (3.25), we have

$$V(\hat{\beta}_\ell^{\text{fe}}) = \sigma_{\eta,\ell}^2 \left(\frac{1}{G_0} + \frac{1}{G_1} \right) = \sigma_{\eta,\ell}^2 \frac{G_0 + G_1}{G_0 G_1}.$$

⁸See also MacKinnon and Webb (2020), who also consider a permutation-based approach in the context of difference-in-differences. Their test statistic differs from that we propose below, and does not lead to valid inference asymptotically under heteroskedasticity.

effect assumption:

$$\text{For all treated } g, \text{TE}_{g,T_0+\ell} = \text{ATT}_\ell. \quad (3.26)$$

This is a strong condition, but note that it mechanically holds if $G_1 = 1$. Under this condition, we can draw inference on ATT_ℓ without imposing the normality condition above. To test that $\text{ATT}_\ell = c$, we use the fact that under (3.21), (3.23), (3.26) and the null hypothesis,

$$\xi_{g,c} := Y_{g,T_0+\ell} - Y_{g,T_0} - cD_g$$

has the same distribution for all groups, irrespective of whether they belong to the treatment or to the control group. Letting π denote an arbitrary permutation of $\{1, \dots, G\}$ and \bar{D} denote the proportion of treated groups, this implies that the distribution of

$$\mathcal{T}_\pi := \frac{\sum_{g=1}^G (D_g - \bar{D}) \xi_{\pi(g),c}}{\sum_{g=1}^G (D_g - \bar{D})^2 \xi_{\pi(g),c}^2}$$

does not depend on π . Let \mathcal{T} denote the test statistic above using $\pi = \text{Id}$, the identity permutation, and let $q_\pi(1 - \alpha)$ denote the empirical quantile of order $1 - \alpha$ of the $(|\mathcal{T}_\pi|)_{\pi \in \mathcal{G}}$, where \mathcal{G} is the set of permutations of $\{1, \dots, G\}$. A test of level α of the null hypothesis can be obtained using the critical region $\{|\mathcal{T}| > q_\pi(1 - \alpha)\}$.⁹ In practice, the set \mathcal{G} is often too large to compute \mathcal{T}_π for all $\pi \in \mathcal{G}$. However, the test keeps its properties if we replace \mathcal{G} by \mathcal{G}' obtained by drawing at random (without replacement) $N - 1$ permutations from $\mathcal{G} \setminus \{\text{Id}\}$, and adding Id to \mathcal{G}' . Also, we can obtain a valid CI on ATT_ℓ by, basically, inverting this test. The following algorithm indicates a fast way to obtain \bar{c} , the upper bound of this interval (the lower bound can be obtained similarly).¹⁰

1. Fix N and create \mathcal{G}' as above.
2. Find an initial value \bar{c}_1 (resp. $\bar{c}_2 > \bar{c}_1$) for which the permutation test $\text{ATT}_\ell = \bar{c}_1$ is not rejected (resp. $\text{ATT}_\ell = \bar{c}_2$ is rejected). Usually, we can choose $\bar{c}_1 = \hat{\beta}_\ell^{\text{fe}}$ and \bar{c}_2 twice the upper bound of the CI based on normality.
3. Fix $\varepsilon < \bar{c}_2 - \bar{c}_1$ small (e.g., $\varepsilon = 10^{-6}$).

⁹Such a test has actually a level slightly lower than α . To ensure exactness, one has to randomize in case $|\mathcal{T}| > q_\pi(1 - \alpha)$. See, e.g., p.1211 of DiCiccio and Romano (2017) for details.

¹⁰Inverting the test may not result in a confidence region that is an interval. However, the construction we propose is a superset of that obtained by test inversion, and it is thus conservative.

4. While $\bar{c}_2 - \bar{c}_1 > \varepsilon$, do:

- (a) Let $\bar{c} = (\bar{c}_1 + \bar{c}_2)/2$.
- (b) Perform the permutation test of $\text{ATT}_\ell = \bar{c}$.
- (c) If the test is rejected, let $\bar{c}_2 = \bar{c}$. Otherwise, let $\bar{c}_1 = \bar{c}$.

Importantly, while these permutation tests and CIs are valid in finite samples under strong conditions, Theorem 3.1 of DiCiccio and Romano (2017) implies that they are also valid asymptotically under, basically, the sole condition that $\overline{D} \rightarrow p \in (0, 1)$.

3.3.3.3 Steps that researchers may follow when $G_0 < 40$ or $G_1 < 40$, and simulations tailored to their application indicate that HC2-BM CIs may not have good coverage

One should start by testing (3.23), using a “permutation pre-trends” test. With the exception of that of Ferman and Pinto (2019), all the CIs in Sections 3.3.3.1 and 3.3.3.2 rely on (3.23). If $T_0 \geq 2$, (3.23) is testable, since it implies that for each $\ell \in \{-1, \dots, -(T_0 - 1)\}$, $(Y_{g, T_0 - \ell} - Y_{g, T_0})_{g: D_g=1}$ and $(Y_{g, T_0 - \ell} - Y_{g, T_0})_{g: D_g=0}$ have the same distribution. Then, we can use a “permutation pre-trends” test to check whether (3.23) holds or not (Bickel, 1969).¹¹

Recommendations when the “permutation pre-trends” test of (3.23) is not rejected.

1. If $G_1 = 1$, we recommend using the permutation-based CIs introduced in the second paragraph of Section 3.3.3.2. Indeed, when $G_1 = 1$, these CIs do not assume homogeneous treatment effects, so they only rely on (3.21) and (3.23), the same conditions as those underlying the CIs of Conley and Taber (2011), but they remain valid even if G_0 is small.
2. If $G_1 > 1$ and $G_0 \geq 50$, we recommend using the CIs of Conley and Taber (2011). The cut-off value of 50 for G_0 comes from the simulations in Conley and Taber (2011), where they do not consider G_0 smaller than 50. Investigating if their CIs remain valid for G_0 smaller than 50 is an interesting avenue for future research.

¹¹Bickel (1969) suggests using the Kolmogorov-Smirnov two-sample test statistic together with a critical value obtained by permutation. This test has nominal level in finite samples. Note that this is a test of (3.23) for a given ℓ . To jointly test (3.23) for all $\ell \in \{-1, \dots, -(T_0 - 1)\}$, one may perform the test for each ℓ separately, and use a Bonferroni adjustment to obtain the joint test’s p-value.

3. If $G_1 > 1$ and $G_0 < 50$, we recommend one of the two CIs discussed in Section 3.3.3.2 above, depending on whether assuming normal disturbances or homogeneous treatment effects seems more credible.

Recommendations when the “permutation pre-trends” test of (3.23) is rejected. If i) $G_0 \geq 50$, ii) one has reasons to believe that the violation of (3.23) only comes from heteroskedasticity, and iii) one also has reasons to believe that heteroskedasticity is mainly driven by some groups’ observable characteristics (for instance: control and treated groups have different populations), then we recommend using the CIs of Ferman and Pinto (2019).

3.4 Limitations of pre-trend tests

3.4.1 We can only test for parallel trends before but not after treatment

While Assumptions NA and PT are partly testable, those assumptions are not fully testable: we can only test for parallel trends before but not after treatment. Even if treated and control groups are on parallel trends before the date when the treatment starts, that does not necessarily mean that without treatment, they would have remained on parallel trends after that date. Accordingly, parallel-trends tests remain suggestive.

3.4.1.1 *Concomitant shocks*

Even when a pre-trends test is not rejected, Kahn-Lang and Lang (2020) argue that researchers should use their contextual knowledge, to assess whether a shock happening at the same time as or after the treatment could have led to differential trends in the absence of treatment. For instance, was there an economic recession at the time of treatment, that could have affected treated and control groups differently and led to differential counterfactual trends? In the compulsory licensing example, researchers could use their contextual knowledge to assess if the sudden spike in the treatment group’s patenting in 1932, visible in Figure 3.2, may not have been due to a shock affecting the treated subclasses and unrelated to the compulsory licensing treatment. To suggestively test whether it is plausible that counterfactual trends would have

been different after treatment even if they were parallel before, one can identify covariates that significantly predict post-treatment outcome trends in the control group, but do not predict pre-treatment outcome trends. Then, if treated and control groups differ on those covariates, it could be that treated and controls were on parallel trends before treatment but would still have experienced differential trends after treatment. Formalizing this proposal, which is related to one by Kahn-Lang and Lang (2020), is an interesting avenue for future research.

3.4.1.2 Other policies

A second, related concern is that policies are rarely implemented in isolation. For instance, Hoehn-Velasco, Penglase, Pesko and Shahid (2024) show that in the US, unilateral divorce laws, which we will return to in Chapter 6, were often adopted at the same time as other policies, in particular legal abortion laws. Then, post-treatment differential trends between treated and control groups may reflect the total effect of all those treatments, rather than just the effect of the treatment under consideration. There again, researchers should document whether other treatments likely to affect the outcome of interest also changed around the time where the treatment of interest changed.

Overall, even when pre-trend tests are not rejected, researchers should present statistical and qualitative evidence that parallel trends would have continued in the post-treatment period. Providing sound guidelines as to how to do so is an important avenue for future research: this issue has received little attention while it is an important one, as we will reiterate in the book’s conclusion.

3.4.2 Pre-trend tests often lack power

Tests of parallel trends are often underpowered in published economics papers. Roth (2022) evaluates the power of parallel trends tests in a sample of 12 papers published in the American Economic Review, American Economic Journal: Applied Economics, and American Economic Journal: Economic Policy between 2014 and June 2018, that contain the phrase “event study”, whose data is publicly available, and that estimated Regression (3.6). For each paper,

he collects the event-study coefficients $(\hat{\beta}_\ell^{\text{fe}})_{\ell \in \{-(T_0-1), \dots, -1, 1, \dots, T_1\}}$ and their estimated variance-covariance matrix $\hat{\Sigma}$, and then runs the following simulations. In each simulation draw, he generates coefficients $(\hat{\beta}_\ell^{s,fe})_{\ell \in \{-(T_0-1), \dots, -1, 1, \dots, T_1\}}$ from a normal distribution with variance-covariance matrix $\hat{\Sigma}$, and where

$$E(\hat{\beta}_\ell^{s,fe}) = \gamma\ell + 1\{\ell \geq 1\}\hat{\beta}_\ell^{\text{fe}},$$

for some real number $\gamma \neq 0$. Interpret the DGP in those simulations. Does the parallel-trends assumption hold? What is the value of ATT_ℓ ?

For $\ell \in \{-(T_0 - 1), \dots, -1\}$, $E(\hat{\beta}_\ell^{s,fe}) = \gamma\ell \neq 0$, so parallel trends fails in this DGP. Treated and control groups experience their own linear trends, and the difference between the linear trends of treated and control groups is equal to γ . ATT_ℓ is equal to $\hat{\beta}_\ell^{\text{fe}}$, the actual value of the estimated ATT_ℓ in each paper. In each simulation draw, Roth mimicks a pre-trends test, which is rejected if there is at least one $\ell \in \{-(T_0 - 1), \dots, -1\}$ such that $|\hat{\beta}_\ell^{s,fe}| > 1.96\hat{\sigma}_\ell$, where $\hat{\sigma}_\ell$ is the estimated standard error of $\hat{\beta}_\ell^{\text{fe}}$.¹² He evaluates the power of the pre-trend tests across a range of values for γ , until finding the value $\gamma_{0.5}$ such that the pre-trends test is rejected 50% of the time. In each of the 12 papers he considers, $\gamma_{0.5}$ represents the differential linear trends between treatment and control groups that have 50% chances of being detected by the researcher. Finally, he evaluates $\frac{1}{T_1} \sum_{\ell=1}^{T_1} \gamma_{0.5}\ell$ and compares that quantity to $\frac{1}{T_1} \sum_{\ell=1}^{T_1} \hat{\beta}_\ell^{\text{fe}}$. What is the purpose of that comparison?

Under differential linear trends that have 50% chances of being detected, $\frac{1}{T_1} \sum_{\ell=1}^{T_1} \gamma_{0.5}\ell$ is the bias of the event-study estimator of the ATT. It is interesting to compare the magnitude of this

¹²As discussed above, this is not a valid procedure to test for pre-trends. However, Roth seeks to reproduce current practice, and articles in his survey that formally test for pre-trends use that procedure.

potential bias with the magnitude of the actual event-study estimate of the ATT, to assess if differential trends that have high chances of not being detected by the researcher can account for a large share of the estimated treatment effect. Results are compelling. His appendix Figure D1, reproduced below, shows that in 7 papers out of 12, under differential linear trends that have 50% chances of being detected, the bias in the event-study estimate of the ATT, the green circle, is no smaller than a half of the actual estimate of the ATT, the blue square. In other words, in 7 papers out of 12, the authors had 50% chances of failing to detect differential trends large enough to account for at least a half of their estimated ATT. Based on his findings, Roth (2022) recommends that practitioners run simulations similar to his, and provides the Stata and R packages `pretrends` for that purpose. Thus, researchers can assess the power of pre-trend tests in their application, and whether they could fail to detect differential trends large enough to account for a substantial fraction of their estimated ATT.

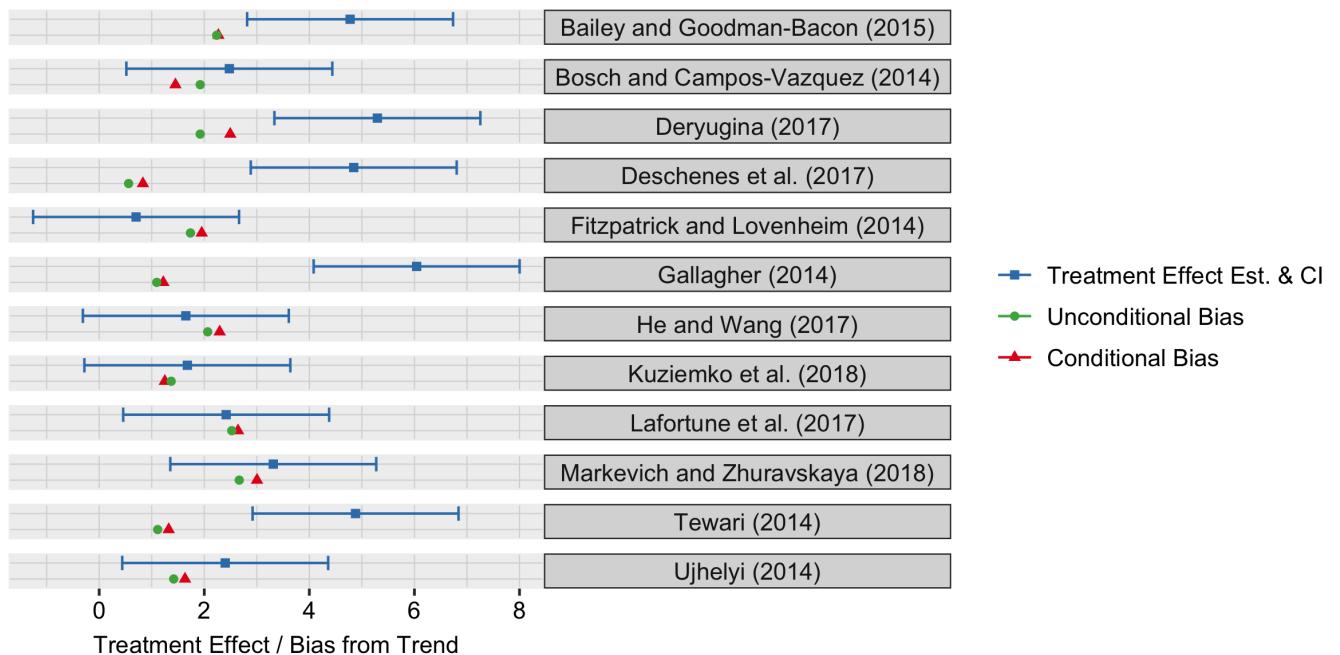


Figure 3.3: Power of pre-trend tests in 12 published economics papers: re-production of Figure D1 in Roth (2022).

Application to the compulsory licensing example. Using the `moser_voena_didtextbook` dataset, run the following lines of code:

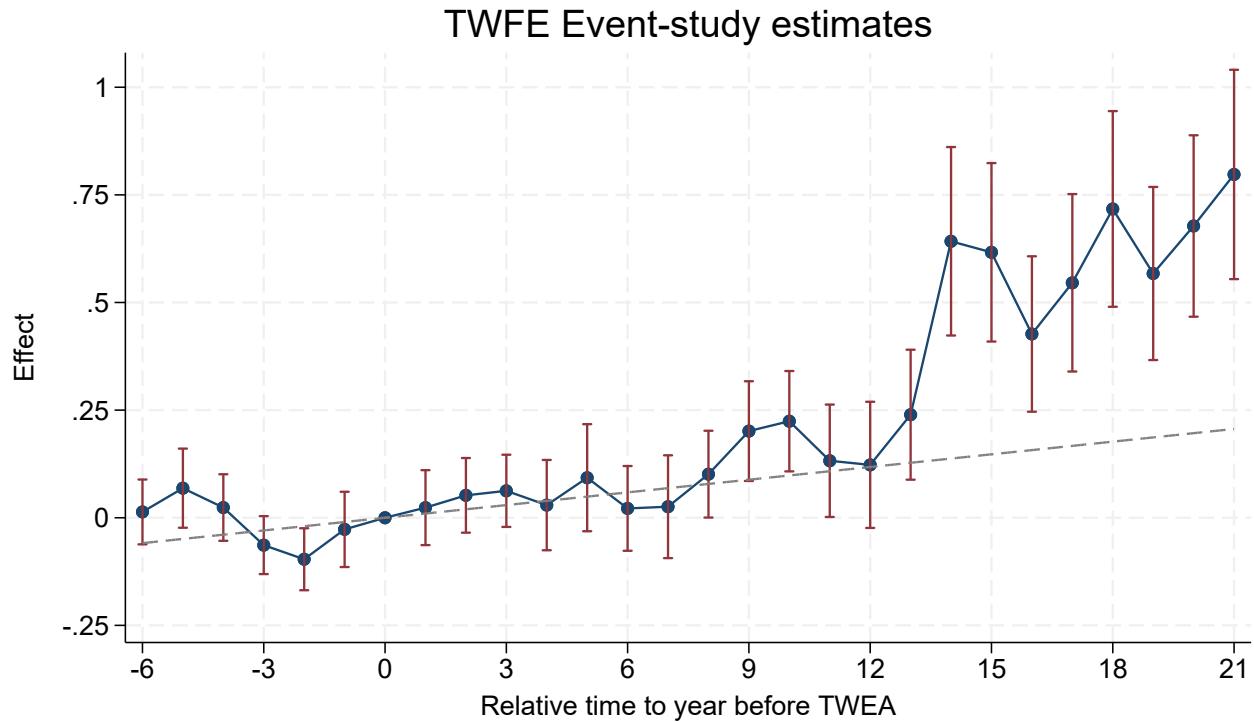
```
local github https://raw.githubusercontent.com
net install pretrends, from('github'/mcaceresb/stata-pretrends/main) replace
reghdfe patents reltimeminus* reltimeplus*, absorb(treatmentgroup year) cluster(subclass)
pretrends power 0.5, numpre(6)
```

This outputs the linear differential trends between treated and control subclasses one would have 50% chances of detecting, in view of the variance-covariance matrix of the first six pre-trends estimates of the ES TWFE regression.¹³ Interpret the result. Could it be the case that our estimated effects of compulsory licensing are entirely generated by differential linear trends we are unable to detect?

We could fail to detect differential linear trends of 0.010. Figure 3.4 shows that extrapolating such differential linear trends throughout the post-treatment period, they can account for all or almost all of the estimated short-term effects of the treatment, but they can only account for around 25% of the estimated effects after 14 years of exposure or more.

¹³The **pretrends** command takes a long time to run with more pre-trends estimates, which is why we only leverage six estimates.

Figure 3.4: Effects of compulsory licensing on US innovation: bias in the TWFE ES regression arising from a hard-to-detect differential linear trend



Note: This figure shows the estimated effects of compulsory licensing on patents, using years 1912 to 1939 of the data from Moser and Voena (2012), and the TWFE event-study regression in (3.6). Standard errors are clustered at the patent subclass level. 95% confidence intervals are shown in red. The grey dotted line shows the differential linear trend one would have 50% chances of detecting, in view of the variance-covariance matrix of the pre-trends estimates, computed using the `pretrends` Stata command.

3.4.3 Pre-trend tests might exacerbate the bias from violations of no anticipation and parallel trends*

Tests of parallel trends may lead to a pre-testing problem. Parallel trends tests are often used as a way to decide whether the analysis should be continued, or which specification should be reported: researchers may add control variables, add group-specific linear trends, change the definition of their control group, etc., until the parallel-trends test is not rejected. This

means that often times, the vector of estimated effects we observe $(\hat{\beta}_\ell^{\text{fe}})_{\ell \in \{1, \dots, T_1\}}$ is conditional on values of the pre-trends coefficients $(\hat{\beta}_\ell^{\text{fe}})_{\ell \in \{-(T_0-1), \dots, -1\}}$ such that the pre-trends test is not rejected. Let Pub be an indicator equal to 1 when that event is realized, where Pub stands for publishable. If $(\hat{\beta}_\ell^{\text{fe}})_{\ell \in \{1, \dots, T_1\}}$ and $(\hat{\beta}_\ell^{\text{fe}})_{\ell \in \{-(T_0-1), \dots, -1\}}$ are not independent, we may have that for $\ell \in \{1, \dots, T_1\}$

$$E(\hat{\beta}_\ell^{\text{fe}} | \text{Pub} = 1) \neq E(\hat{\beta}_\ell^{\text{fe}}),$$

so such pre-testing could lead to a bias of $\hat{\beta}_\ell^{\text{fe}}$, on top of the potential bias that may come from differential trends. Pre-testing could also lead to distorted inference: the distribution of $\hat{\beta}_\ell^{\text{fe}}$ conditional on not rejecting the pre-test may differ from its unconditional distribution, and standard critical values used to construct confidence intervals and tests, which are derived from the unconditional distribution, may not be valid anymore.

If trends are parallel, pre-testing does not lead to a bias and does not distort inference. Reassuringly, Proposition 1 in Roth (2022) shows that when trends are parallel, this additional bias is equal to zero: under parallel trends, testing for pre-trends and estimating the treatment effect only if the pre-trends test is not rejected does not lead to a bias. Furthermore, as shown by de Chaisemartin and D'Haultfœuille (2024), it readily follows from the Gaussian correlation inequality (Royen, 2014) that under parallel trends, conditional on not rejecting the pre-trends test, confidence intervals for treatment effects are conservative. Therefore, pre-testing cannot lead to over-reject null hypotheses.

If trends are not parallel, pre-testing might exacerbate the bias of $\hat{\beta}_\ell^{\text{fe}}$, though this phenomenon seems modest in practice. Proposition 2 in Roth (2022) shows that if trends are not parallel, differential trends widen over time, and the estimators of the pre-trends and actual effects are positively correlated and homoscedastic, then pre-testing leads to a bias which goes in the same direction as the bias coming from differential trends, thus exacerbating it. In Figure D1 in Roth (2022), reproduced above, the red triangles represent $E(\hat{\beta}_\ell^{\text{fe}} | \text{Pub} = 1)$, while the green circles represent $E(\hat{\beta}_\ell^{\text{fe}})$, under the same pre-trends as above. In practice, does the potential bias exacerbation phenomenon uncovered by Roth (2022) seem modest or serious?

This bias exacerbation phenomenon is relatively modest in the 12 papers reviewed by the author: in most cases, the red triangles are close to the green circles.

3.5 An alternative estimator of the event-study effects

3.5.1 Imputation estimator

For $\ell \in \{1, \dots, T_1\}$, let

$$\widehat{\beta}_\ell^{\text{imp}} = \frac{1}{G_1} \sum_{g:D_g=1} \left(Y_{g,T_0+\ell} - \frac{1}{T_0} \sum_{t=1}^{T_0} Y_{g,t} \right) - \frac{1}{G_0} \sum_{g:D_g=0} \left(Y_{g,T_0+\ell} - \frac{1}{T_0} \sum_{t=1}^{T_0} Y_{g,t} \right). \quad (3.27)$$

Explain the difference between $\widehat{\beta}_\ell^{\text{fe}}$ and $\widehat{\beta}_\ell^{\text{imp}}$.

Both estimators are DID estimators comparing the outcome evolution of treated and control groups. In $\widehat{\beta}_\ell^{\text{fe}}$, the “before” period in the “before-after” comparison is $t = T_0$, the last period before treated groups get treated. Instead, in $\widehat{\beta}_\ell^{\text{imp}}$ the “before” period is actually an average of all pre-treatment periods, from period 1 to T_0 . One can show that under Assumptions NA and PT, $\widehat{\beta}_\ell^{\text{imp}}$ is also unbiased for ATT_ℓ .

3.5.2 Three numerical equivalences

A TWFE ES regression to compute $\widehat{\beta}_\ell^{\text{imp}}$. $\widehat{\beta}_\ell^{\text{fe}}$ can be obtained from a TWFE ES regression with $t = T_0$ as the omitted/reference period. With that in mind, which TWFE ES regression could we use to estimate $(\widehat{\beta}_\ell^{\text{imp}})_{\ell \in \{1, \dots, T_1\}}$?

A TWFE ES regression where all pre-treatment periods are omitted:

$$Y_{g,t} = \hat{\alpha}_0^{\text{imp}} + \hat{\alpha}_1^{\text{imp}} D_g + \sum_{t' > T_0} \hat{\gamma}_{t'}^{\text{imp}} 1\{t = t'\} + \sum_{\ell > 0} \hat{\beta}_{\ell}^{\text{imp}} 1\{t = T_0 + \ell\} D_g + \hat{\epsilon}_{g,t}. \quad (3.28)$$

One can also show that $\hat{\beta}_{\ell}^{\text{imp}}$ is numerically equivalent to the coefficients from a regression with FEs for all periods but interactions of D_g and relative-time after but not before treatment:

$$Y_{g,t} = \hat{\alpha}_0^{\text{imp}} + \hat{\alpha}_1^{\text{imp}} D_g + \sum_{t'=1}^T \hat{\gamma}_{t'}^{\text{imp}} 1\{t = t'\} + \sum_{\ell > 0} \hat{\beta}_{\ell}^{\text{imp}} 1\{t = T_0 + \ell\} D_g + \hat{\epsilon}_{g,t}. \quad (3.29)$$

An imputation procedure to compute $\hat{\beta}_{\ell}^{\text{imp}}$. There is a third, numerically-equivalent way to compute $\hat{\beta}_{\ell}^{\text{imp}}$. First, one estimates a TWFE regression of the outcome on group and time FEs in the sample of untreated (g, t) cells (namely, (g, t) s such that either $D_g = 0$ or $t \leq T_0$). Let $\hat{Y}_{g,t}(\mathbf{0}_t)$ denote the predicted or imputed counterfactual outcome of treated cell (g, t) according to that regression: $\hat{Y}_{g,t}(\mathbf{0}_t)$ is the predicted outcome that a treated cell would have had without treatment, according to the TWFE model estimated on the untreated cells. Then, one can show that

$$\hat{\beta}_{\ell}^{\text{imp}} = \frac{1}{G_1} \sum_{g: D_g=1} \left(Y_{g,T_0+\ell} - \hat{Y}_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) \right).$$

Intuitively, to estimate the effect of ℓ periods of exposure to treatment, one merely subtracts the predicted outcome that treated cells would have had without treatment to their actual outcome. This third numerical equivalence is the reason why we refer to $\hat{\beta}_{\ell}^{\text{imp}}$ as an imputation estimator.

3.5.3 Comparing the properties of $\hat{\beta}_{\ell}^{\text{fe}}$ and $\hat{\beta}_{\ell}^{\text{imp}}$

Intuitively, do you expect that $V(\hat{\beta}_{\ell}^{\text{imp}}) > V(\hat{\beta}_{\ell}^{\text{fe}})$, $V(\hat{\beta}_{\ell}^{\text{imp}}) < V(\hat{\beta}_{\ell}^{\text{fe}})$, or $V(\hat{\beta}_{\ell}^{\text{imp}}) = V(\hat{\beta}_{\ell}^{\text{fe}})$?

$V(\hat{\beta}_{\ell}^{\text{imp}}) \leq V(\hat{\beta}_{\ell}^{\text{fe}})$ if errors are not serially correlated. $\hat{\beta}_{\ell}^{\text{fe}}$ use groups' T_0 outcome, the last period before treatment onset, as the baseline outcome. On the other hand, $\hat{\beta}_{\ell}^{\text{imp}}$ uses their

average outcome from period 1 to T_0 . As $\hat{\beta}_\ell^{\text{imp}}$ uses more data than $\hat{\beta}_\ell^{\text{fe}}$, one may expect the former estimator to be more precise than the latter. Actually, whether this is the case depends on the data generating process. Recall that in (2.4), we showed that under Assumptions NA and PT,

$$Y_{g,t}(\mathbf{0}_t) = \alpha_g + \gamma_t + \varepsilon_{g,t}, \quad E[\varepsilon_{g,t}] = 0,$$

with $\varepsilon_{g,t} = Y_{g,t}(\mathbf{0}_t) - E(Y_{g,t}(\mathbf{0}_t))$. Then, Borusyak et al. (2024) show that if the treatment effects are non stochastic and the errors $\varepsilon_{g,t}$ are independent and identically distributed (i.i.d.) across both g and t , $\hat{\beta}_\ell^{\text{imp}}$ is the best linear unbiased estimator (BLUE) of ATT_ℓ . Thus,

$$V[\hat{\beta}_\ell^{\text{imp}}] \leq V[\hat{\beta}_\ell^{\text{fe}}], \quad (3.30)$$

as $\hat{\beta}_\ell^{\text{fe}}$ is also a linear unbiased estimator of ATT_ℓ . However, this result relies on a strong assumption, namely that within group g the errors $\varepsilon_{g,t}$ are uncorrelated over time. This rules out within-group serial correlations that are likely to be present in many empirical settings, and that Bertrand et al. (2004) recommend accounting for when conducting inference in DID studies.

$V(\hat{\beta}_\ell^{\text{imp}}) \geq V(\hat{\beta}_\ell^{\text{fe}})$ if errors follow a random walk. Inequality (3.30) heavily relies on this no-serial-correlation assumption. Instead of assuming independent errors, assume that in each group errors follow a random walk: $\varepsilon_{g,t} = \varepsilon_{g,t-1} + u_{g,t}$, with $u_{g,t}$ i.i.d.. This means that in each group, errors are very strongly positively correlated over time. Then, Harmon (2022) shows that $\hat{\beta}_\ell^{\text{fe}}$ is the BLUE estimator of ATT_ℓ , thus implying that

$$V[\hat{\beta}_\ell^{\text{imp}}] \geq V[\hat{\beta}_\ell^{\text{fe}}]. \quad (3.31)$$

Some intuition for (3.31) goes as follows. Assume that $T_0 = 2$. Then, as treatment effects are assumed to be non-stochastic, the stochastic part of $\hat{\beta}_1^{\text{imp}}$ is a difference between independent averages of

$$\begin{aligned} Y_{g,3}(\mathbf{0}_3) - \frac{1}{2}(Y_{g,2}(\mathbf{0}_2) + Y_{g,1}(\mathbf{0}_1)) &= \gamma_3 - \frac{1}{2}(\gamma_2 + \gamma_1) + \varepsilon_{g,3} - \frac{1}{2}(\varepsilon_{g,2} + \varepsilon_{g,1}) \\ &= \gamma_3 - \frac{1}{2}(\gamma_2 + \gamma_1) + u_{g,3} + \frac{1}{2}u_{g,2}, \end{aligned}$$

where the first equality follows from (2.4) and the second from the random walk assumption. On the other hand, the stochastic part of $\hat{\beta}_1^{\text{fe}}$ is the difference between independent averages of

$$Y_{g,3}(\mathbf{0}_3) - Y_{g,2}(\mathbf{0}_2) = \gamma_3 - \gamma_2 + u_{g,3}.$$

Then, it is easy to see that $V \left[\widehat{\beta}_1^{\text{imp}} \right] > V \left[\widehat{\beta}_1^{\text{fe}} \right]$.

Sensitivity to violations of Assumptions NA and PT. Under Assumptions NA and PT, both $\widehat{\beta}_\ell^{\text{fe}}$ and $\widehat{\beta}_\ell^{\text{imp}}$ are unbiased for ATT_ℓ . However, in non-randomized natural experiments, Assumptions NA and PT do not hold by construction, and pre-trend tests of those assumptions may lack power (see below). Then, it becomes important to assess the sensitivity of $\widehat{\beta}_\ell^{\text{fe}}$ and $\widehat{\beta}_\ell^{\text{imp}}$ to violations of those assumptions. If Assumption PT does not exactly hold and the discrepancy between groups' trends gets larger over longer horizons, as would for instance happen when there are group-specific linear trends, then $\widehat{\beta}_\ell^{\text{imp}}$ is more biased than $\widehat{\beta}_\ell^{\text{fe}}$, because it compares treated and control groups' outcome evolutions over a longer horizon. On the other hand, if Assumption NA fails due to anticipation effects arising at T_0 , just before the treatment starts, $\widehat{\beta}_\ell^{\text{imp}}$ may be less biased than $\widehat{\beta}_\ell^{\text{fe}}$, because $\widehat{\beta}_\ell^{\text{imp}}$ gives less weight than $\widehat{\beta}_\ell^{\text{fe}}$ to period T_0 . However, violations of Assumptions NA and PT may not be equally problematic. Often times, $\widehat{\beta}_\ell^{\text{imp}}$ and $\widehat{\beta}_\ell^{\text{fe}}$ can be immunized against anticipation effects, by redefining the date when the treatment starts as the date when it was announced. On the other hand, it is often harder to immunize those estimators against violations of Assumption PT.

In general, we recommend using $\widehat{\beta}_\ell^{\text{fe}}$, though using one or the other estimator should not really matter. At the end of the day, our recommendation is to rather use $\widehat{\beta}_\ell^{\text{fe}}$. Note that under Assumptions NA and PT, $\widehat{\beta}_\ell^{\text{imp}}$ and $\widehat{\beta}_\ell^{\text{fe}}$ are both unbiased for ATT_ℓ , so both estimators should be close and using one or the other should not make a large difference: if they significantly differ that implies that Assumption NA or PT must fail.

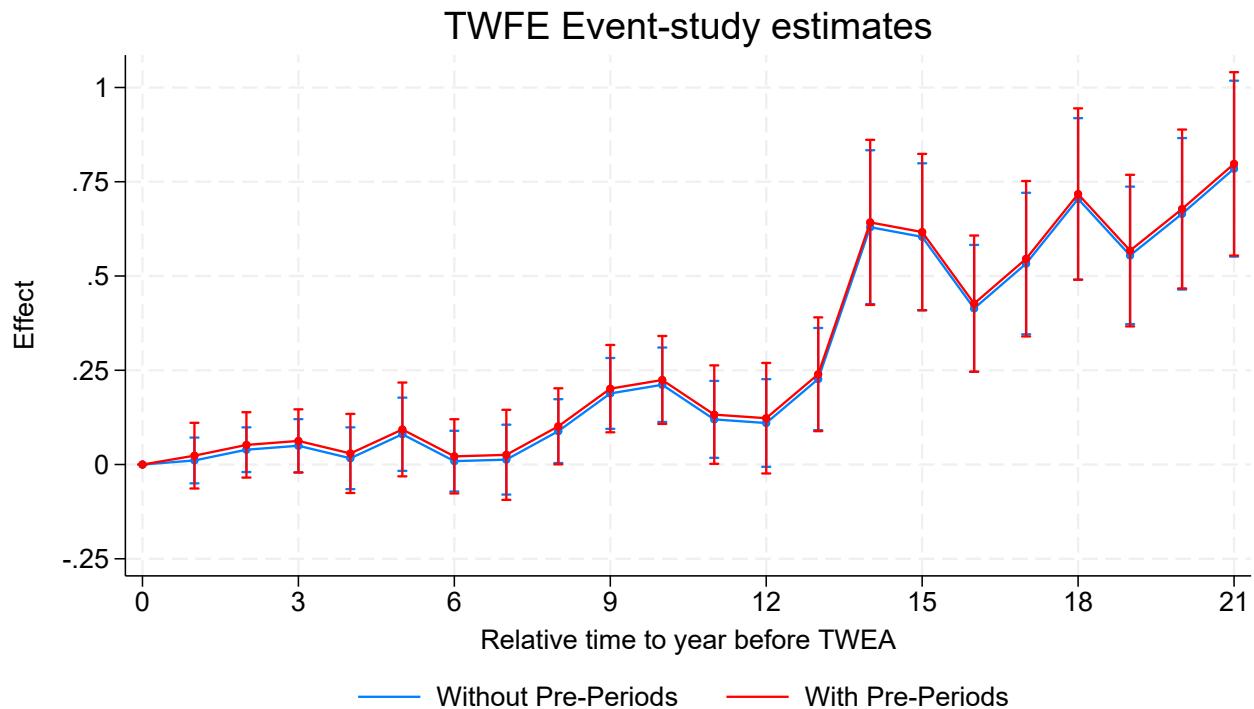
3.5.4 Application to the compulsory licensing example

Using the `moser_voena_didtextbook` dataset, estimate the event-study TWFE regression in (3.28) with patents as the outcome. Are the estimated effects of compulsory licensing in this regression similar to those we obtained estimating (3.6)? Are the effects more precisely estimated?

```
reg patents i.yearpost treatmentgroup reftimeplus*, cluster(subclass)
```

Figure 3.5 below shows the regression's coefficients, as well as those obtained estimating (3.6). The two regressions yield very similar point estimates. Those from (3.28) are slightly more precise.

Figure 3.5: Effects of compulsory licensing on US innovation: comparison of two TWFE ES estimators



Note: This figure shows the estimated effects of compulsory licensing on patents, using years 1900 to 1939 of the data from Moser and Voena (2012), and the TWFE event-study regressions in (3.6) and (3.28). Standard errors are clustered at the patent subclass level. The 95% confidence intervals rely on a normal approximation.

3.6 Estimating heterogeneous treatment effects

3.6.1 Estimating the correlation between treatment effects and some covariates

Target parameter: best-linear predictor. Assume that one wants to assess if the group-specific effects of ℓ periods of exposure to treatment $\text{TE}_{g,T_0+\ell}$ are correlated with a $K \times 1$ vector of time-invariant covariates X_g . Let $X_{g,k}$ denote the k th coordinate of X_g . We assume that $X_{g,1} = 1$: the first coordinate of X_g is a constant. For instance, in an analysis at the level of Chilean communes, one may want to investigate if communes' effects are correlated with their poverty rate and/or the share of their population with a college degree. For any matrix A let A^T denote its transpose. Let

$$\beta_{\ell,X} = \left(\sum_{g:D_g=1} X_g X_g^T \right)^{-1} \sum_{g:D_g=1} X_g \text{TE}_{g,T_0+\ell}$$

be the coefficient on X_g in an infeasible regression of $\text{TE}_{g,T_0+\ell}$ on X_g among treated groups, and let $\beta_{\ell,X,k}$ denote the k th coordinate of $\beta_{\ell,X}$. $X_g^T \beta_X$ is often referred to as the best linear predictor of $\text{TE}_{g,T_0+\ell}$ given X_g . If $K = 2$, X_g contains only one non-constant variable $X_{g,2}$, and one has

$$\beta_{\ell,X,2} = \frac{\sum_{g:D_g=1} (X_{g,2} - \bar{X}_{.,2}) \text{TE}_{g,T_0+\ell}}{\sum_{g:D_g=1} (X_{g,2} - \bar{X}_{.,2})^2},$$

where $\bar{X}_{.,2}$ is the average of $X_{g,2}$ across treated groups. If $X_{g,2}$ is binary, $\beta_{\ell,X,2}$ further simplifies to

$$\frac{1}{G_{1,1}} \sum_{g:D_g=1, X_{g,2}=1} \text{TE}_{g,T_0+\ell} - \frac{1}{G_{1,0}} \sum_{g:D_g=1, X_{g,2}=0} \text{TE}_{g,T_0+\ell},$$

the difference between the ATT of treated groups with $X_{g,2} = 1$ and $X_{g,2} = 0$, where $G_{d,x}$ denotes the number of groups such that $D_g = d, X_{g,2} = x$.

Estimator when $K = 2$ and $X_{g,2}$ is binary. If $K = 2$ and $X_{g,2}$ is binary, estimating $\beta_{\ell,X,2}$ is straightforward: one can just estimate the TWFE ES regression in (3.6) separately among groups such that $X_{g,2} = 1$ and $X_{g,2} = 0$, and take the difference between the coefficients on $1\{t = T_0 + \ell\}D_g$ in the two regressions. This is equivalent to running a TWFE ES regression of $Y_{g,t}$ on $D_g, X_{g,2}, D_g X_{g,2}$, time FEs, time FEs interacted with $X_{g,2}$, $(1\{t = T_0 + \ell\}D_g)_{\ell \in \{-(T_0-1), \dots, T_1\}, \ell \neq 0}$,

and $(1\{t = T_0 + \ell\}D_g X_{g,2})_{\ell \in \{-(T_0 - 1), \dots, T_1\}, \ell \neq 0}$. The coefficient on $1\{t = T_0 + \ell\}D_g X_{g,2}$ is numerically equivalent to the difference between the coefficients on $1\{t = T_0 + \ell\}D_g$ in the two separate regressions. This estimator is unbiased for $\beta_{\ell,X,2}$ under the following assumption: for $x \in \{0, 1\}$,

$$\begin{aligned} & E \left[\frac{1}{G_{1,x}} \sum_{g:D_g=1, X_{g,2}=x} (Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})) \right] \\ &= E \left[\frac{1}{G_{0,x}} \sum_{g:D_g=0, X_{g,2}=x} (Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})) \right]. \end{aligned} \quad (3.32)$$

This assumption requires that treated and control groups with the same value of $X_{g,2}$ have the same average expected outcome evolution without treatment, a conditional parallel-trends assumption. When X_g takes a small number of values relative to the sample size, a similar strategy can be used to estimate the differences between the ATT across groups with different values of X_g . However, this type of estimation strategy is no longer applicable when X_g takes a large number of values, as is for instance the case if some of its coordinates are continuously distributed.

Estimator in the general case. In Section 4.1.3.6 of the next chapter, we will discuss methods to estimate $\beta_{\ell,X}$ under a conditional parallel-trends assumption, when X_g takes a large number of values. For now, let us introduce an alternative method which relies on a different assumption and is extremely simple to implement. Let

$$\hat{\beta}_{\ell,X} = \left(\sum_{g:D_g=1} X_g X_g^T \right)^{-1} \sum_{g:D_g=1} X_g (Y_{g,T_0+\ell} - Y_{g,T_0})$$

denote the coefficient on X_g in a regression of $Y_{g,T_0+\ell} - Y_{g,T_0}$ on a constant and X_g among treated groups. Let $\beta_{\ell,X}^{\Delta Y(\mathbf{0})}$ be the coefficient in a regression of $E[Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0}(\mathbf{0}_{T_0})]$ on X_g among treated groups. If the covariates are non stochastic or conditioned upon, and if for all $k \in \{2, \dots, K\}$,

$$\beta_{k,\ell,X}^{\Delta Y(\mathbf{0})} = 0, \quad (3.33)$$

meaning that untreated outcome evolutions of treated groups are uncorrelated with the non-constant variables in X_g , then one can show that $\hat{\beta}_{\ell,X,k}$ is unbiased for $\beta_{\ell,X,k}$. Interestingly, $\hat{\beta}_{\ell,X}$ can be computed even if there is no control group such that $D_g = 0$. Thus, under (3.33) control groups are not necessary to estimate the correlation between treatment effects and some covariates, a result that seems to have been first noted in Shahn (2023) and in this textbook.

Assessing the plausibility of (3.33), the assumption underlying $\hat{\beta}_{\ell,X}$. First, note that under Assumption PT, $E [Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0}(\mathbf{0}_{T_0})]$ is constant so (3.33) holds: (3.33) is weaker than the strong parallel-trends condition in Assumption PT. Then, if X_g contains only one non-constant variable and that this variable is binary, (3.33) reduces to

$$E \left[\frac{1}{G_{1,1}} \sum_{g:D_g=1, X_{g,2}=1} (Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})) \right] = E \left[\frac{1}{G_{1,0}} \sum_{g:D_g=1, X_{g,2}=0} (Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})) \right],$$

a parallel-trends assumption, between treated groups with $X_{g,2} = 1$ and treated groups with $X_{g,2} = 0$. That condition is neither stronger nor weaker than the conditional parallel-trends assumption in (3.32). Finally, for $\ell \leq T_0 - 1$, (3.33) is placebo testable, by regressing $Y_{g,T_0-\ell} - Y_{g,T_0}$ on X_g in the sample of treated groups, and testing if the coefficient on X_g is equal to zero.

Comparison with current practice. To estimate heterogeneous treatment effects, applied researchers often use TWFE regressions with an interaction term between the treatment or relative time since treatment and the covariates. For instance, assume that $K = 2$ and one runs a TWFE ES regression of $Y_{g,t}$ on a full set of group FEs, time FEs, $(1\{t = T_0 + \ell\}D_g)_{\ell \in \{-(T_0-1), \dots, T_1\}, \ell \neq 0}$, and $(1\{t = T_0 + \ell\}D_g X_{g,2})_{\ell \in \{-(T_0-1), \dots, T_1\}, \ell \neq 0}$. If $T = 2$ and $T_0 = T_1 = 1$, one can show that the coefficient on $1\{t = T_0 + 1\}D_g X_{g,2}$ is numerically equivalent to $\hat{\beta}_{1,X,2}$ above, so this estimation method yields an unbiased estimator of $\beta_{1,X,2}$ under (3.33). When $T > 2$, the coefficients on $1\{t = T_0 + \ell\}D_g X_{g,2}$ are not numerically equivalent to $\hat{\beta}_{\ell,X,2}$, and we are not aware of a paper studying what these coefficients estimate under a parallel-trends assumption.

3.6.2 Estimating the variance of group-specific effects.*

In this section, our target parameter is

$$v_\ell^2 := \frac{1}{G_1} \sum_{g:D_g=1} (\text{TE}_{g,T_0+\ell} - \text{ATT}_\ell)^2,$$

the variance of the effects $(\text{TE}_{g,T_0+\ell})_{g:D_g=1}$. Assume that treatment effects $Y_{g,T_0+\ell}(\mathbf{0}_{T_0}, \mathbf{1}_\ell) - Y_{g,T_0}(\mathbf{0}_{T_0+\ell})$ are not random, as in (3.21). Then,

$$V(Y_{g,T_0+\ell} - Y_{g,T_0}) = V(Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0}(\mathbf{0}_{T_0})) = V(\eta_{g,\ell}),$$

where we recall that

$$\eta_{g,\ell} = Y_{g,T_0+\ell}(\mathbf{0}_{g,T_0+\ell}) - E[Y_{g,T_0+\ell}(\mathbf{0}_{g,T_0+\ell})] - (Y_{g,T_0}(\mathbf{0}_{g,T_0}) - E[Y_{g,T_0}(\mathbf{0}_{g,T_0})]).$$

If one further assumes that $V(\eta_{g,\ell})$ does not depend on g , a weaker assumption than (3.23), the assumption made by Conley and Taber (2011), then it readily follows from (3.14) and (3.15) that

$$\hat{v}_\ell^2 := \frac{G_1 - 1}{G_1} (\hat{\sigma}_{\ell,1}^2 - \hat{\sigma}_{\ell,0}^2)$$

is unbiased for v_ℓ^2 .¹⁴ \hat{v}_ℓ^2 is essentially a comparison of the variances of the outcome's long-differences in the treatment and control groups.¹⁵ Omitting the $(G_1 - 1)/G_1$ term, which will often be close to one, one can use, say, the **sdttest** Stata command to compute \hat{v}_ℓ and test the null that $v_\ell = 0$.

Application to the compulsory licensing example. Using the `moser_voena_didtextbook` dataset, run the following line of code:

```
sdttest diffpatentswrt1918 if year==1932, by(treatmentgroup)
```

Can you reject the null that $v_{14} = 0$? Is there evidence of treatment effect heterogeneity after 14 years of exposure to treatment?

We strongly reject the null that $v_{14} = 0$. Moreover, $\hat{v}_{14} = 0.239$, which is more than a third of $\hat{\beta}_{14}^{\text{fe}} = 0.642$. If the group-specific effects $\text{TE}_{g,T_0+\ell}$ were normally distributed, their first and

¹⁴A similar result can be obtained with random treatment effects: if $Y_{g,T_0+\ell}(\mathbf{0}_{T_0}, \mathbf{1}_\ell) - Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell})$ is uncorrelated with $\eta_{g,\ell}$, \hat{v}_ℓ^2 is consistent, as $G_0, G_1 \rightarrow \infty$, for

$$\tilde{v}_\ell^2 := E \left[\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,T_0+\ell}(\mathbf{0}_{T_0}, \mathbf{1}_\ell) - Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - \text{ATT}_\ell)^2 \right],$$

a treatment-effect variance accounting both for heterogeneous treatment effects between groups, and for heterogeneous treatment effects within groups due to the randomness of $Y_{g,T_0+\ell}(\mathbf{0}_{T_0}, \mathbf{1}_\ell) - Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell})$.

¹⁵A drawback of \hat{v}_ℓ^2 is that it can be negative. One could consider instead $\max[0, \hat{v}_\ell^2]$, but this latter estimator is upward biased.

last decile would be respectively 0.336 and 0.948, thus indicating a substantial amount of effect heterogeneity.

Unbiased estimation of v_ℓ^2 relies on stronger assumptions than unbiased estimation of ATT_ℓ : the strong parallel-trends condition in Assumption PT instead of the weaker one in (2.5), and the assumption that $V(\eta_{g,\ell})$ does not depend on g . How could you run a pre-trends test of those assumptions?

By considering $\hat{\sigma}_{\ell,1}^2 - \hat{\sigma}_{\ell,0}^2$, for $\ell \leq -1$. This is similar in spirit to a standard pre-trends test, except that we compare the variances of the outcome evolutions in the treatment and control groups prior to the treatment, rather than comparing their averages. Using the `moser_voena_didtextbook` dataset, we run the following line of code:

```
sdtest diffpatentswrt1918 if year==1904, by(treatmentgroup),
```

thus giving us a placebo estimator symmetric to our estimator of $v_{14} = 0$. We strongly reject the null that $\sigma_{-14,1}^2 - \hat{\sigma}_{-14,0}^2 = 0$, thus casting doubt on Assumption PT and the assumption that $V(\eta_{g,\ell})$ does not depend on g . However, for every $\ell \leq -1$, we find that $\hat{\sigma}_{\ell,1}^2 - \hat{\sigma}_{\ell,0}^2 < 0$, with a difference between the two variances that is large and stable over time. This suggests that for $\ell \geq 1$, our estimators of the treatment effect variance \hat{v}_ℓ^2 might be downward biased: perhaps treatment effects are even more heterogeneous than suggested by \hat{v}_ℓ^2 .

3.6.3 Estimating the distribution of group-specific effects*

Under Assumption NA and the strong parallel-trends condition in Assumption PT, one can unbiasedly estimate the group-level treatment effects

$$\text{TE}_{g,T_0+\ell} = E [Y_{g,T_0+\ell}(\mathbf{0}_{T_0}, \mathbf{1}_\ell) - Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell})],$$

for instance using

$$\widehat{\text{TE}}_{g,T_0+\ell} := Y_{g,T_0+\ell} - Y_{g,T_0} - \frac{1}{G_0} \sum_{g': D_{g'}=0} (Y_{g',T_0+\ell} - Y_{g',T_0}),$$

a DID estimator comparing the T_0 to $T_0 + \ell$ outcome evolution of g to that of all control groups. Then, one may consider using $(\widehat{\text{TE}}_{g,T_0+\ell})_{g:D_g=1}$ to estimate the distribution of the effects $(\text{TE}_{g,T_0+\ell})_{g:D_g=1}$. Under the independent groups framework in Assumption IND, an estimator may be consistent if it averages outcomes from a number of groups that goes to infinity when $G \rightarrow \infty$. For any value of G , $\bar{Y}_{g,T_0+\ell} - \bar{Y}_{g,T_0}$ averages the outcomes of only one group, so $\widehat{\text{TE}}_{g,T_0+\ell}$ is not consistent. As the estimators $(\widehat{\text{TE}}_{g,T_0+\ell})_{g:D_g=1}$ are not consistent, using the distribution of those estimators to estimate the distribution of treatment effects across groups would be misleading: roughly speaking, one needs to remove the estimation error from the distribution of $(\widehat{\text{TE}}_{g,T_0+\ell})_{g:D_g=1}$. Arellano and Bonhomme (2012) propose a deconvolution method that can be applied to recover the density of group-specific treatment effects in DID models, though the resulting estimator relies on the strong assumption that the treatment effects are independent of the errors $(\varepsilon_{g,1}, \dots, \varepsilon_{g,T})$ in (2.4), and will often converge at a slow rate (see Bonhomme and Robin, 2010).

3.7 Non-linear DID

3.7.1 Limited dependent variables

In this section, to simplify the exposition we assume that $T = 2$, $T_0 = T_1 = 1$, and that Assumption ND holds. Also, for all $(d, t) \in \{0, 1\}^2$, and any variable $X_{g,t}$ let $\bar{X}_t^d = \frac{1}{G_d} \sum_{g:D_g=d} X_{g,t}$ denote the average of $X_{g,t}$ across groups with $D_g = d$ at period t .

With a binary outcome, the parallel-trends assumption can yield problematic predictions. Researchers are often interested in limited dependent variables, that can only take a limited set of values. To fix ideas, let us assume for now that $Y_{g,t}$ is binary. Under Assumption PT, $\bar{Y}_1^1 + \bar{Y}_2^0 - \bar{Y}_1^0$ is supposed to estimate $E[\bar{Y}_2^1(0)]$, the expected average outcome without treatment in the treatment group at period two. **This estimator fails to satisfy a desirable property, which one?**

This estimator is not guaranteed to be included between zero and one, while by construction, $E[\bar{Y}_2^1(0)]$ must be included between zero and one. For instance, if the average outcome is equal to 0.8 in the treatment group at period one, and the average outcome increases from 0.4 to 0.7 in the control group, then the treatment group's estimated counterfactual outcome at period two is equal to 1.1, which is not possible.

An alternative parallel-trends assumption. To alleviate this concern, Blundell, Costa-Dias, Meghir and Van Reenen (2004) and Wooldridge (2023) propose to replace Assumption PT by the following condition:

$$\begin{aligned} & L^{-1}(E[\bar{Y}_2^1(0)]) - L^{-1}(E[\bar{Y}_1^1(0)]) \\ &= L^{-1}(E[\bar{Y}_2^0(0)]) - L^{-1}(E[\bar{Y}_1^0(0)]), \end{aligned} \quad (3.34)$$

for a known, strictly increasing function L , taking values in $[0, 1]$. Note that if L is the identity function, then (3.34) is equivalent to Assumption PT. With two pre-treatment periods or more, one can run pre-trend tests of (3.34), comparing the evolution of L^{-1} (average outcome) in the treatment and control groups, before the treatment starts. [Show that \$E\[\bar{Y}_2^1\(0\)\]\$ is identified under \(3.34\).](#)

$$\begin{aligned} & E[\bar{Y}_2^1(0)] \\ &= L(L^{-1}(E[\bar{Y}_2^1(0)])) \\ &= L(L^{-1}(E[\bar{Y}_1^1(0)]) + L^{-1}(E[\bar{Y}_2^1(0)]) - L^{-1}(E[\bar{Y}_1^1(0)])) \\ &= L(L^{-1}(E[\bar{Y}_1^1(0)]) + L^{-1}(E[\bar{Y}_2^0(0)]) - L^{-1}(E[\bar{Y}_1^0(0)])) \\ &= L(L^{-1}(E[\bar{Y}_1^1]) + L^{-1}(E[\bar{Y}_2^0]) - L^{-1}(E[\bar{Y}_1^0])), \end{aligned} \quad (3.35)$$

where the third equality follows from (3.34).

Estimation of the ATT under (3.34). From what precedes, a natural estimator of the ATT is

$$\hat{\beta}_{\text{bin}}^{\text{fe}} := \bar{Y}_2^1 - L \left(L^{-1}(\bar{Y}_1^1) + L^{-1}(\bar{Y}_2^0) - L^{-1}(\bar{Y}_1^0) \right).$$

One can show that

$$\hat{\beta}_{\text{bin}}^{\text{fe}} = L(\hat{\alpha}_0 + \hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\alpha}_3) - L(\hat{\alpha}_0 + \hat{\alpha}_1 + \hat{\alpha}_2), \quad (3.36)$$

where $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3)$ is the maximum likelihood estimator of a binary choice model assuming $P(Y_{g,t} = 1) = L(\alpha_0 + D_g \alpha_1 + 1\{t = 2\} \alpha_2 + D_g 1\{t = 2\} \alpha_3)$. If L is the logistic (resp. normal) cdf, for instance, $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3)$ can be obtained by a logit (resp. probit) regression. Such regressions were first considered by Puhani (2012). In (3.36), $\hat{\beta}_{\text{bin}}^{\text{fe}}$ only depends on logit (resp. probit) coefficients, so standard logit (resp. probit) postestimation routines can be used to estimate its variance. We recommend using this estimator with a binary outcome: it is almost as simple to compute as $\hat{\beta}^{\text{fe}}$, drawing inference on it is also simple, and it also relies on a partly-testable parallel-trends assumption, which may be more plausible than that underlying $\hat{\beta}^{\text{fe}}$.

Non-negative dependent variables. With non-negative dependent variables, one can use a Poisson regression of $Y_{g,t}$ on an intercept, D_g , $1\{t = 2\}$, and $D_g 1\{t = 2\}$ to estimate the ATT. Wooldridge (2023) shows that this estimator relies on (3.34) with $L^{-1}() = \ln()$, a parallel-trends assumption on the logarithm of the average untreated outcome.

3.7.2 Sensitivity to functional form*

In this section and in the next, to simplify the exposition we assume that $T = 2$ and $T_0 = T_1 = 1$. Then, because groups can be treated for at most one period of time, Assumption ND can be imposed without loss of generality. Also, we adopt a sampling-based perspective, where the G groups we observe are an independent and identically distributed sample from a larger population, as assumed in Assumption IID. Then, the g subscript can be dropped, the design is no longer conditioned upon, and we make the following parallel-trends assumption:

$$E_u [Y_2(0) - Y_1(0)|D = 1] = E_u [Y_2(0) - Y_1(0)|D = 0], \quad (3.37)$$

where we recall that $E_u[\cdot]$ denotes unconditional expectations, that are not conditional to the design.

3.7.2.1 The parallel-trends assumption may not be invariant to functional form

The parallel-trends assumption may not be invariant to functional form. If (3.37) holds, then parallel trends also holds for any linear or affine transformation of the outcome: parallel trends is invariant to linear or affine transformations. On the other hand, (3.37) does not necessarily imply that for any strictly increasing function $h(\cdot)$,

$$E_u[h(Y_2(0)) - h(Y_1(0))|D = 1] = E_u[h(Y_2(0)) - h(Y_1(0))|D = 0] :$$

the parallel-trends assumption is not invariant to functional form. For instance, it may hold in levels but not in logs, and conversely. A practical consequence is that DID estimators can be sensitive to functional form: for instance, they can give opposite results with Y and $\ln(Y)$. An early example of this phenomenon is Meyer, Viscusi and Durbin (1995). They use increases in the benefits workers receive when injured that took place in the 1980s in Kentucky and Michigan, to study the effect of the benefits' amount on injury duration. While they do not find an effect on injuries' duration, they do find an effect on the log of injuries' duration.

Building up intuition. Consider a simple numerical example with two-periods and two groups, where group one is the control group and group two is the treated group, and $Y_{1,t} = t$ while $Y_{2,t} = 2t$. Then, the DID estimator on $Y_{g,t}$ is equal to

$$2 \times 2 - 2 \times 1 - (1 \times 2 - 1 \times 1) = 1 :$$

from period one to two, the outcome increases by one more unit in the treatment than in the control group, which might suggest that the treatment has a positive effect. Yet, the DID estimator on $\ln(Y_{g,t})$ is equal to

$$\ln(4) - \ln(2) - (\ln(2) - \ln(1)) = 0 :$$

the outcome increases by 100% in the treatment and in the control group, which might, equally reasonably, suggest that the treatment has no effect. [Which feature of this numerical example](#)

makes DID estimators sensitive to functional form? Can you think of another numerical example where DID estimators would be insensitive to functional form?

In the numerical example, the sensitivity to functional form comes from the very different outcome levels at period one in the treatment and control groups. If the treatment and control group have the same outcome at period one ($Y_{2,1} = Y_{1,1}$), then

$$h(Y_{2,2}) - h(Y_{2,1}) - (h(Y_{1,2}) - h(Y_{1,1})) = h(Y_{2,2}) - h(Y_{1,2}),$$

an estimator whose sign is the same for any strictly increasing function $h(\cdot)$. This suggests that sensitivity to functional form is more likely to be a concern when treated and control groups have very different pre-treatment outcome levels, as is for instance the case in Meyer et al. (1995), than when their pre-treatment outcome levels are similar.

3.7.2.2 A necessary and sufficient condition to have parallel-trends for any functional form

Roth and Sant'Anna (2023) show that

$$E_u [h(Y_2(0)) - h(Y_1(0))|D = 1] = E_u [h(Y_2(0)) - h(Y_1(0))|D = 0]$$

for any strictly increasing function $h(\cdot)$ if and only if, for all $y \in \mathbb{R}$,

$$\begin{aligned} & P_u(Y_2(0) \leq y|D = 1) - P_u(Y_1(0) \leq y|D = 1) \\ &= P_u(Y_2(0) \leq y|D = 0) - P_u(Y_1(0) \leq y|D = 0) : \end{aligned} \quad (3.38)$$

parallel trends holds for any functional form if and only if the cumulative distribution function (cdf) of the untreated outcome follows parallel trends in the treatment and control groups. (3.38) is a strong condition. As shown by Roth and Sant'Anna (2023), this condition also has a testable implication. It implies that

$$\begin{aligned} P_u(Y_2(0) \leq y|D = 1) &= P_u(Y_1(0) \leq y|D = 1) + P_u(Y_2(0) \leq y|D = 0) - P_u(Y_1(0) \leq y|D = 0) \\ &= P_u(Y_1 \leq y|D = 1) + P_u(Y_2 \leq y|D = 0) - P_u(Y_1 \leq y|D = 0). \end{aligned} \quad (3.39)$$

The right-hand-side of the previous display is the outcome's cdf at period one in the treatment group plus the change in the outcome's cdf in the control group, a function that is identified and can be estimated from the data, replacing probabilities by sample proportions. If (3.38) holds, $y \mapsto P_u(Y_1 \leq y|D = 1) + P_u(Y_2 \leq y|D = 0) - P_u(Y_1 \leq y|D = 0)$ identifies the cdf of $Y_2(0)$ in the treatment group. Then, as a cdf is weakly increasing, $y \mapsto P_u(Y_1 \leq y|D = 1) + P_u(Y_2 \leq y|D = 0) - P_u(Y_1 \leq y|D = 0)$ should be weakly increasing in y . But because the increasing function $P_u(Y_1 \leq y|D = 0)$ enters with a negative sign in it, that function may or may not be increasing, hence the testability of (3.38). Kim and Wooldridge (2024) propose a test, computed by the `ddid` (Kim, 2024) Stata package. When the test is rejected, parallel trends is sensitive to functional form. An alternative way of assessing the plausibility of (3.38) is by testing if the same condition holds in pre-treatment periods. However, as the null involves functions rather than scalars, the testing problem is more complicated than with standard pre-trends test, though one may be able to construct a Kolmogorov-Smirnov or Cramer-Von Mises test. One could also combine a pre-trends test of (3.38) with the monotonicity test of (3.39).

3.7.2.3 When should researchers worry about sensitivity to functional form?

Average treatment effects are also sensitive to functional form. First it is important to realize that sensitivity to functional form is not a unique feature of the parallel-trends assumption. Average treatment effects are also sensitive to functional form. In general, for a non-linear and strictly increasing function $h(\cdot)$, $E_u[h(Y_2(d))|D = 1] \neq h(E_u[Y_2(d)|D = 1])$. Then, one may have that $E_u[Y_2(1) - Y_2(0)|D = 1]$ and $E_u[h(Y_2(1)) - h(Y_2(0))|D = 1]$ are of a different sign. Therefore, a positive DID in levels and a negative one in logs either means that parallel trends is violated for at least one of the two functional forms, or that parallel trends holds for both functional forms but $\text{ATT} > 0$ while $\text{ATT}_{\ln} := E_u(\ln(Y_2(1)) - \ln(Y_2(0))|D = 1) < 0$.

Sensitivity to functional form is not an issue if the research question dictates a specific functional form. Assume that $Y_2(0) > 0$, and the researcher's target parameter is

$$\text{ASE} := E_u(Y_2(1)/Y_2(0) - 1|D = 1),$$

the average across all treated groups of the relative outcome change in response to treatment, the so-called average semi-elasticity. Due to the ratio inside the average, this parameter cannot

be estimated using a standard DID. However, using the fact that $Y_2(1)/Y_2(0) - 1 \approx \ln(Y_2(1)) - \ln(Y_2(0))$ if $Y_2(1)/Y_2(0) - 1$ is close to zero, ATT_{\ln} may provide a reasonable approximation of ASE. Then, the researcher should use a DID estimator on the log outcome. That estimator is unbiased if parallel trends holds in logs, even if parallel trends fails for other functional forms.

Sensitivity to functional form is not an issue if the researcher is indifferent between several functional forms, and pre-trend tests can detect functional forms for which parallel trends fails. Assume instead that the researcher's goal is to estimate $E_u(h(Y_2(1)) - h(Y_2(0)) | D = 1)$ for one strictly increasing function $h(\cdot)$ belonging to a set \mathcal{H} , but they do not have a preference over the elements of \mathcal{H} . For instance, one may either want to estimate ATT or ATT_{\ln} , with no strong preference between the two. Such indifference could be justified by the fact that even if they are interested in the ASE, if (3.37) holds in levels but not in logs, the researcher can always estimate ATT, and normalize their estimator by

$$\frac{1}{G_1} \sum_{g:D_g=1} Y_1 + \frac{1}{G_0} \sum_{g:D_g=0} (Y_2 - Y_1),$$

an unbiased estimator of $E(Y_2(0)|D = 1)$ under (3.37), thus yielding an estimator of

$$\text{SEA} := \frac{\text{ATT}}{E_u(Y_2(0)|D = 1)} = \frac{E_u(Y_2(1)|D = 1)}{E_u(Y_2(0)|D = 1)} - 1,$$

the so-called semi-elasticity of the average. Neither the SEA nor the ATT_{\ln} are exactly equal to the ASE, but both may provide a reasonable approximation, and there may not be a strong argument to prefer one over the other. Then, the researcher can use a DID estimator on Y if parallel-trends holds in levels, or a DID on $\ln(Y)$ if parallel-trends holds in logs. Pre-trend tests can help them determine if parallel-trends holds for any or both of these functional forms. If there is only one functional form for which pre-trends tests are not rejected, then the researcher should use that functional form.

Instances where sensitivity to functional form is an issue. If the researcher's goal is to estimate

$$\text{ATT}_h := E_u(h(Y_2(1)) - h(Y_2(0)) | D = 1)$$

for any strictly increasing function $h(\cdot)$, then parallel trends has to hold for any $h(\cdot)$, and sensitivity to functional form is an issue. Such an estimation goal, while much more ambitious

than just trying to estimate ATT and/or ATT_{\ln} , may be justified: as will become clear later, meeting this goal is one way to ensure that one can estimate all quantile treatment effects, namely how the treatment affects the entire distribution of treated groups' period-two outcome, rather than just their average outcome. Sensitivity to functional form is also an issue if the researcher is indifferent between several functional forms, but pre-trend tests cannot detect functional forms for which parallel trends fails, either because they lack power, or because one is not ready to assume that parallel trends prior to treatment imply parallel trends post treatment. One may also worry that pre-testing for the right functional form could bias the DID estimator with the functional form selected at the outset of the pre-tests, and could distort inference. Investigating whether this is a legitimate concern in realistic settings is an interesting avenue for future research.

3.7.3 Estimating quantile treatment effects*

Defining quantiles and quantile treatment effects. For any $(d, t) \in \{0, 1\} \times \{1, 2\}$ and for any generic variable X , let $x \mapsto F_{X_t|D=d}(x)$ denote the cdf of X in treatment group d at time t . For any weakly increasing function f let f^{-1} denote its generalized inverse, $f^{-1}(x) = \inf\{y : f(y) \geq x\}$. Then, for any $\tau \in [0, 1]$, $F_{X_t|D=d}^{-1}(\tau)$ is the quantile of order τ of X in treatment group d at time t . For instance, $F_{X_t|D=d}^{-1}(0.5)$ is the median of X in treatment group d at time t . Finally, let

$$\text{QTE}(\tau) = F_{Y_2(1)|D=1}^{-1}(\tau) - F_{Y_2(0)|D=1}^{-1}(\tau)$$

denote the quantile treatment effect (QTE) of order τ in the treatment group at period two. For instance, $\text{QTE}(0.5) = F_{Y_2(1)|D=1}^{-1}(0.5) - F_{Y_2(0)|D=1}^{-1}(0.5)$ is the difference between the median outcome in the treatment group at period two with and without treatment. One has

$$\text{ATT} = \int_0^1 \text{QTE}(\tau) d\tau :$$

the ATT is the integral of the QTE function.

Estimating QTEs under a parallel-trends assumption on the cdf. (3.39) shows that $y \mapsto F_{Y_2(0)|D=1}(y)$, the cdf of the treated group's period-two untreated outcome, is identified

under the cdf parallel-trends condition in (3.38). As a variable's quantile function is just the generalized inverse of its cdf, if $y \mapsto F_{Y_1(0)|D=1}(y) + F_{Y_2(0)|D=0}(y) - F_{Y_1(0)|D=0}(y)$ is weakly increasing then it follows from (3.39) that

$$F_{Y_2(0)|D=1}^{-1}(\tau) = \left(F_{Y_1|D=1} + F_{Y_2|D=0} - F_{Y_1|D=0} \right)^{-1}(\tau) :$$

all quantiles of $Y_2(0)$ in the treatment group are identified. Therefore,

$$\text{QTE}(\tau) = F_{Y_2|D=1}^{-1}(\tau) - \left(F_{Y_1|D=1} + F_{Y_2|D=0} - F_{Y_1|D=0} \right)^{-1}(\tau) : \quad (3.40)$$

QTEs are identified under (3.38). Kim and Wooldridge (2024) propose to use

$$\widehat{\text{QTE}}^{\text{cdf}}(\tau) = \widehat{F}_{Y_2|D=1}^{-1}(\tau) - \left(\widehat{F}_{Y_1|D=1} + \widehat{F}_{Y_2|D=0} - \widehat{F}_{Y_1|D=0} \right)^{-1}(\tau)$$

to estimate the QTEs. Those estimators are computed by the `ddid` Stata package.

Estimating QTEs under a parallel-trends assumption on the quantile function. Another popular DID-like approach to estimate QTEs is the quantile DID estimator, initially proposed by Poterba, Venti and Wise (1995) and Meyer et al. (1995), where one applies DID to each quantile: to estimate $\text{QTE}(\tau)$, one compares the evolution of the quantile of order τ in the treatment and control groups:

$$\widehat{\text{QTE}}^{\text{quant}}(\tau) = \widehat{F}_{Y_2|D=1}^{-1}(\tau) - \widehat{F}_{Y_1|D=1}^{-1}(\tau) - \left(\widehat{F}_{Y_2|D=0}^{-1}(\tau) - \widehat{F}_{Y_1|D=0}^{-1}(\tau) \right).$$

To compute $\widehat{\text{QTE}}^{\text{quant}}(\tau)$, one can just run a quantile regression of order τ of the outcome on a treatment group indicator, a period-two indicator, and the interaction of the two indicators. The coefficient on the interaction is equal to $\widehat{\text{QTE}}^{\text{quant}}(\tau)$. For a thorough exposition of quantile regressions, see Koenker (2005). Quantile regressions can be performed using the Stata commands `qreg` and `sqreg` or the package `qreg2`, or using the R command `quantreg`. The identifying assumption rationalizing the quantile DID estimator is a parallel-trends assumption on the untreated outcome's quantiles:

$$F_{Y_2(0)|D=1}^{-1}(\tau) - F_{Y_1(0)|D=1}^{-1}(\tau) = F_{Y_2(0)|D=0}^{-1}(\tau) - F_{Y_1(0)|D=0}^{-1}(\tau). \quad (3.41)$$

Unlike (3.38), (3.41) is not invariant to functional form. On the other hand, (3.38) needs to hold for all y . Then, one can estimate the full QTE function $\tau \mapsto \text{QTE}(\tau)$, but that estimation relies

on an assumption on the evolution of the entire distribution of the untreated outcome. Instead, one may only impose (3.41) for specific quantiles (e.g. the median). Then, one can only estimate specific QTEs, but now estimation only relies on restrictions on the evolution of specific features of the distribution of the untreated outcome, in the spirit of the standard DID estimator that only imposes restrictions on the evolution of the average. An additional advantage of only using the quantile DID estimator at specific quantiles is that then, pre-trend tests are straightforward: they do not involve functionals, just vectors of quantile regression coefficients similar to the aforementioned one, but estimated in the pre-treatment periods.

Estimating QTEs with a Changes-in-Changes (CIC) estimator. A last popular DID-like estimator of QTEs is the CIC estimator of Athey and Imbens (2006):

$$\widehat{\text{QTE}}^{\text{cic}}(\tau) = \widehat{F}_{Y_2|D=1}^{-1}(\tau) - \widehat{F}_{Y_2|D=0}^{-1}\left(\widehat{F}_{Y_1|D=0}\left(\widehat{F}_{Y_1|D=1}^{-1}(\tau)\right)\right).$$

The identifying assumptions rationalizing that estimator is that for all $t \in \{1, 2\}$, $Y_t(0) = h_t(U_t)$ for a strictly increasing function $h_t(\cdot)$, and for all $d \in \{0, 1\}$ $U_2|D=d \sim U_1|D=d$, where \sim denotes equality in distribution. Under those assumptions, the τ th quantile of $Y_t(0)$ may not follow the same evolution in the treatment and in the control group. On the other hand, those assumptions ensure that if τ and τ' are such that at period one, the τ th quantile in the treatment group corresponds to the τ' th quantile in the control group, then without treatment the period-one-to-two change in the treatment group's τ th quantile would have been the same as that of the control group's τ' th quantile. Then, under the CIC assumptions, one has parallel-trends between the treatment and control group quantiles, not for the same quantile order τ but for quantile orders τ and τ' for which treated and control groups' outcomes are equal at period one. Like (3.38), the identifying assumptions underlying the CIC estimator are invariant to functional form, but the CIC assumptions are not parallel-trends assumptions: for instance they do not imply that the average untreated outcome of the treated and control groups are on parallel trends. Like (3.38), the CIC assumptions impose restrictions on the evolution of the untreated outcome's entire distribution. The CIC estimators are computed by the `fuzzydid` (de Chaisemartin, D'Haultfœuille and Guyonvarch, 2019) and `cic` (Kranker, 2019) Stata packages, and by the `qte` (Callaway, 2023) R package.

3.8 Instrumental-variable DID estimators*

Motivation. There are instances where one may prefer making a parallel-trends assumption with respect to an instrument rather than the treatment. For instance, one may be interested in estimating the price-elasticity of a good. If prices respond to demand shocks, the counterfactual consumption trends of units experiencing and not experiencing a price change may not be the same, so a parallel-trends assumption with respect to prices may not hold. But instead, one can make a parallel-trends assumption with respect to taxes, used as an instrument for prices.

Set-up. In this section, we assume that $T = 2$ and $T_0 = T_1 = 1$. Then, because groups can be treated for at most one period of time, Assumption ND can be imposed without loss of generality. Also, we adopt a sampling-based perspective, where the G groups we observe are an independent and identically distributed sample from a larger population, as assumed in Assumption IID. Then, the g subscript can be dropped, and the design is no longer conditioned upon. In line with the focus of this chapter, we restrict attention to classical instrumental-variable (IV) DID designs, with a binary instrument Z_t , such that a subset of units receive the instrument at period two: $Z_t = 1\{t = 2\}Z$, where Z is an indicator for the “instrumented” group. Finally, we assume that treatment is binary, and we let $D_t(1)$ and $D_t(0)$ denote a unit’s potential treatments at period t with and without the instrument.

Assumptions. de Chaisemartin (2010) considers the following assumptions:

$$E_u [Y_2(D_2(0)) - Y_1(D_1(0))|Z = 1] = E_u [Y_2(D_2(0)) - Y_1(D_1(0))|Z = 0], \quad (3.42)$$

$$E_u [D_2(0) - D_1(0)|Z = 1] = E_u [D_2(0) - D_1(0)|Z = 0], \quad (3.43)$$

and

$$D_2(1) \geq D_2(0). \quad (3.44)$$

(3.42) requires that on average, instrumented and uninstrumented groups have the same outcome evolutions from period one to two, in the counterfactual where the instrumented group does not receive the instrument. (3.42) is a “reduced-form” version of the parallel-trends condition in (2.10). (3.43) requires that on average, instrumented and uninstrumented groups have the same

treatment evolutions from period one to two, in the counterfactual where the instrumented group does not receive the instrument. Finally, (3.44) is a monotonicity condition, analogous to that introduced by Imbens and Angrist (1994). It requires that receiving the instrument cannot decrease groups' treatment.

Identification and estimation. Proposition 1 in de Chaisemartin (2010) shows that if (3.42), (3.43), and (3.44) hold, then

$$\begin{aligned} & E_u(Y_2(1) - Y_2(0)|D_2(1) > D_2(0), Z = 1) \\ &= \frac{E_u(Y_2 - Y_1|Z = 1) - E_u(Y_2 - Y_1|Z = 0)}{E_u(D_2 - D_1|Z = 1) - E_u(D_2 - D_1|Z = 0)} : \end{aligned} \quad (3.45)$$

$E_u(Y_2(1) - Y_2(0)|D_2(1) > D_2(0), Z = 1)$, the average treatment effect at period two across compliers in the instrumented group, is identified by a so-called Wald-DID, a ratio whose numerator compares the outcome evolutions of the instrumented and uninstrumented groups, while its denominator compares the treatment evolutions of the two groups. See also Hudson et al. (2015) for a closely-related result. To estimate the Wald-DID, one can replace expectations by sample averages. Equivalently, one can use a so-called 2SLS-TWFE regression of $Y_{g,t}$ on an intercept, Z_g , $1\{t = 2\}$, and $D_{g,t}$, using $1\{t = 2\}Z_g$ as the excluded instrument for $D_{g,t}$.

Reduced-form parallel-trends restricts treatment-effect heterogeneity. Assume that the instrumented and uninstrumented groups have the same untreated-outcome evolutions:

$$E_u[Y_2(0) - Y_1(0)|Z = 1] = E_u[Y_2(0) - Y_1(0)|Z = 0]. \quad (3.46)$$

After some algebra, (3.42) and (3.46) imply that

$$\begin{aligned} & E_u[Y_2(1) - Y_2(0) - (Y_1(1) - Y_1(0))|D_1 = 1, Z = 1] P_u(D_1 = 1|Z = 1) \\ &+ E_u[Y_2(1) - Y_2(0)|D_2(0) - D_1(0) = 1, Z = 1] P_u(D_2(0) - D_1(0) = 1|Z = 1) \\ &- E_u[Y_2(1) - Y_2(0)|D_2(0) - D_1(0) = -1, Z = 1] P_u(D_2(0) - D_1(0) = -1|Z = 1) \\ &= E_u[Y_2(1) - Y_2(0) - (Y_1(1) - Y_1(0))|D_1 = 1, Z = 0] P_u(D_1 = 1|Z = 0) \\ &+ E_u[Y_2(1) - Y_2(0)|D_2(0) - D_1(0) = 1, Z = 0] P_u(D_2(0) - D_1(0) = 1|Z = 0) \\ &- E_u[Y_2(1) - Y_2(0)|D_2(0) - D_1(0) = -1, Z = 0] P_u(D_2(0) - D_1(0) = -1|Z = 0). \end{aligned} \quad (3.47)$$

Under parallel trends on the first stage, the conditions below are sufficient for (3.47) to hold:

$$\forall z \in \{0, 1\}, E_u [Y_2(1) - Y_2(0)|D_1 = 1, Z = z] = E_u [Y_1(1) - Y_1(0)|D_1 = 1, Z = z] \quad (3.48)$$

$$\begin{aligned} \forall (z, z', \delta, \delta') \in \{0, 1\}^2 \times \{-1, 1\}^2, & E_u [Y_2(1) - Y_2(0)|D_2(0) - D_1(0) = \delta, Z = z] \\ & = E_u [Y_2(1) - Y_2(0)|D_2(0) - D_1(0) = \delta', Z = z']. \end{aligned} \quad (3.49)$$

(3.48) requires that the LATE of groups treated at period one is constant over time, both in the instrumented and uninstrumented groups. (3.49) requires that the LATE of groups naturally switching into treatment over time ($D_2(0) - D_1(0) = 1$) is the same as the LATE of groups switching out ($D_2(0) - D_1(0) = -1$), and that those LATEs are the same in the instrumented and uninstrumented groups. Conversely, (3.47) may fail if (3.48) or (3.49) fails. Thus, when combined with (3.46), (3.42) may fail if the treatment effect of groups treated at period one changes over time, or if switchers in and out have heterogeneous effects, or if switchers have heterogeneous effects in the instrumented and uninstrumented groups. Then, to have that (3.42) does not restrict effects' heterogeneity, (3.46) has to fail: groups have to be on parallel trends in the counterfactual where they do not receive the instrument, but they have to experience differential trends in the counterfactual where they remain untreated. Such a scenario might be hard to rationalize, so we view reduced-form parallel-trends as restricting effects' heterogeneity.

Is IV-DID “as credible as” DID? (3.45) shows that the IV-DID estimand relies on two parallel-trends assumptions, that are both partly testable via pre-trends test when one has data from several pre periods where no one receives the instrument. However, as discussed earlier in this chapter, pre-trends tests are only suggestive tests of parallel trends and are sometimes underpowered. Moreover, the previous paragraph also shows that the reduced-form parallel-trends assumption imposes restrictions on treatment-effect heterogeneity that may often be implausible. By contrast, the standard parallel-trends assumption underlying the DID estimand does not impose such restrictions. This leads us to consider IV-DID as a less credible research design than DID. In Chapter 8 (see Section 8.4.8.5), we will consider an alternative reduced-form parallel-trends assumption, which imposes less restrictions on treatment-effect heterogeneity than (3.42).

Bibliographic notes. The realization that the “reduced-form” parallel-trends assumption in de Chaisemartin (2010) and Hudson et al. (2015) actually restricts effect heterogeneity explains the apparent discrepancy between those two papers and Blundell and Costa-Dias (2009) and de Chaisemartin and D’Haultfœuille (2018), two papers that show that IV-DID estimators rely on restrictions on effects’ heterogeneity under the standard parallel-trends assumption on the untreated outcome.

3.9 Further topics*

3.9.1 Imbalanced panels

Naive TWFE ES estimators. In many applications, the data is actually an imbalanced panel, meaning that there are cells (g, t) for which $Y_{g,t}$ is missing, either because the variable could not be collected, or because group g does not exist at time t . Let $O_{g,t}$ be an indicator for (g, t) cells for which $Y_{g,t}$ is observed, which we treat as a non-stochastic/conditioned upon quantity. Let $G_{1,t}$ and $G_{0,t}$ respectively denote the number of groups such that $D_g = 1, O_{g,t} = 1$ and $D_g = 0, O_{g,t} = 1$. Then, let

$$\widetilde{\text{ATT}}_\ell^{\text{imb}} = \frac{1}{G_{1,T_0+\ell}} \sum_{g:D_g=1, O_{g,T_0+\ell}=1} \text{TE}_{g,T_0+\ell}$$

be the analogue of ATT_ℓ , for the subsample of treatment groups observed at $T_0 + \ell$. If we ignore missingness, we estimate a TWFE ES regression in the subsample of (g, t) cells for which $Y_{g,t}$ is observed. One can show that the coefficient on $1\{t = T_0 + \ell\}D_g$ in this naive regression is equal to

$$\begin{aligned} & \frac{1}{G_{1,T_0+\ell}} \sum_{g:D_g=1, O_{g,T_0+\ell}=1} Y_{g,T_0+\ell} - \frac{1}{G_{1,T_0}} \sum_{g:D_g=1, O_{g,T_0}=1} Y_{g,T_0} \\ & - \left(\frac{1}{G_{0,T_0+\ell}} \sum_{g:D_g=0, O_{g,T_0+\ell}=1} Y_{g,T_0+\ell} - \frac{1}{G_{0,T_0}} \sum_{g:D_g=0, O_{g,T_0}=1} Y_{g,T_0} \right). \end{aligned} \quad (3.50)$$

Assume that the following condition holds:

$$\begin{aligned} & E \left[\frac{1}{G_{1,T_0+\ell}} \sum_{g:D_g=1, O_{g,T_0+\ell}=1} Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - \frac{1}{G_{1,T_0}} \sum_{g:D_g=1, O_{g,T_0}=1} Y_{g,T_0}(\mathbf{0}_{T_0}) \right] \\ & = E \left[\frac{1}{G_{0,T_0+\ell}} \sum_{g:D_g=0, O_{g,T_0+\ell}=1} Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - \frac{1}{G_{0,T_0}} \sum_{g:D_g=0, O_{g,T_0}=1} Y_{g,T_0}(\mathbf{0}_{T_0}) \right]. \end{aligned} \quad (3.51)$$

Under the parallel-trends assumption in (3.51), the estimator in (3.50) is unbiased for $\widetilde{\text{ATT}}_\ell^{\text{imb}}$. However, (3.51) is not an easy-to-rationalize assumption: one could have that Assumption PT holds, yet (3.51) fails. Intuitively, this is because the average trends in the left- and right-hand sides of (3.51) mix time trends on the untreated outcomes with composition effects, as missing groups differ between periods T_0 and $T_0 + \ell$.

Alternative estimators. For any $\ell \in \{1, \dots, T_1\}$, let D_g^ℓ be an indicator equal to 1 if $D_g = 1, O_{g,T_0} = 1, O_{g,T_0+\ell} = 1$, to 0 if $D_g = 0, O_{g,T_0} = 1, O_{g,T_0+\ell} = 1$, and missing otherwise, and let $G_{1,\ell}$ and $G_{0,\ell}$ respectively denote the number of groups such that $D_g^\ell = 1$ and $D_g^\ell = 0$. Then, let

$$\text{ATT}_\ell^{\text{imb}} = \frac{1}{G_{1,\ell}} \sum_{g:D_g^\ell=1} E [Y_{g,T_0+\ell}(\mathbf{0}_{T_0}, \mathbf{1}_\ell) - Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell})]$$

be the analogue of ATT_ℓ , for the subsample of treatment groups observed at T_0 and $T_0 + \ell$. Similarly, let

$$\widehat{\beta}_\ell^{\text{fe,imb}} = \frac{1}{G_{1,\ell}} \sum_{g:D_g^\ell=1} (Y_{g,T_0+\ell} - Y_{g,T_0}) - \frac{1}{G_{0,\ell}} \sum_{D_g^\ell=0} (Y_{g,T_0+\ell} - Y_{g,T_0})$$

be a DID estimator comparing the T_0 to $T_0 + \ell$ outcome evolution of treatment and control groups observed both at T_0 and $T_0 + \ell$, thus avoiding composition effects. That estimator can easily be obtained from a regression of $Y_{g,t}$ on an intercept, D_g , and $1\{t = T_0 + \ell\}D_g$, restricting the sample to $t = T_0$ or $t = T_0 + \ell$ and groups for which Y_{g,T_0} and $Y_{g,T_0+\ell}$ are both observed.

Assume that the following condition holds:

$$\begin{aligned} & E \left[\frac{1}{G_{1,\ell}} \sum_{g:D_g^\ell=1} (Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0}(\mathbf{0}_{T_0})) \right] \\ & = E \left[\frac{1}{G_{0,\ell}} \sum_{D_g^\ell=0} (Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - Y_{g,T_0})(\mathbf{0}_{T_0}) \right]. \end{aligned} \quad (3.52)$$

$\widehat{\beta}_\ell^{\text{fe,imb}}$ is unbiased for $\text{ATT}_\ell^{\text{imb}}$ under (3.52). (3.52) is easier to rationalize than (3.51): if Assumption PT holds, then (3.52) also holds, so (3.52) is weaker than Assumption PT. Note that to estimate the effect of ℓ periods of exposure to treatment among groups observed at period $T_0 + \ell$ but not at period T_0 , one can compute a DID estimator comparing the $T_0 - 1$ to $T_0 + \ell$ outcome evolutions of treated and control groups, in the subsample observed at $T_0 - 1$ and $T_0 + \ell$ but not at T_0 . One can also compute a DID estimator comparing the $T_0 - 2$ to $T_0 + \ell$ outcome evolutions of treated and control groups, in the subsample observed at $T_0 - 2$ and $T_0 + \ell$ but not at T_0 and $T_0 - 1$, etc. Then, one can compute a weighted average of $\widehat{\beta}_\ell^{\text{fe,imb}}$ and of all those DIDs, with weights proportional to the number of treated groups in each DID. This yields an estimator of the effect of ℓ periods of exposure among all treatment groups observed at period $T_0 + \ell$ and at least at one period strictly before $T_0 + 1$.

Recommendation. We recommend using $\widehat{\beta}_\ell^{\text{fe,imb}}$. The assumption underlying that estimator, (3.52), seems more plausible to us than (3.51), the assumption underlying naive TWFE ES estimators. Moreover, (3.52) seems more in line with a standard parallel-trends assumption than (3.51): (3.51) could fail even if Assumption PT holds.

3.9.2 Weighting

Weighted TWFE regressions. In many applications, researchers weight their TWFE regression. One can show that the coefficient on $1\{t = T_0 + \ell\}$ in the TWFE ES regression in (3.6) weighted by $W_{g,t}$ is equal to

$$\widehat{\beta}_\ell^{\text{fe},W} := \sum_{g:D_g=1} W_{g,T_0+\ell}^n Y_{g,T_0+\ell} - \sum_{g:D_g=1} W_{g,T_0}^n Y_{g,T_0} - \left(\sum_{g:D_g=0} W_{g,T_0+\ell}^n Y_{g,T_0+\ell} - \sum_{g:D_g=0} W_{g,T_0}^n Y_{g,T_0} \right),$$

where $W_{g,t}^n = W_{g,t} / \sum_{g':D_{g'}=D_g} W_{g',t}$ are normalized weights summing to one at each date in the treatment and in the control group. Under the parallel-trends condition in Assumption PT, do we have that $\widehat{\beta}_\ell^{\text{fe},W}$ is unbiased for $\text{ATT}_\ell^W := \sum_{g:D_g=1} W_{g,T_0+\ell}^n \text{TE}_{g,\ell}^r$, a weighted version of ATT_ℓ ?

If the weights are time-varying, $\hat{\beta}_\ell^{\text{fe},W}$ may not be unbiased for ATT_ℓ^W under the parallel-trends condition in Assumption PT. Intuitively, this is because we may have $W_{g,T_0+\ell}^n \neq W_{g,T_0}^n$, and therefore $\hat{\beta}_\ell^{\text{fe},W}$ does not compare a weighted average of outcome evolutions $Y_{g,T_0+\ell} - Y_{g,T_0}$ in the treatment and in the control group. Still, $\hat{\beta}_\ell^{\text{fe},W}$ is unbiased for ATT_ℓ^W under a “parallel trends of a weighted average” assumption:

$$\begin{aligned} & E \left[\sum_{g:D_g=1} W_{g,T_0+\ell}^n Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - \sum_{g:D_g=1} W_{g,T_0}^n Y_{g,T_0}(\mathbf{0}_{T_0}) \right] \\ & = E \left[\sum_{g:D_g=0} W_{g,T_0+\ell}^n Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - \sum_{g:D_g=0} W_{g,T_0}^n Y_{g,T_0}(\mathbf{0}_{T_0}) \right]. \end{aligned} \quad (3.53)$$

On the other hand, if the weights are time-invariant, $\hat{\beta}_\ell^{\text{fe},W}$ does compare a weighted average of outcome evolutions $Y_{g,T_0+\ell} - Y_{g,T_0}$ in the treatment and in the control group, and it is unbiased for ATT_ℓ^W under Assumption PT. Overall, weighting changes the effect that the regression estimates, and it may change the identifying assumption underlying it if the weights are time-varying.

Two arguments in favor of weighting. To fix ideas, suppose one has a panel data of Malian communes, and one considers weighting the regression by the population of commune g at period t . In that context, interpret ATT_ℓ and ATT_ℓ^W . Is one of these two parameters more “interesting” than the other?

ATT_ℓ is the average effect of ℓ periods of exposure to treatment across Malian communes, where each commune receives the same weight. ATT_ℓ^W is the average effect of ℓ periods of exposure to treatment across the Malian population, because each commune is weighted by its population. In as much as we care more about people than communes, ATT_ℓ^W is more interesting than ATT_ℓ . This would argue in favor of weighting the regression. Another argument in favor of weighting is when survey data is used, and survey weights correspond to the inverse of the probability to appear in the sample. Then, not using such weights may lead to a selection bias, as shown by Wooldridge (2007) and Solon, Haider and Wooldridge (2015).

A cautionary note on weighting. Weighting also has one drawback: it has been shown, theoretically and through simulations, that with weighted regressions, confidence intervals relying on asymptotic approximations need larger sample sizes to be reliable (see e.g. Cameron and Miller, 2015; Carter, Schnepel and Steigerwald, 2017). This is especially true when a few groups receive a very large weight, as is likely to be the case when one weights by population. Then, a few groups have a disproportionate amount of influence on the regression coefficients, and the regression's effective sample size is much smaller than its nominal sample size. Carter et al. (2017) propose a measure of the regression's effective sample size G^* , computed by the `clusteff` Stata package (Lee and Steigerwald, 2018). They recommend relying on asymptotic approximations only if G^* is larger than 50. As we saw earlier in this chapter, in the DID context the reliability of asymptotic approximations depends on the numbers of treated and control groups, not just on the total number of groups. To our knowledge, measures of the effective number of treated and control groups have not been proposed yet. Doing so is an interesting avenue for future research.

When the weights are affected by the treatment, we recommend using time-invariant weights, determined prior to treatment. Sometimes, the weighting variable may be affected by the treatment. Typically, $W_{g,t}$ is the population of group g at time t , and the treatment may affect that population, something one can assess by estimating an unweighted TWFE ES regression, with $W_{g,t}$ as the outcome variable. When $W_{g,t}$ is affected by the treatment, (3.53) might be implausible. Let $W_{g,t}^n(0)$ and $W_{g,t}^n(1)$ denote counterfactual values of $W_{g,t}^n$ without and with treatment. To rationalize (3.53), one for instance needs to assume that

$$\begin{aligned} E \left[\sum_{g:D_g=1} W_{g,T_0+\ell}^n(0) Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - \sum_{g:D_g=1} W_{g,T_0}^n(0) Y_{g,T_0}(\mathbf{0}_{T_0}) \right] \\ = E \left[\sum_{g:D_g=0} W_{g,T_0+\ell}^n(0) Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) - \sum_{g:D_g=0} W_{g,T_0}^n(0) Y_{g,T_0}(\mathbf{0}_{T_0}) \right], \end{aligned} \quad (3.54)$$

$$E \left[\sum_{g:D_g=1} (W_{g,T_0+\ell}^n(1) - W_{g,T_0+\ell}^n(0)) Y_{g,T_0+\ell}(\mathbf{0}_{T_0+\ell}) \right] = 0. \quad (3.55)$$

While (3.54) can be placebo tested by computing weighted pre-trend estimators, (3.55) cannot be placebo tested, thus making (3.53) not fully placebo-testable, and at odds with standard DID logic. Therefore, when the weights are affected by the treatment, we recommend weighting the

regression by time-invariant weights determined before treatment. For instance, one can weight the regression by $W_{g,1}$, or by W_{g,T_0} .

3.9.3 Accounting for and estimating spillover effects

Geographic spillovers to control groups close to treated ones. There are many applications where the effect of the treatment might spill over onto untreated groups geographically close to a treated group. For example, if a county increases its minimum wage, individuals from contiguous counties may decide to go work there to benefit from the higher minimum wage. Similarly, if a coal-fired power plant adopts an emissions-control technology, this will reduce air pollution in the county where the plant is located, but as air pollution travels this can also reduce air pollution in neighboring counties downwind of the treated one. This leads to a violation of the SUTVA assumption. Then, a treated unit may experience both a direct treatment effect, arising from its own treatment, but it may also experience an indirect effect, arising from the treatment of treated units located close to it. The sum of such direct and indirect effects is sometimes referred to as the treatment’s total effect. If a parallel-trends assumption holds but SUTVA fails, Clarke (2017) and Butts (2021c) show that a TWFE regression that ignores such violations will not estimate the average total effect of the treatment across all treated groups. Instead, this regression will estimate the average total effect, minus the proportion of untreated groups that are affected by treated groups’ treatment, times the average indirect effect of the treatment across those “affected” untreated groups. Assuming that both effects are of the same sign but that the indirect effect is closer to zero than the total effect, the TWFE regression suffers from an attenuation bias. If one is ready to assume that a subset of untreated groups is not affected by treatment groups’ treatment, for instance because they are geographically far to all treated groups, then the population can be partitioned into treated groups, affected untreated groups, and unaffected untreated groups. Under a parallel-trends assumption, Butts (2021c) shows that a TWFE regression restricting the sample to treated and to unaffected untreated groups estimates the average total effect of the treatment across all treated groups. This rationalizes the common approach in applied work of excluding groups located close to the treatment area. Then, Butts (2021c) also shows that a TWFE regression restricting the sample to affected

untreated groups and to unaffected untreated groups estimates the average indirect effect of the treatment across all affected untreated groups.

Market spillovers to substitutes or complement products. There are also many applications where a treatment affects a product A, and the treatment's effect spills over onto an untreated product B that is a substitute or a complement of A. For instance, product A might start receiving a subsidy, thus leading producers to decrease its price. This may reduce demand for and prices of substitute product B. A commonly used strategy in such instances is to move the analysis at the market rather than at the product level, though this also shifts the research question to measuring the aggregate, market-level effects of the subsidy. If some untreated markets are observed and one can reasonably rule out spillovers across markets, then to estimate the treatment's direct effect one can use a DID estimator comparing the evolution of demand for and/or price of product A in treated and untreated markets. Similarly, to estimate the spillover effect one can use a DID estimator comparing the evolution of demand for and/or price of product B in treated and untreated markets.

3.10 Next steps

de Chaisemartin and D'Haultfoeuille (2025) reviewed the 100 most cited papers published by the *American Economic Review* from 2015 to 2019, and found that 26 have estimated at least one TWFE regression. Of those 26 papers, only two have an absorbing and binary treatment with no variation in treatment timing. This suggests that classical DID designs are more the exception than the norm, and the results we have seen so far only apply to a minority of the research designs encountered by social scientists analyzing natural experiments. Accordingly, in the following chapters we will consider more complex designs. We will show that in such designs, TWFE regressions rely on much stronger assumptions than in the classical design, which will lead us to consider more robust estimators. But before considering more complex designs, in the next chapter we consider relaxations of the parallel-trends assumption in the classical design.

3.11 Appendix*

3.11.1 Caveats in interpreting $\ell \mapsto \text{ATT}_\ell$: further details

One has

$$\begin{aligned}\text{ATT}_2 &= \frac{1}{G_1} \sum_{g:D_g=1} E[Y_{g,T_0+2}(\mathbf{0}_{T_0}, 1, 1) - Y_{g,T_0+2}(\mathbf{0}_{T_0+2})] \\ &= \frac{1}{G_1} \sum_{g:D_g=1} E[Y_{g,T_0+2}(\mathbf{0}_{T_0}, 0, 1) - Y_{g,T_0+2}(\mathbf{0}_{T_0+2})] \\ &\quad + \frac{1}{G_1} \sum_{g:D_g=1} E[Y_{g,T_0+2}(\mathbf{0}_{T_0}, 1, 1) - Y_{g,T_0+2}(\mathbf{0}_{T_0}, 0, 1)],\end{aligned}$$

while

$$\text{ATT}_1 = \frac{1}{G_1} \sum_{g:D_g=1} E[Y_{g,T_0+1}(\mathbf{0}_{T_0}, 1) - Y_{g,T_0+1}(\mathbf{0}_{T_0+1})].$$

Therefore, $\text{ATT}_2 > \text{ATT}_1$ implies that either

$$\frac{1}{G_1} \sum_{g:D_g=1} E[Y_{g,T_0+2}(\mathbf{0}_{T_0}, 1, 1) - Y_{g,T_0+2}(\mathbf{0}_{T_0}, 0, 1)] > 0,$$

meaning that being treated for two periods produces a larger effect than being treated for one period, or

$$\frac{1}{G_1} \sum_{g:D_g=1} E[Y_{g,T_0+2}(\mathbf{0}_{T_0}, 0, 1) - Y_{g,T_0+2}(\mathbf{0}_{T_0+2})] > \frac{1}{G_1} \sum_{g:D_g=1} E[Y_{g,T_0+1}(\mathbf{0}_{T_0}, 1) - Y_{g,T_0+1}(\mathbf{0}_{T_0+1})],$$

meaning that the effect of being treated for one period is larger at $T_0 + 2$ than at $T_0 + 1$.

3.11.2 The Frisch-Waugh-Lovell theorem

We do not give here the most general statement of this theorem, but this version is sufficient for our purposes. The second equality below, which is less commonly presented, is used at several places in the book.

Theorem 4 *Let $\hat{\theta}_1$ be the coefficient on a scalar variable $X_{i,1}$, in an ordinary least squares regression of a scalar variable $(Y_i)_{1 \leq i \leq n}$ on a vector of $K+1$ variables $(1, X_{i,1}, X_{i,2}, \dots, X_{i,K})_{1 \leq i \leq n}$.*

One has that

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n \hat{u}_i Y_i}{\sum_{i=1}^n \hat{u}_i^2} = \frac{\sum_{i=1}^n \hat{u}_i Y_i}{\sum_{i=1}^n \hat{u}_i X_{i,1}},$$

where \hat{u}_i is the residual from a regression of $(X_{i,1})_{1 \leq i \leq n}$ on $(1, X_{i,2}, \dots, X_{i,K})_{1 \leq i \leq n}$.

Proof. Let $\hat{\theta} = (\hat{\theta}_0, \dots, \hat{\theta}_K)'$ denote the OLS estimator of the regression of $(Y_i)_{1 \leq i \leq n}$ on $(X_i)_{1 \leq i \leq n}$, with $X_i := (1, \dots, X_{i,K})'$ and $(\hat{\varepsilon}_i)_{1 \leq i \leq n}$ be the corresponding residuals. Then,

$$Y_i = \hat{\theta}_0 + X_{i,1}\hat{\theta}_1 + \sum_{j=2}^K X_{i,j}\hat{\theta}_j + \hat{\varepsilon}_i. \quad (3.56)$$

As a residual, \hat{u}_i is uncorrelated with the regressors $(1, X_{i,2}, \dots, X_{i,K})'$. Hence, for all $j = 2, \dots, K$,

$$\sum_{i=1}^n \hat{u}_i = \sum_{i=1}^n \hat{u}_i X_{i,j} = 0. \quad (3.57)$$

Also as a residual, $\hat{\varepsilon}_i$ is uncorrelated with any linear combination of the regressors X_i . Since \hat{u}_i is one such combination,

$$\sum_{i=1}^n \hat{u}_i \hat{\varepsilon}_i = 0. \quad (3.58)$$

Combining (3.56)-(3.58), we obtain

$$\sum_{i=1}^n \hat{u}_i Y_i = \sum_{i=1}^n \hat{u}_i X_{i,1} \hat{\theta}_1.$$

The second equality of the theorem follows. To obtain the first, let $\widehat{X}_{i,1}$ denote the prediction of $X_{i,1}$ from the regression of $(X_{i,1})_{1 \leq i \leq n}$ on $(1, X_{i,2}, \dots, X_{i,K})_{1 \leq i \leq n}$, so that $X_{i,1} = \widehat{X}_{i,1} + \hat{u}_i$. Since $\widehat{X}_{i,1}$ is a linear combination of $(1, X_{i,2}, \dots, X_{i,K})$, we obtain, from (3.57),

$$\sum_{i=1}^n \hat{u}_i X_{i,1} = \sum_{i=1}^n \hat{u}_i (\widehat{X}_{i,1} + \hat{u}_i) = \sum_{i=1}^n \hat{u}_i^2.$$

The first equality of the theorem follows.

3.11.3 Variance estimation with independent but non-identically distributed variables

We prove (3.15), by proving that for any independent but not necessarily identically distributed variables $(U_g)_{g=1,\dots,G}$,

$$E \left[\frac{1}{G-1} \sum_{g=1}^G (U_g - \bar{U})^2 \right] = \frac{1}{G} \sum_{g=1}^G V(U_g) + \frac{1}{G-1} \sum_{g=1}^G (E(U_g) - E(\bar{U}))^2, \quad (3.59)$$

where $\bar{U} = \frac{1}{G} \sum_{g=1}^G U_g$. First, we have

$$\begin{aligned} \frac{1}{G-1} \sum_{g=1}^G (U_g - \bar{U})^2 &= \frac{1}{G-1} \sum_{g=1}^G [U_g - E(\bar{U}) + E(\bar{U}) - \bar{U}]^2 \\ &= \frac{1}{G-1} \sum_{g=1}^G [U_g - E(\bar{U})]^2 - \frac{G}{G-1} [\bar{U} - E(\bar{U})]^2. \end{aligned} \quad (3.60)$$

Moreover, by independence of the $(U_g)_{g=1,\dots,G}$,

$$\frac{G}{G-1} E \left[(\bar{U} - E(\bar{U}))^2 \right] = \frac{G}{G-1} V(\bar{U}) = \frac{1}{G(G-1)} \sum_{g=1}^G V(U_g). \quad (3.61)$$

Besides,

$$\begin{aligned} E[(U_g - E(\bar{U}))^2] &= E[(U_g - E(U_g))^2] + (E(U_g) - E(\bar{U}))^2 + 2(E(U_g) - E(\bar{U}))E[U_g - E(U_g)] \\ &= V(U_g) + (E(U_g) - E(\bar{U}))^2. \end{aligned} \quad (3.62)$$

Then, taking expectations in (3.60) and plugging in (3.61)-(3.62), we obtain

$$\begin{aligned} E \left[\frac{1}{G-1} \sum_{g=1}^G (U_g - \bar{U})^2 \right] &= \frac{1}{G-1} \sum_{g=1}^G V(U_g) + \frac{1}{G-1} \sum_{g=1}^G (E(U_g) - E(\bar{U}))^2 - \frac{1}{G(G-1)} \sum_{g=1}^G V(U_g) \\ &= \frac{1}{G} \sum_{g=1}^G V(U_g) + \frac{1}{G-1} \sum_{g=1}^G (E(U_g) - E(\bar{U}))^2. \end{aligned}$$

Chapter 4

Alternatives to the parallel-trends assumption

Throughout this chapter, we assume that we are in Design CLA: $D_{g,t} = 1\{t > T_0\}D_g$, as we have assumed in Chapter 3. Unless otherwise noted, we also assume that Assumption ND holds: the treatment does not have dynamic effects. Then, $\text{TE}_{g,t}$ reduces to

$$\text{TE}_{g,t} = E(Y_{g,t}(1) - Y_{g,t}(0)),$$

the ATE in cell (g, t) . That restriction is not of essence and simplifies the exposition.

4.1 TWFE and DID estimators with control variables

4.1.1 Conditional parallel-trends

Unless otherwise indicated, in this section we assume that the data contains only two time periods: $T = 2$, $T_0 = T_1 = 1$.

Non-parametric conditional parallel-trends assumptions. Let X_g denote a $K \times 1$ vector of control variables. In this section, the design and the controls $(X_g)_{g \in \{1, \dots, G\}}$ are implicitly conditioned upon, so we treat the controls as non-stochastic. Introducing controls allows us

to replace Assumption PT (which, under Assumption ND, reduces to (2.1)) by the following condition.

Assumption CPT (*Conditional parallel trends*) *There exists a function $m : x \mapsto m(x)$ such that $E [Y_{g,2}(0) - Y_{g,1}(0)] = m(X_g)$.*

Interpret Assumption CPT.

Assumption CPT is a conditional parallel-trends assumption. It implies that if two groups g_1 and g_2 are such that $X_{g_1} = X_{g_2}$, then g_1 and g_2 have the same expected evolution of their untreated outcome:

$$X_{g_1} = X_{g_2} \Rightarrow E [Y_{g_1,2}(0) - Y_{g_1,1}(0)] = E [Y_{g_2,2}(0) - Y_{g_2,1}(0)].$$

But under Assumption CPT, groups with different values of their control variables may have different expected evolutions of their untreated outcomes. Thus, Assumption CPT may be more plausible than Assumption PT. It has for instance been considered by Heckman, Ichimura and Todd (1997), Blundell et al. (2004), Abadie (2005), and Sant'Anna and Zhao (2020).

Conditional parallel-trends with a linear functional form. One could also strengthen Assumption CPT, assuming that $m : x \mapsto m(x)$ is linear:

Assumption LPT (*Conditional parallel trends, with a linear functional form*) *There exists a real number γ_2 and a $K \times 1$ vector θ such that $E [Y_{g,2}(0) - Y_{g,1}(0)] = \gamma_2 + X'_g \theta$.*

Conditional parallel trends with respect to covariates in levels or in first differences? The control variables X_g may be a function of a vector of time-varying covariates $X_{g,t}$, where by covariates we mean variables that cannot be affected by the treatment. One could have $X_g = X_{g,1}$: one controls for the baseline value of the covariates. One could also have $X_g = X_{g,2} - X_{g,1}$: one controls for the first differences of the covariates. Or one could have $X_g = (X'_{g,1}, (X_{g,2} - X_{g,1})')'$: one controls for the baseline value and the first difference of the covariates.

With $X_g = X_{g,1}$, Assumption CPT means that groups with the same baseline levels of the covariates experience parallel trends. With $X_g = X_{g,2} - X_{g,1}$, Assumption CPT means that groups with the same changes of the covariates experience parallel trends. Whether Assumption CPT is more plausible with levels and/or first-differences is context specific and depends on whether differential trends between groups are more likely to be due to differences in the level of their covariates, to the change of their covariates, or to both. Still, there is one important difference between controlling for covariates in levels or in first-difference. The requirement that the covariates be unaffected by the treatment is fairly plausible when one controls for $X_{g,1}$: the treatment is not available yet at period one, so this is only ruling out anticipation effects on the covariates. On the other hand, when one controls for $X_{g,2} - X_{g,1}$, this requirement becomes stronger: $X_{g,2}$ could be affected by the treatment, in which case controlling for $X_{g,2} - X_{g,1}$ could lead to a so-called “bad controls” problem (see Section 3.2.3 of Angrist and Pischke, 2009). Then, researchers controlling for covariates in first difference need to be able to plausibly rule out treatment effects on covariates.

4.1.2 TWFE regressions with control variables

A TWFE regression with control variables. Assume that one controls for $X_g = X_{g,1}$, and one is ready to assume that Assumption LPT holds. [Propose a TWFE regression to estimate the ATT.](#)

$$Y_{g,t} = \sum_{g'=1}^G \hat{\alpha}_{g'} 1\{g = g'\} + \hat{\gamma}_2 1\{t = 2\} + X'_{g,1} \hat{\theta} 1\{t = 2\} + \hat{\beta}_X^{\text{fe}} D_{g,t} + \hat{\epsilon}_{g,t}. \quad (4.1)$$

With $T = 2$, the regressions in (4.1) and (3.1) are very similar, except that by including $X'_{g,1} 1\{t = 2\}$, the one in (4.1) allows groups with different values of $X_{g,1}$ to experience different trends. In designs with a binary treatment and no variation in treatment timing, we have seen that $\hat{\beta}^{\text{fe}}$ is unbiased for the ATT under Assumption PT. Then, it would seem natural to assume that in the

same designs, $\hat{\beta}_X^{\text{fe}}$ is unbiased for the ATT under Assumption LPT. Actually, Theorem 5 below shows this is not necessarily the case.

$\hat{\beta}_X^{\text{fe}}$ may be biased if treatment effects are heterogeneous and correlated with the covariates. Assume that $X_{g,1}$ is of dimension one. Let $\mu_{X,1} = \frac{1}{G_1} \sum_{g:D_g=1} X_{g,1}$, $\mu_{X,0} = \frac{1}{G_0} \sum_{g:D_g=0} X_{g,1}$, $\sigma_{X,1}^2 = \frac{1}{G_1} \sum_{g:D_g=1} (X_{g,1} - \mu_{X,1})^2$, and $\sigma_{X,0}^2 = \frac{1}{G_0} \sum_{g:D_g=0} (X_{g,1} - \mu_{X,0})^2$ respectively denote the average and the variance of the covariate in the treatment and in the control group. Let $w_1 = \frac{G_1 \sigma_{X,1}^2}{G_0 \sigma_{X,0}^2 + G_1 \sigma_{X,1}^2}$, and let

$$\beta_X = \frac{\frac{1}{G_1} \sum_{g:D_g=1} (X_{g,1} - \mu_{X,1}) \text{TE}_{g,2}}{\sigma_{X,1}^2}$$

be the coefficient from an unfeasible regression of $\text{TE}_{g,2}$ on $X_{g,1}$ in the treatment group.

Theorem 5 *In Design CLA, if Assumptions NA and LPT hold and $X_{g,1}$ is of dimension one,*

$$E(\hat{\beta}_X^{\text{fe}}) = \text{ATT} + w_1(\mu_{X,1} - \mu_{X,0})\beta_X = \sum_{g:D_g=1} W_g \text{TE}_{g,2},$$

where $W_g = \frac{1}{G_1} + \frac{w_1(\mu_{X,1} - \mu_{X,0})(X_{g,1} - \mu_{X,1})}{G_1 \sigma_{X,1}^2}$. One has that $\sum_{g:D_g=1} W_g = 1$, but one could have that $W_g < 0$ for some g .

The first equality of Theorem 5, which to our knowledge was first shown in this textbook, shows that if treatment effects are heterogeneous and correlated with the baseline covariate $X_{g,1}$ ($\beta_X \neq 0$), and if $X_{g,1}$ is correlated with the treatment group indicator D_g ($\mu_{X,1} - \mu_{X,0} \neq 0$), then $\hat{\beta}_X^{\text{fe}}$ is biased for the ATT. The proof of that first equality is given in this chapter's appendix. Intuitively, the problem comes from the fact that (4.1) tries to do two things at the same time: estimate θ , the “effect” of $X_{g,1}$ on the evolution of the untreated outcome, and estimate the treatment’s effect. As (4.1) is estimated on the full sample of treated and untreated groups, $\hat{\theta}$ captures both the correlation between $X_{g,1}$ and the evolution of the untreated outcome, but also the correlation between $X_{g,1}$ and the treatment’s effect. Then, $\hat{\theta}$ is biased for θ , which ultimately biases $\hat{\beta}_X^{\text{fe}}$.

A decomposition of $\hat{\beta}_X^{\text{fe}}$ as a weighted sum of treatment effects where weights sum to one but may not all be positive. The second equality of Theorem 5 shows that $\hat{\beta}_X^{\text{fe}}$

is unbiased for a linear combination of the effects $\text{TE}_{g,t}$ across all treatment groups, where the loadings multiplying the treatment effects sum to one but may not all be positive. Throughout this textbook, linear combinations of effects with loadings summing to one are referred to as “weighted sums of effects”, while linear combinations of effects with positive loadings summing to one are referred to as “weighted averages of effects” or “convex combinations of effects”. This second equality directly follows from the first equality in the proposition, once noted that β_X is a weighted sum of the effects $\text{TE}_{g,t}$, with weights $\frac{X_{g,1} - \mu_{X,1}}{\sigma_{X,1}^2 G_1}$ that sum to zero, while the ATT is a weighted average of the same effects with weights $\frac{1}{G_1}$ summing to one. This second equality is a special case of Theorem S4 in the Web Appendix of de Chaisemartin and D'Haultfoeuille (2020). That result shows that beyond the special case considered here with a single time-invariant covariate, under a parallel-trends assumption TWFE regressions with covariates always estimate weighted sums of effects where weights sum to one, but where some weights may be negative. Due to the negative weights, one could have that $\text{TE}_{g,t} \geq 0$ for all (g, t) , but $E(\widehat{\beta}_X^{\text{fe}}) < 0$: the TWFE regression does not satisfy the so-called no-sign reversal property, an issue we will return to in the following chapter.

4.1.3 DID estimators with control variables

4.1.3.1 Parametric estimators

Linear outcome regression. The TWFE regression with control variables is biased because it tries to do too many things at the same time (jointly estimate the effect of the controls X_g on the untreated-outcome's evolution and the treatment's effect). Instead, one can use a two-step estimation procedure, inspired from Heckman et al. (1997), where in a first step one estimates θ in Assumption LPT, before estimating the treatment effect. In the first step, which regression should we run to estimate θ ?

We can regress $Y_{g,2} - Y_{g,1}$ on the covariates, restricting the sample to groups such that $D_g = 0$:

$$Y_{g,2} - Y_{g,1} = \hat{\gamma}_2^{or} + X'_g \hat{\theta}^{or} + \hat{\epsilon}_g. \quad (4.2)$$

Then, propose an unbiased estimator of the ATT.

Let

$$\text{DID}_{X,\text{or}} = \frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,2} - Y_{g,1} - (\hat{\gamma}_2^{or} + X'_g \hat{\theta}^{or})).$$

Here is the intuition underlying $\text{DID}_{X,\text{or}}$. Under Assumption LPT, groups' outcome evolution without treatment is a linear function of the control variables. We estimate the coefficients in that linear function by regressing control groups' outcome evolution on the covariates. We then compute $\hat{\gamma}_2^{or} + X'_g \hat{\theta}^{or}$, groups' predicted outcome evolution without treatment, based on that regression. Finally, DID_X subtracts from treated groups' actual outcome evolution their predicted outcome evolution without treatment, to recover their treatment effect.

Theorem 6 *In Design CLA, if Assumptions NA and LPT hold, $E(\text{DID}_{X,\text{or}}) = \text{ATT}$.*

Proof of Theorem 6. Under Assumption LPT, one can show that $\hat{\gamma}_2^{or}$ and $\hat{\theta}^{or}$ are unbiased for γ_2 and θ . Then,

$$\begin{aligned} E(\text{DID}_{X,\text{or}}) &= \frac{1}{G_1} \sum_{g:D_g=1} (E(Y_{g,2}(1) - Y_{g,1}(0)) - (E(\hat{\gamma}_2^{or}) + X'_g E(\hat{\theta}^{or}))) \\ &= \frac{1}{G_1} \sum_{g:D_g=1} E(Y_{g,2}(1) - Y_{g,2}(0)) \\ &\quad + \frac{1}{G_1} \sum_{g:D_g=1} (E(Y_{g,2}(0) - Y_{g,1}(0)) - (\gamma_2 + X'_g \theta)) \\ &= \text{ATT} + \frac{1}{G_1} \sum_{g:D_g=1} (\gamma_2 + X'_g \theta - (\gamma_2 + X'_g \theta)) \\ &= \text{ATT}. \end{aligned}$$

The second equality follows from adding and subtracting $Y_{g,2}(0)$ and from the unbiasedness of $\hat{\gamma}_2^{or}$ and $\hat{\theta}^{or}$. The third equality follows from the definition of ATT and Assumption LPT **QED**.

Propensity-score reweighting. Under Assumption CPT, one can also use propensity-score reweighting to estimate the ATT. In a first step, one regresses the treatment group indicator D_g on X_g , using a logit or probit model. Let $\hat{p}(X_g)$ denote group's g predicted probability to be a treatment group according to this regression, its so-called propensity score. Then, Abadie (2005) proposes to use

$$\text{DID}_{X,\text{ps}} := \frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,2} - Y_{g,1}) - \frac{1}{G_0} \sum_{g:D_g=0} \frac{\hat{p}(X_g)}{1 - \hat{p}(X_g)} \frac{G_0}{G_1} (Y_{g,2} - Y_{g,1})$$

to estimate the ATT. Intuitively, $\text{DID}_{X,\text{ps}}$ compares the outcome evolution of treated and control groups, after reweighting control groups. As $x \mapsto x/(1-x)$ is increasing in x on $(0, 1)$, the reweighting gives more weight to control groups that have a larger value of $\hat{p}(X_g)$, namely to control groups who, based on their X_g , have a larger predicted probability of being treated. Thus, the reweighting gives more weight to control groups that “look like” treatment groups. Actually, one can show that the reweighting ensures that asymptotically, the distribution of X_g is the same in the treatment group and in the reweighted control group, so that $\text{DID}_{X,\text{ps}}$ is consistent for the ATT. Importantly, $\text{DID}_{X,\text{ps}}$ relies on the assumption that the population propensity score $p(X_g)$ be bounded away from one: $p(X_g) \leq 1 - \epsilon$ for some $\epsilon > 0$, the so-called overlap condition. Groups with an estimated propensity score too close to one can be dropped from the estimation, to decrease the estimator's variance. One can for instance follow the popular rule proposed by Crump, Hotz, Imbens and Mitnik (2009) of trimming treatment and control groups with a propensity score larger than 0.9.

4.1.3.2 Doubly robust and/or non-parametric estimators

Parametric doubly-robust estimator. $\text{DID}_{X,\text{or}}$ is inconsistent if Assumption CPT holds, but the functional form for $m(X_g)$ in Assumption LPT is misspecified.¹ Likewise, $\text{DID}_{X,\text{ps}}$ with a parametric logit or probit model is also inconsistent if the true propensity score $p(X_g) := P(D_g = 1|X_g)$ does not follow the chosen parametric model.² To alleviate these concerns, Sant'Anna and

¹Similarly, Caetano, Callaway, Payne and Rodrigues (2022) show that under Assumption CPT alone, $\widehat{\beta}_X^{\text{fe}}$ may be biased even if the treatment effect is homogeneous.

²For instance, a probit model is correctly specified if there exists a $K \times 1$ vector β_0 such that $P(D_g = 1|X_g) = \Phi(X'_g \beta_0)$, where Φ is the cumulative distribution function of a standard normal, and the model is misspecified if there does not exist a β_0 satisfying this condition.

Zhao (2020) propose to use a so-called doubly-robust DID estimator, that combines outcome regression and propensity-score reweighting. Specifically, one of their estimators of the ATT is

$$\text{DID}_{X,\text{dr}} := \frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,2} - Y_{g,1} - (\hat{\gamma}_2^{or} + X_g' \hat{\theta}^{or})) - \frac{1}{G_0} \sum_{g:D_g=0} \frac{\hat{p}(X_g)}{1 - \hat{p}(X_g)} \frac{G_0}{G_1} (Y_{g,2} - Y_{g,1} - (\hat{\gamma}_2^{or} + X_g' \hat{\theta}^{or})),$$

where $(\hat{\gamma}_2^{or}, \hat{\theta}^{or})$ are the coefficients from the linear regression in (4.2), and where $\hat{p}(X_g)$ comes from a probit or logit regression of D_g on X_g . Sant'Anna and Zhao (2020) show that $\text{DID}_{X,\text{dr}}$ is consistent for the ATT if either $m(X_g)$ follows the linear functional form in Assumption LPT, or the parametric model for the propensity score is correctly specified. Thus, $\text{DID}_{X,\text{dr}}$ offers some robustness against misspecification, though it still requires that one of the two parametric models be correctly specified: if $m(X_g)$ does not follow the linear functional form in Assumption LPT and the parametric model for $p(X_g)$ is incorrectly specified, $\text{DID}_{X,\text{dr}}$ is inconsistent.

Non-parametric estimator when the control variables take a small number of values.

When X_g takes a small number of values relative to the sample size, it is straightforward to estimate ATT fully non-parametrically. For instance, one can estimate $m(X_g)$ non-parametrically, by regressing $Y_{g,2} - Y_{g,1}$ on FEs for all values that X_g can take, restricting the sample to groups such that $D_g = 0$. Then, one uses

$$\text{DID}_{X,\text{or-np}} := \frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,2} - Y_{g,1} - \hat{m}(X_g))$$

to estimate the ATT. This strategy can for instance be used to allow for state-specific trends in a county-level analysis, or industry-specific trends in a firm-level analysis. Then, a full set of state or industry FEs needs to be included in the regression in (4.2). The resulting estimator relies on the assumption that counties in the same state (resp. firms in the same sector) would have experienced the same outcome evolutions without treatment. On the other hand, counties in different states (resp. firms in different sectors) are allowed to experience different evolutions.

Non-parametric estimator when the control variables take a large number of values.

Assume that one wants to control for variables taking a large number of values, such as continuous variables, without imposing functional-form assumptions on $m(X_g)$ and the propensity score $p(X_g)$. To achieve that purpose, a recent literature proposes to use debiased machine learning (DML) methods, namely the combination of machine learning (ML) estimators (e.g. Lasso,

random forest, neural networks) of $m(X_g)$ and $p(X_g)$, doubly-robust estimation, and cross-fitting (see, e.g., Chang, 2020; Lu et al., 2019, for early applications of DML to DID). For instance, one may use the following algorithm:

1. Randomly divide the G groups into two subsamples \mathcal{I}_1 and \mathcal{I}_2 .
2. Use an ML method to predict $Y_{g,2}(0) - Y_{g,1}(0)$ and D_g given X_g in subsample 1, and let $\widehat{m}_{\text{ml}}^{(1)}(X_g)$ and $\widehat{p}_{\text{ml}}^{(1)}(X_g)$ denote the corresponding estimators.
3. Compute a doubly-robust ATT estimator in subsample 2, based on the estimators of $m(X_g)$ and $p(X_g)$ computed in subsample 1:

$$\begin{aligned} \text{DID}_{X,\text{dr-ml}}^{(2)} := & \frac{1}{\#\{g \in \mathcal{I}_2, D_g = 1\}} \sum_{g \in \mathcal{I}_2, D_g=1} (Y_{g,2} - Y_{g,1} - \widehat{m}_{\text{ml}}^{(1)}(X_g)) \\ & - \frac{1}{\#\{g \in \mathcal{I}_2, D_g = 0\}} \sum_{g \in \mathcal{I}_2, D_g=0} \frac{\widehat{p}_{\text{ml}}^{(1)}(X_g)}{1 - \widehat{p}_{\text{ml}}^{(1)}(X_g)} \frac{\#\{g \in \mathcal{I}_2, D_g = 0\}}{\#\{g \in \mathcal{I}_2, D_g = 1\}} (Y_{g,2} - Y_{g,1} - \widehat{m}_{\text{ml}}^{(1)}(X_g)). \end{aligned}$$

4. Revert the roles of subsamples 1 and 2: compute $\text{DID}_{X,\text{dr-ml}}^{(1)}$ a doubly-robust ATT estimator in subsample 1, based on estimators of $m(X_g)$ and $p(X_g)$ computed in subsample 2.
5. Finally, one uses

$$\text{DID}_{X,\text{dr-ml}} := 1/2 \text{ DID}_{X,\text{dr-ml}}^{(1)} + 1/2 \text{ DID}_{X,\text{dr-ml}}^{(2)}$$

to estimate the ATT.

4.1.3.3 Controlling for group-specific linear trends

Identifying assumption: common deviations from linear trends. In this section, let us momentarily assume that $T = 5$ and $T_0 = 3$: the data contains five periods, and treated groups become treated at period four. We consider the following assumption.

Assumption CDLT (*Common deviations from linear trends*) For all $t \geq 2$, $E [Y_{g,t}(0) - Y_{g,t-1}(0)] = \gamma_t + \lambda_g$.

Assumption CDLT allows groups to experience group-specific linear trends λ_g , but requires that between each pair of consecutive periods, all groups have the same expected deviation from their

linear trend. Under Assumption CDLT, the standard parallel-trends assumption holds, but for the first-differenced outcome: for all $t \geq 3$,

$$E [Y_{g,t}(0) - Y_{g,t-1}(0) - (Y_{g,t-1}(0) - Y_{g,t-2}(0))] = \tilde{\gamma}_t.$$

DID estimators. Under Assumption CDLT, Mora and Reggio (2019) show that one can use

$$\begin{aligned} \text{DID}_{X,\text{tr-lin}} := & \frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,4} - Y_{g,3} - (Y_{g,3} - Y_{g,2})) \\ & - \frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,4} - Y_{g,3} - (Y_{g,3} - Y_{g,2})) \end{aligned} \quad (4.3)$$

to estimate ATT_1 . Intuitively, instead of comparing first-differences of the outcome in the treatment and in the control group, $\text{DID}_{X,\text{tr-lin}}$ compares second-differences. To estimate ATT_2 , one can use

$$\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,5} - Y_{g,3} - 2(Y_{g,3} - Y_{g,2})) - \frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,5} - Y_{g,3} - 2(Y_{g,3} - Y_{g,2})), \quad (4.4)$$

where $Y_{g,3} - Y_{g,2}$ has to be multiplied by two to account for the fact that $Y_{g,5} - Y_{g,3}$ is a long difference that incorporates $2\lambda_g$, the linear trend of group g . To test Assumption CDLT, one can compute the following pre-trends estimator:

$$\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,2} - Y_{g,1} - (Y_{g,3} - Y_{g,2})) - \frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,2} - Y_{g,1} - (Y_{g,3} - Y_{g,2})), \quad (4.5)$$

whose expectation is equal to zero under Assumption CDLT.

TWFE estimators. Alternatively, one could run a TWFE regression with group-specific linear trends:

$$Y_{g,t} = \sum_{g'=1}^G \hat{\alpha}_{g'} 1\{g = g'\} + \sum_{t'=1}^T \hat{\gamma}_{t'} 1\{t = t'\} + \sum_{g'=1}^G t \hat{\lambda}_{g'} 1\{g = g'\} + \hat{\beta}_{\text{tr-lin}}^{\text{fe}} D_{g,t} + \hat{\epsilon}_{g,t}. \quad (4.6)$$

Is $\hat{\beta}_{\text{tr-lin}}^{\text{fe}}$ unbiased for the ATT under Assumption CDLT? If not, what is the issue with the TWFE regression in (4.6)?

As any TWFE estimator with control variables, $\hat{\beta}_{\text{tr-lin}}^{\text{fe}}$ may be biased if treatment effects are heterogeneous and correlated with the covariates. In this specific case, as (4.6) is estimated on the

full sample, the linear trends $\hat{\lambda}_g$ reflect both groups' outcome evolutions prior to treatment, but also their outcome evolutions after the treatment. Assume for instance that the treatment effect increases linearly with length of exposure. This generates a linear trend after treatment which, for treated groups, contaminates $\hat{\lambda}_g$, the estimator of groups' linear trend without treatment. Then, $\hat{\lambda}_g$ is biased for λ_g , which ultimately biases $\hat{\beta}_{\text{tr-lin}}^{\text{fe}}$. With TWFE event-study estimators, further issues arise when introducing group-specific linear trends in the regression, as then the group-specific linear trends can be collinear with the relative-time indicators.

4.1.3.4 Triple-difference estimators*

Set up. In this section, let us momentarily assume that one has outcome data at a more disaggregated level than the (g, t) cells, say at the (i, g, t) level, where i indexes, say, individuals, while g could represent, say, municipalities. In each cell, some individuals are eligible for treatment and will therefore be treated if g is a treatment group and $t > T_0$, and some individuals are ineligible and therefore remain untreated even if g is a treatment group and $t > T_0$. Let $X_{i,g,t}$ denote the eligibility status of individual i in cell (g, t) . For instance, one may have that only females are eligible for treatment, so that $X_{i,g,t} = 1$ for females and $X_{i,g,t} = 0$ for males. Then, $D_{i,g,t} = 1\{t > T_0\}D_gX_{i,g,t}$. For $x \in \{0, 1\}$, let $Y_{x,g,t}$ denote the average outcome across individuals satisfying $X_{i,g,t} = x$ in cell (g, t) . In the example, $Y_{1,g,t}$ denotes the average outcome of females in cell (g, t) , while $Y_{0,g,t}$ is the average outcome of males. Let $Y_{x,g,t}(0)$ denote the average outcome without treatment across individuals satisfying $X_{i,g,t} = x$.

Identifying assumption. We consider the following assumption:

$$E [Y_{1,g,2}(0) - Y_{0,g,2}(0) - (Y_{1,g,1}(0) - Y_{0,g,1}(0))] \text{ does not depend on } g. \quad (4.7)$$

(4.7) is a parallel-trends assumption on the difference between the untreated outcomes of the eligible and ineligible subgroups: it requires that all groups experience the same evolution of that difference. (4.7) is neither stronger or weaker than Assumption PT. On the other hand, (4.7) is weaker than assuming that eligibles and ineligibles are on parallel trends:

$$E [Y_{x,g,2}(0) - Y_{x,g,2}(0)] \text{ does not depend on } g$$

for all $x \in \{0, 1\}$. This may be why researchers often feel that triple-difference estimators rely on a weaker assumption than DIDs.

Triple-difference estimator. Under (4.7), the following “triple-difference” estimator

$$\text{TRI-D} := \frac{1}{G_1} \sum_{g:D_g=1} (Y_{1,g,2} - Y_{0,g,2} - (Y_{1,g,1} - Y_{0,g,1})) - \frac{1}{G_0} \sum_{g:D_g=0} (Y_{1,g,2} - Y_{0,g,2} - (Y_{1,g,1} - Y_{0,g,1}))$$

is unbiased for the ATT. Intuitively, TRI-D compares the evolution of the eligible versus ineligible difference in treated and control groups from period one to two. To compute TRI-D, one can regress $Y_{i,g,t}$ on an intercept, D_g , $X_{i,g,t}$, $1\{t = 2\}$, $D_g X_{i,g,t}$, $D_g 1\{t = 2\}$, $X_{i,g,t} 1\{t = 2\}$, and $D_{i,g,t}$.

Pre-trends test. If another pre-period, period zero, is available, one can run a pre-trends test of (4.7), by comparing the evolution of the eligible versus ineligible difference in treated and control groups from period zero to one.

4.1.3.5 When to include control variables in the estimation?

More or less credible results with control variables? Let us first clear up a common misconception. Increasing statistical precision cannot be, in and of itself, a reason to introduce control variables in a DID analysis. In this chapter’s appendix, we show that if one makes a parallel-trends assumption under which the DID estimators with and without covariates are both consistent, as well as an homoscedasticity assumption, the asymptotic variance of the estimator without covariates is always smaller than that of the estimator with covariates. This is in contrast with RCTs, where controlling for covariates is not needed for identification but can improve precision, and we also give some intuition for that difference in the appendix. Then, the only reason to introduce control variables is that doing so allows the researcher to work under a conditional parallel-trends assumption, which might be more plausible than the unconditional one, thus increasing the credibility of one’s results. At the same time, introducing covariates could also diminish results’ credibility. Choosing which control variables to include gives researchers the possibility to engage in specification searching and p-hacking, as different controls may lead to different results.³ The recent development of automated machine-learning based methods for

³Prespecifying the controls one will use is often not feasible in observational studies where DID estimators are used: there, it is difficult to credibly document when the researcher first accessed the data.

variable selection may attenuate these concerns, without making them disappear: these methods still require that the researcher specify a dictionary of potential controls to choose from, and different dictionaries can lead to different results.

Practical recommendations. In view of the above, we believe that if the pre-trend coefficients $\hat{\beta}_\ell^{\text{fe}}$ in the ES TWFE regression without controls in (3.6) are precisely estimated and not significantly different from zero, there may not be a compelling reason to include controls in the estimation: there is no indication of a violation of the unconditional parallel-trends assumption, so moving to a conditional assumption may not be warranted. If on the other hand pre-trend coefficients without controls are significant, large, or imprecise, there is a more compelling argument to include some controls, if pre-trends are less significant, smaller, or more precisely estimated with covariates. Still, this means that controls are only included after a pre-test has been conducted. It has been shown in other contexts that pre-testing/model-selection steps can bias the post-test estimator and distort inference (Leeb and Pötscher, 2003). Assessing if this issue is quantitatively important in DID estimation is an active area of research (Roth, 2022).

4.1.3.6 Investigating heterogeneous treatment effects under a conditional parallel-trends assumption

Target parameter: best-linear predictor. In this section, our target parameter is

$$\beta_{1,X} = \left(\sum_{g:D_g=1} X_g X_g^T \right)^{-1} \sum_{g:D_g=1} X_g \text{TE}_{g,2},$$

the coefficient on X_g in an infeasible regression of $\text{TE}_{g,2}$ on X_g among treated groups, which we already considered in Section 3.6.1 of the previous chapter. Therein, we saw that if X_g contains only one non-constant variable and that this variable is binary, then one can use a simple linear regression to estimate $\beta_{1,X}$ under a conditional parallel-trends assumption. We now show that when X_g contains several and/or non-binary variables, $\beta_{1,X}$ can also be estimated under conditional parallel trends.

Estimator. Propose an estimator that is unbiased for $\beta_{1,X}$ under Assumption LPT.

Under Assumption LPT,

$$\hat{\beta}_{1,X}^{\text{cpt}} = \left(\sum_{g:D_g=1} X_g X_g^T \right)^{-1} \sum_{g:D_g=1} X_g (Y_{g,2} - Y_{g,1} - (\hat{\gamma}_2^{or} + X_g' \hat{\theta}^{or}))$$

is unbiased for $\beta_{1,X}$. Indeed, one can show that

$$E(Y_{g,2} - Y_{g,1} - (\hat{\gamma}_2^{or} + X_g' \hat{\theta}^{or})) = \text{TE}_{g,2}$$

under Assumption LPT. Then, unbiasedness of $\hat{\beta}_{1,X}^{\text{cpt}}$ readily follows. In Section 3.6 of the previous chapter, we had proposed $\hat{\beta}_{1,X}$, another estimator of $\beta_{1,X}$ when X_g contains several and/or non-binary variables. [Explain the pros and cons of \$\hat{\beta}_{1,X}\$ and \$\hat{\beta}_{1,X}^{\text{cpt}}\$.](#)

$\hat{\beta}_{1,X}$ and $\hat{\beta}_{1,X}^{\text{cpt}}$ rely on different, non-nested and placebo-testable identifying assumptions. $\hat{\beta}_{1,X}$ is a one-step regression estimator so inference based on that estimator is straightforward. On the other hand, $\hat{\beta}_{1,X}^{\text{cpt}}$ is a two-step regression estimator, and we are not aware of R or Stata commands that estimate its asymptotic variance.

Estimating the conditional ATT function? Instead of estimating the best linear predictor of treated groups' treatment effects, one may be interested in estimating the conditional ATT (CATT) function, namely the function mapping groups' covariates to their treatment effect. If the CATT function is linear, it coincides with the best linear predictor, but if the CATT function is not linear the two functions differ. Few estimators of the CATT function under a conditional parallel-trends assumption have been proposed. An exception is Lu, Nie and Wager (2019), but the estimators in that paper are not implemented in a Stata or R command.

4.1.4 Controlling for the lagged outcome?

4.1.4.1 AR(1) model for the outcome without treatment*

In this section, we no longer assume that Assumption ND holds: lagged treatments may have an effect on the current outcome. We also no longer assume that $T = 2$.

An AR(1) model with fixed effects. A common justification to control for the lagged outcome is that the untreated outcome may follow an AR(1) model. For instance, assume that for $t \geq 2$

$$Y_{g,t}(\mathbf{0}_t) = \alpha_g + \gamma_t + \rho Y_{g,t-1}(\mathbf{0}_{t-1}) + \varepsilon_{g,t}, \quad (4.8)$$

for some α_g , γ_t , ρ and $\varepsilon_{g,t}$ satisfying $E[\varepsilon_{g,t}] = 0$. With $\rho = 0$, (4.8) boils down to the TWFE model for the never-treated outcome in (2.4), which we saw is equivalent to the parallel-trends assumption. Then, let us further assume that the effects of the current treatment and of the lagged outcome on the current outcome are constant: for $t \geq 2$, for all (d_1, \dots, d_t) in $\{0, 1\}^t$,

$$Y_{g,t}(d_1, \dots, d_t) = \alpha_g + \gamma_t + d_t \delta + \rho Y_{g,t-1}(d_1, \dots, d_{t-1}) + \varepsilon_{g,t}, \quad t = 1, \dots, T. \quad (4.9)$$

Combined with (4.8), this implies that for $t \geq 2$,

$$Y_{g,t} = \alpha_g + \gamma_t + \delta D_{g,t} + \rho Y_{g,t-1} + \varepsilon_{g,t}. \quad (4.10)$$

When $T \geq 3$, this motivates a TWFE regression controlling for $Y_{g,t-1}$.

Under the AR(1) model with fixed effects, whether we need to control for the lagged outcome depends on whether the outcome process has reached a steady-state. For simplicity, assume that $|\rho| < 1$ and (4.8) holds for all $t \in \mathbb{Z}$, including for negative time periods, even if we still assume that $Y_{g,t}$ is observed at $t = 1, \dots, T$ only. Then, replacing $Y_{g,t-1}(\mathbf{0}_{t-1})$ by $\alpha_g + \gamma_{t-1} + \rho Y_{g,t-2}(\mathbf{0}_{t-2}) + \varepsilon_{g,t-1}$ in (4.8) and iterating for $Y_{g,t-2}(\mathbf{0}_{t-2})$, $Y_{g,t-3}(\mathbf{0}_{t-3})$, ..., we obtain

$$Y_{g,t}(0_t) = \tilde{\alpha}_g + \tilde{\gamma}_t + \tilde{\varepsilon}_{g,t}, \quad (4.11)$$

where $\tilde{\alpha}_g := \alpha_g / (1 - \rho)$, $\tilde{\gamma}_t := \sum_{k=0}^{\infty} \rho^k \gamma_{t-k}$ and $\tilde{\varepsilon}_{g,t} := \sum_{k=0}^{\infty} \rho^k \varepsilon_{g,t-k}$ (we assume that the two series converge; this holds for instance if $|\gamma_t| \leq M$ for all $t \in \mathbb{Z}$ and $(\varepsilon_{g,t})_{t \in \mathbb{Z}}$ is stationary). Moreover, since $E[\tilde{\varepsilon}_{g,t}] = 0$, (4.11) is equivalent to the TWFE model for the never-treated outcome in (2.4), so the parallel-trends assumption holds unconditionally, and it is unnecessary to control for the lagged outcome to estimate consistently the ATT or the $(ATT_\ell)_{\ell=1, \dots, T_1}$. This conclusion only relies on (4.8), and it does not rely on the assumption that treatment effects are homogeneous in (4.9). Moreover, this conclusion still holds if we replace $\rho Y_{g,t-1}(0)$ by any linear combination of past potential outcomes, provided that we can invert the corresponding

relationship, as we did above. On the other hand, this conclusion relies on the assumption that the outcome process started sufficiently long ago so as to assume that it has reached a steady state where we can omit the initial condition. At the other extreme, assume that $T = 2$ and that the outcome process starts at period 1. Then,

$$Y_{g,2}(0_2) - Y_{g,1}(0) = \gamma_2 - \gamma_1 + (\rho - 1)Y_{g,1}(0) + \varepsilon_{g,2} - \varepsilon_{g,1},$$

so the unconditional parallel-trends assumption does not hold, and one needs to control for the lagged outcome.

Under an AR(1) model with fixed effects, controlling for the lagged outcome may lead to a so-called Nickell bias. While identification arguments can motivate controlling for the lagged outcome, doing so can lead to issues when it comes to estimation. Even if we assume $E[\varepsilon_{g,t}|Y_{g,t-1}] = 0$, the TWFE estimator of δ using $Y_{g,t-1}$ as a control is biased, and inconsistent when G diverges to $+\infty$ but T is fixed (Nickell, 1981), a realistic asymptotic approximation in many DID applications where the number of time periods is not very large. Specifically, due to the group FEs, it follows from the Frisch-Waugh-Lowell theorem that the regression is equivalent to a regression without group FEs and where all variables have been demeaned within groups. However, letting $Y_{g,2:T}$ and $Y_{g,1:T-1}$ respectively denote the average of $Y_{g,t}$ from periods 2 to T and from periods 1 to $T - 1$ (and similarly for $D_{g,t}$), we have, under (4.10),

$$Y_{g,t} - Y_{g,2:T} = \gamma_t - \gamma_{2:T} + \delta(D_{g,t} - D_{g,2:T}) + \rho(Y_{g,t-1} - Y_{g,1:T-1}) + \varepsilon_{g,t} - \varepsilon_{g,2:T}.$$

In this equation, which regressor is mechanically correlated to the residual $\varepsilon_{g,t} - \varepsilon_{g,2:T}$?

$Y_{g,t-1} - Y_{g,1:T-1}$ is a function of $(\varepsilon_{g,1}, \dots, \varepsilon_{g,T-1})$ and is therefore mechanically correlated to the residual $\varepsilon_{g,t} - \varepsilon_{g,2:T}$, thus leading to an omitted variable bias that may bias the estimators of ρ and δ .

Solutions to the Nickell bias. Consistent estimators of the coefficients in (4.10) can be obtained by estimating the equation in first difference, instrumenting $Y_{g,t} - Y_{g,t-1}$ by $Y_{g,t-2}$ and/or

earlier lags of the outcome (see, e.g., Arellano and Bond, 1991), but such regressions sometimes suffer from weak instrument problems. Alternatively, one may derive analytic expressions of the coefficients' asymptotic bias to bias-correct them (see, e.g., Kiviet, 1995).⁴ However, those two estimation strategies rely on the homogeneous treatment effect assumption in (4.9), and extending them to allow for heterogeneous treatment effects is not straightforward. Klosin (2024) proposes a bias-correction method that allows for heterogeneous treatment effects along some covariates, but the method does not allow for unrestricted heterogeneity.

4.1.4.2 Self-selection, and parallel trends conditional on the baseline outcome

In this section, we go back to ruling out dynamic effects and assuming that the data contains only two time periods: $T = 2$, $T_0 = T_1 = 1$.

Self-selection and the Ashenfelter dip. Sometimes, one might worry that treated groups choose to get treated after experiencing a negative outcome shock. For instance, Ashenfelter (1978) finds that US workers choosing to receive a post-schooling training program experience an earnings drop before doing so, a so-called Ashenfelter's dip. Then, if outcome shocks are positively correlated over time, it is more likely that treated units would have kept experiencing negative shocks if they had not been treated, thus leading to a violation of Assumption PT.

Parallel-trends conditional on the baseline outcome. In the presence of self-selection, one may replace Assumption PT by a conditional parallel-trends assumption, where one assumes that treated and control groups with the same baseline outcome $Y_{g,1}(0)$ experience parallel trends. Intuitively, treated and control groups with the same $Y_{g,1}(0)$ have experienced similar pre-treatment outcome shocks, thus making it more plausible that they would have kept experiencing the same outcome shocks if the treated had not been treated. Formally, we assume that there exists a function $m : x \mapsto m(x)$ such that

$$E [Y_{g,2}(0) - Y_{g,1}(0)] = m(Y_{g,1}(0)), \quad (4.12)$$

⁴If (4.9) holds, then $\text{ATT}_\ell = \delta(1 - \rho^\ell)/(1 - \rho)$, so having consistent estimators of δ and ρ is sufficient to consistently estimate the ATT_ℓ effects.

the same condition as in Assumption CPT, with the baseline outcome playing the role of the covariates. Like the covariates in the previous sections, the baseline outcomes $(Y_{g,1}(0))_{g \in \{1, \dots, G\}}$ are implicitly conditioned upon in what follows, and we treat them as non-stochastic. [Is Assumption CPT still a parallel-trends assumption?](#)

Letting $\tilde{\gamma}(x) = m(x) + x$, and because $Y_{g,1}(0)$ is conditioned upon, (4.12) is equivalent to

$$E [Y_{g,2}(0)] = \tilde{\gamma}(Y_{g,1}(0)) :$$

groups with the same baseline outcome should have the same expected untreated outcome at period 2. This is equivalent to the conditional independence or ignorability assumption that underlies matching estimators. In fact, comparing the outcomes evolutions $Y_{g,2} - Y_{g,1}$ of treated and control groups with the same $Y_{g,1}$ is equivalent to just comparing their $Y_{g,2}$: a DID estimator conditional on $Y_{g,1}$ is actually a matching estimator (Chabé-Ferret, 2015). Accordingly, (4.12) is also similar to a sequential randomization assumption, where one assumes that treatment is randomly assigned conditional on the lagged outcome (Robins, 1986; Bojinov et al., 2021).

Matching on the baseline outcome might lead to spurious results due to mean-reversion. Mean-reversion phenomena when matching on the baseline outcome were first discussed in McNemar (1940), and further discussed in Chabé-Ferret (2015) and Daw and Hatfield (2018). The following toy model is sufficient to convey the intuition. Assume that the treatment has no effect ($Y_{g,t}(1) = Y_{g,t}(0)$), and treated groups are such that $P(Y_{g,t}(0) = 1) = P(Y_{g,t}(0) = 2) = 1/2$, while control groups are such that $P(Y_{g,t}(0) = 0) = P(Y_{g,t}(0) = 1) = 1/2$, and $Y_{g,2}(0)$ and $Y_{g,1}(0)$ are independent for all g . Thus, for all t , $E(Y_{g,t}(0)) = 1.5$ for treated groups while $E(Y_{g,t}(0)) = 0.5$ for control groups. [In this DGP, should a researcher use a DID estimator, or a matching estimator conditional on the baseline outcome?](#)

The DID estimator is equal to

$$\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,2} - Y_{g,1}) - \frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,2} - Y_{g,1}).$$

For all g , $E(Y_{g,2} - Y_{g,1}) = E(Y_{g,2}(0) - Y_{g,1}(0)) = 0$. Therefore, the expectation of the DID estimator is equal to zero, so it is unbiased. The matching estimator is equal to

$$\frac{1}{G_1} \sum_{g:D_g=1, Y_{g,1}=1} Y_{g,2} - \frac{1}{G_0} \sum_{g:D_g=0, Y_{g,1}=1} Y_{g,2},$$

because the only common value of $Y_{g,1}$ for the treated and control groups is 1. Then, as $Y_{g,2}(0)$ and $Y_{g,1}(0)$ are independent, for every treated group g such that $Y_{g,1} = Y_{g,1}(0) = 1$, $E(Y_{g,2}) = E(Y_{g,2}(0)) = 1.5$, while for every control group g such that $Y_{g,1} = Y_{g,1}(0) = 1$, $E(Y_{g,2}) = E(Y_{g,2}(0)) = 0.5$. Accordingly, the expectation of the matching estimator is equal to $1.5 - 0.5 = 1$: this estimator is biased. Intuitively, the matching estimator compares the period-two outcomes of treated groups with a bad period-one shock and control groups with a good period-one shock. But at period two the treatment groups revert to their higher mean, while control groups revert to their lower mean, thus generating a spurious positive estimate.

Practical recommendations. Of course the previous toy model is very simplistic, but Chabé-Ferret (2015) and Daw and Hatfield (2018) show that mean-reversion can still happen in more realistic models, and Bach, Bozio, Guillouzouic and Malgouyres (2023) is a recent striking example showing that mean-reversion can lead to a substantial bias in an actual empirical application. If more than two time periods are available, researchers that want to control for the lagged outcome should compute a placebo matching estimator, comparing the Y_{g,T_0} outcome of the treated and controls with the same Y_{g,T_0-1} outcome. This can help assess if mean reversion will mechanically bias the actual matching estimator. They may also match on less recent outcome lags than Y_{g,T_0} , and they could also match on several outcome lags: the more lags one matches on, the less likely it is that treated and controls are matched on transitory shocks, the essence of the mean-reversion problem. Here, there is an interesting connection with the synthetic control estimator. To reconstruct treated's counterfactual outcome, that estimator uses a weighted average of controls that match the treated at all pre-treatment periods. The synthetic control estimator is consistent when the number of groups and the number of pre-treatment periods go to infinity, the later requirement ensuring that treated and controls are not matched on transitory shocks.

Applications where the treatment is a function of the baseline outcome. There are many applications where the treatment is a function of the pre-treatment outcome: $D_g = f(Y_{g,T_0})$. For instance, only firms whose number of employees is above a threshold at period T_0 may benefit from a policy, and one is interested in the policy's effect on firms' number of employees at period $T_0 + 1$. In such cases, mean-reversion issues might also bias DID estimators: some firms may happen to be treated just because they experienced a good shock at period T_0 , and could naturally revert to a lower number of employees at period $T_0 + 1$. Similarly, some firms may happen to be untreated just because they experienced a bad shock at period T_0 , and could naturally revert to a higher number of employees at period $T_0 + 1$. Again, a pre-trends estimator comparing the $T_0 - 1$ to T_0 outcome evolutions of firms with a number of employees higher/lower than the threshold at period $T_0 - 1$ can be useful to assess if mean-reversion is likely to bias the actual T_0 -to- $T_0 + 1$ DID estimator.

4.1.5 Computation: Stata and R commands to compute DID estimators with controls

Several Stata and R commands compute DID estimators with control variables. We review several of them thoroughly in Chapter 6. At this stage, we just mention two commands.

The drdid command. $\text{DID}_{X,\text{or}}$, $\text{DID}_{X,\text{ps}}$, and $\text{DID}_{X,\text{dr}}$ are computed by the `dridid` Stata (see Rios-Avila, Sant'Anna and Naqvi, 2021) and R (see Sant'Anna and Zhao, 2022) commands. The syntax of the Stata command is:

```
dridid outcome [controls_var_names], ivar(groupid) time(timeid) treatment(var_name).
```

With time-varying covariates, the control variables inputted to the command have to be defined as $X_g = X_{g,2} - X_{g,1}$ at every date. If the user inputs $X_{g,t}$, the command controls for $X_{g,1}$.

The did_multiplegt_dyn command. $\text{DID}_{X,\text{or}}$ is computed by the `did_multiplegt_dyn` Stata (see de Chaisemartin, Ciccia, D'Haultfœuille, Knau, Malézieux and Sow, 2024b) and R (see de Chaisemartin, Ciccia, D'Haultfœuille, Knau, Malézieux and Sow, 2024a) commands. The syntax of the Stata command is:

```
did_multiplegt_dyn outcome groupid timeid treatment, controls(var_names).
```

With time-invariant covariates, the control variables inputted to the command have to be defined as $X_g \times t$. If the user defines the controls as X_g , the controls are dropped from the estimation. If the control variables take a small number of values relative to the number of groups G , and one wants to control for those variables without assuming a functional form, one can use the `trends_non_param` instead of the `controls` option, inputting the control variables to the option. For instance, if one wants to include sector-specific trends in a firm-level analysis, one just needs to input the sector identifier to the `trends_non_param` option. $DID_{X,\text{tr-lin}}$ is also computed by `did_multiplegt_dyn`. Then, the syntax is:

```
did_multiplegt_dyn outcome groupid timeid treatment, trends_lin.
```

Computing pre-trend and event-study estimators with controls when $T > 2$. Stata and R commands computing DID estimators with control variables can generally be used with more than two time periods. Then, they can compute pre-trends and event-study estimators with controls, while avoiding the issues of TWFE regressions with controls. Those estimators are analogous to the two-periods DID estimators with controls discussed in this chapter, except that to estimate effect ℓ , the covariate-adjusted DID goes from the last period before treatment T_0 to $T_0 + \ell$, instead of going from period one to two. Similarly, the ℓ th pre-trend estimator is a covariate-adjusted DID from T_0 to $T_0 - \ell$. As we will use it in the empirical example in the next section, here is the syntax of the `did_multiplegt_dyn` command, to produce an event-study graph controlling non-parametrically for some variables taking a small number of values:

```
did_multiplegt_dyn outcome groupid timeid treatment,  
effects(#) placebo(#) trends_nonparam(var_names)
```

where `effects(#)` is the number of event-study effects to be estimated, and `placebo(#)` is the number of pre-trend estimators.

Commands computing debiased-machine-learning DID estimators. To compute debiased-machine-learning DID estimators, such as $DID_{X,\text{dr-ml}}$, one can use any available command for DML estimation originally developed for cross-sectional data, defining the outcome as $Y_{g,2} - Y_{g,1}$.

4.1.6 Application to the compulsory licensing example

Using the `moser_voena_didtextbook` dataset, reestimate the TWFE ES regression we estimated in Chapter 3 (patents on a treatment group FE, year FEs, indicators for having been exposed to treatment for 1, 2, ..., 21 years, and indicators for being a treatment group subclass 1, 2, ..., 18 years before 1918), but controlling for a full set of interactions between subclasses' number of patents in 1900 and year FEs. Are pre-trend estimators less significant with these controls? Are the estimated event-study effects with controls very different from those without controls?

```
reghdfe patents reltimeminus* reltimeplus*, absorb(year#patents1900 treatmentgroup)
cluster(subclass)
```

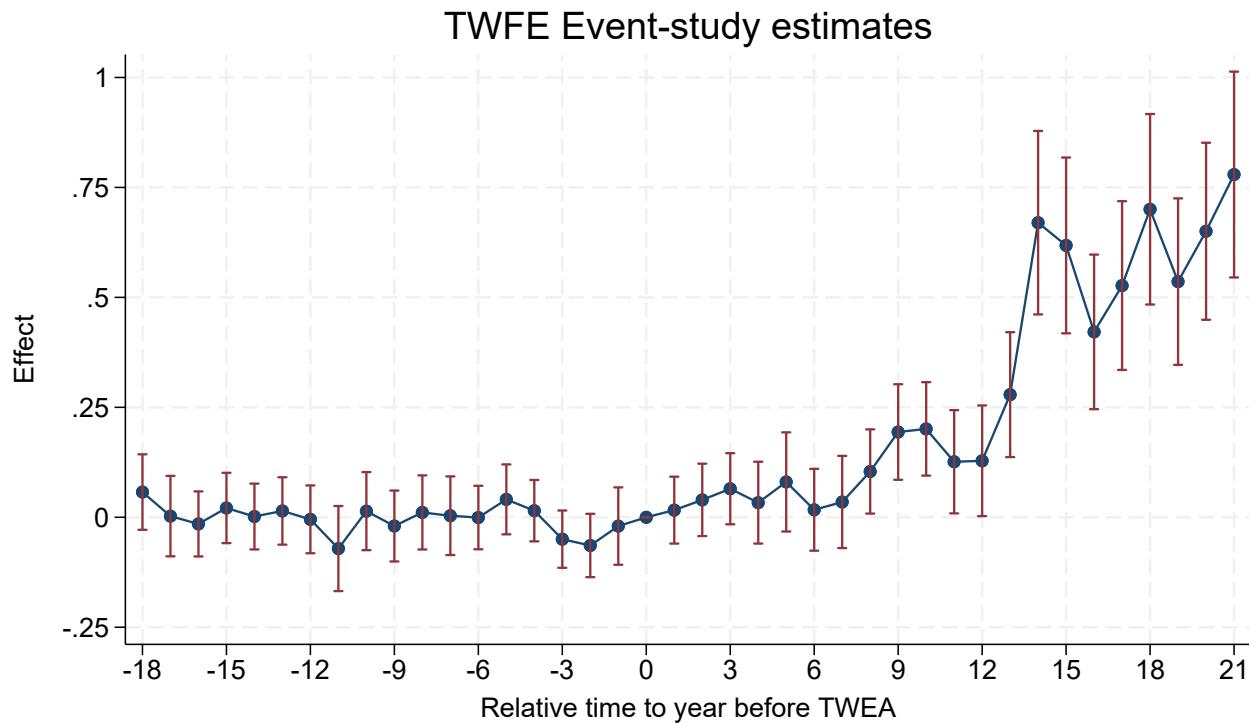
Figure 4.1 below shows the regression's coefficients. The pre-trends and event-study estimators are similar to those in Figure 3.2. An F-test that all pre-trends coefficients are equal to zero is still very strongly rejected ($p\text{-value} < 0.001$), so the data does not lend more support to the parallel-trends assumption conditional on baseline patents than to the unconditional assumption. Now, use the `did_multiplegt_dyn` command to perform the same estimation. Are results very different from those you obtained with the TWFE ES regression with controls?

```
did_multiplegt_dyn patents subclass year tweas,
effects(21) placebo(18) trends_nonparam(patents1900)
```

The results are not shown on Figure 4.1, because they are very close to those from the TWFE regressions. While TWFE regressions with controls can be biased, this is not a concern in this example. This might be because the specific control variable considered here, while correlated with the treatment group indicator, does not vary much: 88% of subclasses had no patent in

1900. In this application, the pre-trend estimators in Figure 3.2 lend support to the unconditional parallel-trends assumption, so there may not be a compelling argument to move away from TWFE estimators without controls. Of course, when there is evidence that the unconditional parallel-trends assumption is violated, controlling for some covariates may be appealing.

Figure 4.1: Effects of compulsory licensing on US innovation, according to a TWFE ES regression with controls



Note: This figure shows the estimated effects of compulsory licensing on patents, as well as pre-trends estimates, using years 1900 to 1939 of the data from Moser and Voena (2012), and a TWFE event-study regression with a full set of interactions between subclasses' number of patents in 1900 and year fixed effects. Standard errors are clustered at the patent subclass level. 95% confidence intervals are shown in red.

4.2 Interactive fixed effects, synthetic controls, and synthetic DID

Interactive fixed effects (IFE), synthetic control (SC), and synthetic DID (SD) estimators are very popular alternatives to DID estimators. We open this section with a list of advantages and drawbacks of these methods relative to DIDs, and a list of topics on which we think that further

research would be particularly useful.

Advantages and drawbacks.

1. **Applicability.** IFE, SC, and SD estimators are not as widely applicable as DIDs, because they require a large number of pre-treatment periods T_0 .
2. **Identification.** IFE, SC, and SD estimators rely on a factor model for the untreated outcome that is substantially weaker than the TWFE model underlying DID estimators. At the same time, IFE, SC, and SD estimators assume that the factor model holds for a large number of time periods, while DID estimators can be used if the more restrictive TWFE model holds for a few periods around the treatment-adoption date.
3. **Estimation.** The SD estimator can sometimes be more precise than the DID estimator. On the other hand, unlike the DID estimator, IFE, SC, and SD estimators require choosing tuning parameters, and there is no theoretically justified way of choosing those parameters.
4. **Placebo tests.** While it is easy to placebo test the parallel-trends and no-anticipation assumptions underlying DIDs via pre-trend tests, testing the assumptions underlying IFE, SC, and SD estimators is less straightforward, an issue that has received little attention.

Avenues for future research.

1. **Applicability.** Proposing estimators relying on a factor model and with proven guarantees even if T_0 is fixed would be a major improvement.⁵ More simulation studies assessing the number of pre-treatment periods that are necessary to reliably use IFE, SC, and SD estimators would also be useful.
2. **Estimation.** It would be very useful to propose theoretically-justified and data-driven choices of the tuning parameters needed to compute IFE, SC, and SD estimators.
3. **Placebo tests.** Providing placebo tests of the IFE, SC, and SD identifying assumptions would be useful. As was done for DID, it is also important to run realistic simulations to

⁵Imbens, Kallus and Mao (2021) and Brown and Butts (2023) propose such estimators, but while their results allow T_0 to be fixed, they require other strong assumptions.

assess if those placebo tests have enough power to detect violations of identifying assumptions that could lead to substantial biases.

Disclaimer. The literature on IFE, SC, and SD is huge, and what follows is very far from being exhaustive: our goal is simply to present the main references in that literature, as well as the most popular Stata and R commands to compute those estimators. For much more thorough reviews of this literature, see for instance Abadie (2021) or Arkhangelsky and Imbens (2024).

4.2.1 Interactive fixed effects

A TWFE model with interactive fixed effects. As shown in (2.4), Assumption PT is equivalent to assuming that the untreated outcome follows a TWFE model:

$$Y_{g,t}(0) = \alpha_g + \gamma_t + \varepsilon_{g,t}, \quad (4.13)$$

with $E(\varepsilon_{g,t}) = 0$. Hsiao, Ching and Ki Wan (2012), Gobillon and Magnac (2016), and Xu (2017) instead assume that the untreated outcome follows a TWFE model augmented with interactive fixed effects (TWFE-IFE):

$$Y_{g,t}(0) = \alpha_g + \gamma_t + \sum_{r=1}^R \lambda_{g,r} f_{t,r} + \varepsilon_{g,t}, \quad (4.14)$$

with $E(\varepsilon_{g,t}) = 0$. $(f_{t,r})_{r \in \{1, \dots, R\}}$ is a vector of period-specific shocks affecting all groups, like the period FE γ_t . The key difference between (4.13) and (4.14) is that (4.14) has group-specific coefficients $(\lambda_{g,r})_{r \in \{1, \dots, R\}}$ in front of the common shocks $(f_{t,r})_{r \in \{1, \dots, R\}}$, as if some group FEs were interacted with period FEs. For instance, assume that $f_{t,1}$ represents the state of the economy at period t . Then, how would you describe groups with a large positive value of $\lambda_{g,1}$? Groups with a smaller positive value of $\lambda_{g,1}$? Groups with a negative value of $\lambda_{g,1}$?

Groups that have a large positive value of $\lambda_{g,1}$ are such that their outcome without treatment responds very positively to a positive economic shock and very negatively to a negative shock:

their outcome is very sensitive to the economic environment. Groups with a smaller positive value of $\lambda_{g,1}$ are less sensitive to the economic environment, and groups with a negative coefficient are counter-cyclical. [Does the parallel-trends assumption hold under \(4.14\)?](#)

By allowing groups to respond differently to common shocks, (4.14) allows for differential trends between groups:

$$E(Y_{g,t}(0) - Y_{g,t-1}(0)) = \gamma_t - \gamma_{t-1} + \sum_{r=1}^R \lambda_{g,r} (f_{t,r} - f_{t-1,r}).$$

One could also assume that (4.14) holds with control variables, thus further relaxing the parallel-trends assumption. $(f_{t,r})_{r \in \{1, \dots, R\}}$ are often called the factors, $(\lambda_{g,r})_{r \in \{1, \dots, R\}}$ are often called the loadings, and IFE models are also often called factor models.

Estimating the $(ATT_\ell)_{\ell \in \{1, \dots, T_1\}}$ and the ATT under a TWFE-IFE model. To estimate the $(ATT_\ell)_{\ell \in \{1, \dots, T_1\}}$ and the ATT under (4.14), Xu (2017), building upon Bai (2009), proposes the following algorithm:⁶

1. Restrict the sample to control groups. In that subsample:

- (a) Run a regression of $Y_{g,t}$ on group and period FEs, let $(\hat{\alpha}_g)_{g:D_g=0}$ and $(\hat{\gamma}_t)_{t \in \{1, \dots, T\}}$ denote the estimated FEs, and let $\hat{\varepsilon}_{g,t}$ be the residuals.
- (b) Find $(\hat{f}_{t,r})_{r \in \{1, \dots, R\}, t \in \{1, \dots, T\}}$ and $(\hat{\lambda}_{g,r})_{r \in \{1, \dots, R\}, g:D_g=0}$, the minimizers of

$$\sum_{g:D_g=0} \sum_{t=1}^T \left(\hat{\varepsilon}_{g,t} - \sum_{r=1}^R \hat{\lambda}_{g,r} \hat{f}_{t,r} \right)^2,$$

see this chapter's appendix for details on how to solve this minimization problem.⁷

⁶Hsiao et al. (2012) and Gobillon and Magnac (2016) propose closely-related procedures.

⁷With covariates $X_{g,t}$, one needs to iterate Step (b) and a step where one estimates the coefficients on $X_{g,t}$ until convergence. See Xu (2017) for details.

2. Then, for each treated group, run a regression of $(Y_{g,t} - \hat{\gamma}_t)_{t \in \{1, \dots, T_0\}}$ on an intercept and $(\hat{f}_{t,r})_{r \in \{1, \dots, R\}, t \in \{1, \dots, T_0\}}$, and let $(\hat{\alpha}_g, (\hat{\lambda}_{g,r})_{r \in \{1, \dots, R\}})_{g: D_g=1}$ denote the coefficients from those G_1 regressions.

3. Finally, for $\ell \in \{1, \dots, T_1\}$, let

$$\hat{\beta}_\ell^{\text{ife}} = \frac{1}{G_1} \sum_{g: D_g=1} \left(Y_{g, T_0 + \ell} - \left(\hat{\alpha}_g + \hat{\gamma}_{T_0 + \ell} + \sum_{r=1}^R \hat{\lambda}_{g,r} \hat{f}_{T_0 + \ell, r} \right) \right)$$

be an estimator of ATT_ℓ , and let

$$\hat{\beta}^{\text{ife}} = \frac{1}{T_1} \sum_{t=T_0+1}^T \hat{\beta}_\ell^{\text{ife}}$$

be an estimator of ATT .

Intuition for the TWFE-IFE estimator. What is the intuition underlying $\hat{\beta}_\ell^{\text{ife}}$? In particular, what are the commonalities and differences between $\hat{\beta}_\ell^{\text{ife}}$ and $\hat{\beta}_\ell^{\text{imp}}$?

Under (4.14), one can use

$$\hat{\alpha}_g + \hat{\gamma}_t + \sum_{r=1}^R \hat{\lambda}_{g,r} \hat{f}_{t,r}$$

to impute the unobserved counterfactual outcome $Y_{g,t}(0)$ of treated (g, t) cells. Then, one can use

$$Y_{g,t} - \left(\hat{\alpha}_g + \hat{\gamma}_t + \sum_{r=1}^R \hat{\lambda}_{g,r} \hat{f}_{t,r} \right),$$

the difference between the cell's observed and imputed outcome, to estimate its treatment effect. Thus, $\hat{\beta}_\ell^{\text{ife}}$ is similar to $\hat{\beta}_\ell^{\text{imp}}$, the imputation estimator of ATT_ℓ discussed in Chapter 3. The difference between the two estimators is that $\hat{\beta}_\ell^{\text{ife}}$ uses a TWFE-IFE model to impute the missing counterfactual outcome, while $\hat{\beta}_\ell^{\text{imp}}$ just uses a TWFE model.

Asymptotic theory for the TWFE-IFE estimator. Liu et al. (2024) show that $\hat{\beta}_\ell^{\text{ife}}$ is consistent when G_0 , T_0 , and G_1 diverge to $+\infty$. Xu (2017) proposes a parametric bootstrap to estimate its variance. However, to our knowledge the validity of this procedure has not been

established yet. In fact, we are not aware of an asymptotic-normality result on $\hat{\beta}_\ell^{\text{ife}}$: the results of Bai (2009), for instance, apply to other parameters.⁸

How many pre-treatment periods should one have to use a TWFE-IFE estimator?

The asymptotic approximation underlying the TWFE-IFE estimator requires that G_0 , T_0 , and G_1 diverge to $+\infty$. This is in contrast with the asymptotic approximation underlying the TWFE estimator, which only requires that G_0 and G_1 diverge to $+\infty$. Then, one should have a reasonably large number of pre-treatment periods T_0 to reliably use the TWFE-IFE estimator, but how large should that number be? The simulations in Xu (2017) suggest that with 15 pre-treatment periods, the confidence interval of the TWFE-IFE estimator has close to nominal coverage. However, those simulations are not calibrated to a real dataset, and there may exist realistic settings where more pre-treatment periods are needed to have good coverage. Assessing the coverage of the confidence interval of the TWFE-IFE estimator in simulations calibrated to real datasets is an interesting avenue for future research.

Weak factors.* Asymptotic results for factor models typically rely on a “strong factor assumption”, which requires that the loadings $\lambda_{g,r}$ and factors $f_{t,r}$ have sufficient variation across g and over t , respectively (Bai, 2009). If that assumption fails IFE estimators can be biased and their confidence intervals (CIs) can be size-distorted. This is an issue, as applied researchers have no way of knowing ex-ante if in their empirical application, factors are strong or weak. Armstrong, Weidner and Zeleniev (2022) propose IFE estimators and confidence intervals that remain valid when factors are weak. However, their estimators and confidence intervals have not been extended yet to a TWFE-IFE imputation estimator allowing for heterogeneous treatment effects, in the spirit of that of Xu (2017).

How to choose the number of factors? Another difference between the TWFE-IFE and TWFE estimators is that unlike the latter, the former requires choosing a tuning parameter, namely the number of factors R . One may use cross-validation to choose R . For instance, one

⁸Bai and Ng (2021) also propose treatment effect estimators under a factor model for the untreated outcome, and they derive an asymptotic-normality result for their estimator. However, their estimator is not computed yet by a Stata or R command, unlike that of Xu (2017).

estimates the TWFE-IFE model with $R = 1$ using all periods except period T , one uses the model to predict the outcome of untreated groups at T , and one computes the model's mean squared error (MSE). One repeats the procedure holding out periods $T - 1, \dots, 1$, and one finally computes the model's average MSE across all periods. Then, one repeats the same procedure for $R = 2$, for $R = 3$, etc. Finally, one chooses the value of R yielding the lowest MSE. Xu (2017) finds that the TWFE-IFE estimator with a cross-validated R works well in simulations, though the paper does not provide a theoretical justification for that procedure.

How to test the TWFE-IFE model? The parallel-trends assumption can be tested via pre-trend tests. [How can one run a similar test of the TWFE-IFE model?](#)

For $\ell \in \{-1, \dots, -(T_0 - 1)\}$, one could define the following pre-trends estimator:

$$\widehat{\beta}_\ell^{\text{ife}} = \frac{1}{G_1} \sum_{g:D_g=1} \left(Y_{g,T_0+\ell} - \left(\widehat{\alpha}_g + \widehat{\gamma}_{T_0+\ell} + \sum_{r=1}^R \widehat{\lambda}_{g,r} \widehat{f}_{T_0+\ell} \right) \right).$$

However, this estimator will be mechanically close to zero, even if (4.14) fails. This placebo compares the actual and imputed outcome of treatment groups before treatment, but treatment groups' outcomes at all pre-treatment periods are used to estimate the imputation parameters, so the actual and imputed outcomes will mechanically be close. Instead, one can recompute the TWFE-IFE estimator, pretending that the treatment took place at period $T_0 + 1 - P$ instead of $T_0 + 1$ (Liu et al., 2024). Then, for $\ell \in \{1, \dots, P\}$, $\widehat{\beta}_\ell^{\text{ife}}$ is an actual placebo estimator, comparing treatment groups' actual and imputed outcomes before treatment, at time periods that were not used to estimate the imputation parameters. This testing strategy comes with one caveat. The TWFE-IFE estimator can only be used when the number of pre-treatment periods is large, but having P placebos reduces to $T_0 - P$ the number of pre-treatment periods one can use to estimate the loadings of the treated groups. This can be an issue if T_0 is already not very large.

Computation: Stata and R commands to compute the TWFE-IFE estimator of Xu (2017). The estimators $(\widehat{\beta}_\ell^{\text{ife}})_{\ell \in \{1, \dots, T_1\}}$ are computed by the `fetc` Stata (see Liu, Wang, Xu,

Liu and Liu, 2022b) and R (see Liu, Wang, Xu, Liu and Liu, 2022a) commands. In Stata, the command's syntax is

```
fect outcome, treat(treatment) unit(groupid) time(timeid) method("ife") r(#) tol(1e-4),
```

where one inputs the number of factors as the argument of the `r` option. `tol(1e-4)` sets the tolerance level to 10^{-4} .⁹ If one wants to use cross-validation to choose the number of factors, one should add the `cv` option. Then, cross-validation is performed to determine the optimal number of factors from $R = 1$ to the number inputted in the `r` option.

4.2.2 Synthetic control and synthetic DID

4.2.2.1 Synthetic control

The synthetic control (SC) estimator. The SC estimator was originally proposed for settings with only one treated group, so in most of this section we assume that $G_1 = 1$, and that group G is the treated group. To simplify the exposition, we also assume that $T_0 + 1 = T$: group G is only treated at period T . Then, the ATT reduces to $E(Y_{G,T}(1) - Y_{G,T}(0))$, and we need to estimate the missing counterfactual outcome $Y_{G,T}(0)$. For that purpose, Abadie, Diamond and Hainmueller (2010), building upon Abadie and Gardeazabal (2003), propose to use

$$\hat{Y}_{G,T}(0) = \sum_{g=1}^{G-1} \hat{w}_g Y_{g,T},$$

where $(\hat{w}_1, \dots, \hat{w}_{G-1})$ are the minimizers of

$$\sum_{t=1}^{T-1} \left(Y_{G,t} - \sum_{g=1}^{G-1} w_g Y_{g,t} \right)^2, \quad (4.15)$$

subject to $w_g \geq 0$ and $\sum_{g=1}^{G-1} w_g = 1$.¹⁰ Give a verbal description of the imputed counterfactual outcome $\hat{Y}_{G,T}(0)$.

⁹By default, the tolerance level is set to 10^{-5} , but this tolerance level yields very noisy results in our empirical application.

¹⁰This minimization can be solved easily as an instance of quadratic programming.

$\hat{Y}_{G,T}(0)$ is the period- T outcome of the weighted average of control groups whose period 1 to $T-1$ outcomes are closest to that of the treatment group, hereafter referred to as the synthetic control. Then,

$$\hat{\beta}^{\text{sc}} = Y_{G,T} - \hat{Y}_{G,T}(0).$$

$\hat{\beta}^{\text{sc}}$ is one of the estimators considered by Abadie et al. (2010), sometimes referred to as the canonical SC estimator (Doudchenko and Imbens, 2016). When there is more than one post-treatment period, one can compute event-study SC estimators, by letting $\hat{\beta}_\ell^{\text{sc}} = Y_{G,T_0+\ell} - \hat{Y}_{G,T_0+\ell}(0)$ for $\ell \in \{1, \dots, T_1\}$, where $\hat{Y}_{G,T_0+\ell}(0) = \sum_{g=1}^{G-1} \hat{w}_g Y_{g,T_0+\ell}$.

Other versions of the SC estimator. Abadie et al. (2010) also consider cases where other pre-intervention characteristics than the pre-treatment outcomes $(Y_{g,1}, \dots, Y_{g,T-1})$ are used in the objective function in (4.15), and where not all pre-treatment outcomes are used in that objective function. However, there is currently little guidance as to how one should choose the pre-intervention characteristics on which to match the treated and controls. This gives researchers opportunities to engage in specification searching (Ferman, Pinto and Possebom, 2020). The canonical SC estimator is not subject to that concern.

Intuition underlying the SC estimator. To convey the intuition underlying the SC estimator, let us assume that the TWFE-IFE model in (4.14) holds. If $Y_{g,t}(0) = \alpha_g + \gamma_t + \sum_{r=1}^R \lambda_{g,r} f_{t,r} + \varepsilon_{g,t}$ and $\sum_{g=1}^{G-1} \hat{w}_g Y_{g,t} \approx Y_{G,t}$ for all $t \leq T_0$, what relationship might there be between the FEs and loadings of the synthetic control and of the treated group?

Under the TWFE-IFE model in (4.14),

$$\sum_{g=1}^{G-1} \hat{w}_g Y_{g,t} \approx Y_{G,t} \Rightarrow \sum_{g=1}^{G-1} \hat{w}_g (\alpha_g + \sum_{r=1}^R \lambda_{g,r} f_{t,r} + \varepsilon_{g,t}) \approx \alpha_G + \sum_{r=1}^R \lambda_{G,r} f_{t,r} + \varepsilon_{G,t}.$$

Then, the fact that the level of the outcome is similar in the treated group and in the synthetic

control might indicate that their FEs are similar:

$$\sum_{g=1}^{G-1} \hat{w}_g \alpha_g \approx \alpha_G. \quad (4.16)$$

Similarly, the fact that their outcome paths are similar might indicate that they react similarly to the common shocks $(f_{t,r})_{r \in \{1, \dots, R\}}$, thus implying that their loadings may also be similar:

$$\sum_{g=1}^{G-1} \hat{w}_g \lambda_{g,r} \approx \lambda_{G,r}. \quad (4.17)$$

Then,

$$\begin{aligned} E(\hat{Y}_{G,T}(0)) &= E\left(\sum_{g=1}^{G-1} \hat{w}_g Y_{g,T}\right) \\ &= E\left(\sum_{g=1}^{G-1} \hat{w}_g \left(\alpha_g + \gamma_T + \sum_{r=1}^R \lambda_{g,r} f_{T,r} + \varepsilon_{g,T}\right)\right) \\ &\approx \alpha_G + \gamma_T + \sum_{r=1}^R \lambda_{G,r} f_{T,r} \\ &= E(Y_{G,T}(0)), \end{aligned}$$

where the approximation follows from (4.16) and (4.17), and from omitting the estimation error in \hat{w}_g combined with $E(\varepsilon_{g,T}) = 0$.

Identifying assumptions underlying the SC estimator. Consistent with this intuition, Ferman (2021) exhibits conditions under which the SC estimator is asymptotically unbiased, under the TWFE-IFE model in (4.14), when both T_0 and G diverge to $+\infty$. Those conditions require that there exist positive weights w_g^* summing to one, for which the weighted average of the control groups' FEs and loadings reproduces the treatment group's FE and loadings when G goes to infinity:

$$\sum_{g=1}^{G-1} w_g^* \alpha_g - \alpha_G \rightarrow 0 \quad (4.18)$$

$$\sum_{g=1}^{G-1} w_g^* \lambda_{g,r} - \lambda_{G,r} \rightarrow 0. \quad (4.19)$$

Moreover, w_g^* should not be too sparse: the number of control groups for which $w_g^* > 0$ should not be too low. One may think of the TWFE-IFE model in (4.14), (4.18), and (4.19) as the identifying assumptions underlying the SC estimator.

Comparing the identifying assumptions underlying the SC and DID estimator. SC requires that there exists a weighted average of control groups that has the same counterfactual trend as the treated group. Instead, DID requires that the simple average of control groups has the same counterfactual trend as the treated group, which is stronger. However, DID estimators with controls can be consistent in instances where the SC estimator is not. For instance, if

$$Y_{g,t}(0) = \alpha_g + \gamma_t + \lambda_g t + \varepsilon_{g,t}, \quad E(\varepsilon_{g,t}) = 0,$$

meaning that the untreated outcome follows a TWFE model with group-specific linear trends, then a DID estimator with linear trends is always consistent for the ATT, while the SC estimator is inconsistent if λ_G does not belong to the convex hull of $(\lambda_g)_{g \in \{1, \dots, G-1\}}$, as then (4.19) fails (Arboleda Cárcamo, 2024). In their Footnote 4, Arkhangelsky, Athey, Hirshberg, Imbens and Wager (2021) sketch an extension of their SC and SD estimators to models with covariates. However, this extension does not allow for heterogeneous treatment effects, and it is unclear whether it is applicable to control variables whose dimensionality grows with the sample size, as is the case of group-specific linear trends.

If one is ready to assume a TWFE-IFE model, then why not use a TWFE-IFE estimator? The identifying assumptions underlying the SC estimator ((4.14), (4.18), and (4.19)) are stronger than those underlying the TWFE-IFE estimator ((4.14)). Yet, there are two arguments to still use the SC estimator, when it comes to estimation. First, unlike the TWFE-IFE estimator, the SC estimator does not require taking a stance on the number of factors. Second, TWFE-IFE estimates the factors and loadings in (4.14), before estimating the treatment's effect. However, estimating the factors and loadings may require imposing stronger assumptions than those needed to estimate the treatment's effect. Instead, the SC estimator bypasses the estimation of the factors and loadings, and directly estimates the treatment effect in a way that accounts for the factor structure in (4.14). This might explain why in simulations, the SC estimator sometimes performs better under (4.14) than estimators trying to estimate the factor structure (Arkhangelsky et al., 2021).

Inference with one treated group. With only one treated group, conducting inference is not straightforward. Abadie et al. (2010) propose a placebo approach where one compares the SC

estimator to the quantiles of the distribution of placebo SC estimators, computed assuming that the treated group was actually one of the control groups. Typically, this type of randomization inference procedure can be justified by assuming that the treated group was chosen at random, but this may not be a plausible assumption in the settings where SC estimators are used. Instead, Chernozhukov, Wüthrich and Zhu (2021) propose an inference procedure that is valid under more realistic assumptions, namely if the TWFE-IFE model in (4.14) holds and the error terms $\varepsilon_{g,t}$ are stationary and not too serially correlated, and if the number of time periods T goes to infinity. In the simple example we consider, where group G is only treated at period T , a p-value of the sharp null that $Y_{G,T}(1) - Y_{G,T}(0) = 0$ can be obtained as follows:

1. Compute $(\tilde{w}_1, \dots, \tilde{w}_{G-1})$, the minimizers of

$$\sum_{t=1}^T \left(Y_{G,t} - \sum_{g=1}^{G-1} w_g Y_{g,t} \right)^2.$$

2. Let $\hat{u}_t^{\text{sc}} = Y_{G,t} - \sum_{g=1}^{G-1} \tilde{w}_g Y_{g,t}$, and let $\hat{F}(x) = \frac{1}{T} \sum_{t=1}^T \mathbb{1}\{|\hat{u}_t^{\text{sc}}| \leq x\}$.

3. The p-value is equal to $1 - \hat{F}(\hat{u}_T^{\text{sc}})$.

In Step 1, one solves almost the same minimization problem as in (4.15), except that one finds the weighted average of control groups whose period 1 to T outcomes are closest to that of the treatment group. Under the null that $Y_{G,T}(1) - Y_{G,T}(0) = 0$, $Y_{G,T} = Y_{G,T}(1) = Y_{G,T}(0)$, so period T can be used in the computation of the SC weights, as if period T too was a pre-treatment period. Then, one computes the SC prediction errors \hat{u}_t^{sc} , and finally the p-value in Step 3 is just the proportion of SC prediction errors \hat{u}_t^{sc} whose absolute value is larger than \hat{u}_T^{sc} . Intuitively, we reject the null that $Y_{G,T}(1) - Y_{G,T}(0) = 0$ if an SC estimator computed under that null yields a much larger prediction error at period T , the period where the null is imposed, than at other periods. Including period T in Step 1 is key: if that period is not included, $|\hat{u}_T^{\text{sc}}|$ will be mechanically larger than the other residuals, just because the weights were chosen to minimize the squared residuals from period 1 to $T-1$ but not at T .¹¹

¹¹One can proceed similarly to test the null that $Y_{G,T}(1) - Y_{G,T}(0) = \theta$ for any $\theta \in \mathbb{R}$, except that one minimizes

$$\sum_{t=1}^T \left(Y_{G,t} - \theta \mathbb{1}\{t = T\} - \sum_{g=1}^{G-1} w_g Y_{g,t} \right)^2$$

Placebo tests with the SC estimator. An intuitive concern with the SC estimator goes as follows. By construction, the outcome of the synthetic control will match that of the treated group almost perfectly in the pre-treatment periods. Then, the outcomes of the two groups may start diverging when the treatment group gets treated, but is this due to the treatment's effect, or just to the fact that post-treatment outcomes are not used to compute the SC weights? To assess whether this concern is legitimate or not, one can recompute the SC estimator, pretending that the treatment took place at period $T_0 + 1 - P$ instead of $T_0 + 1$. If placebo SC estimators at periods $T_0 + 1 - P, \dots, T_0$ are small in comparison to the actual estimators at periods $T_0 + 1, T_0 + 2, \dots$, this lends credibility to the SC estimator. As with the TWFE-IFE estimator, a caveat of this strategy is that having P placebos reduces to $T_0 - P$ the number of pre-treatment periods one can use to estimate the synthetic control weights, which can be an issue if T_0 is already not very large to begin with.

Computation: Stata and R commands to compute SC estimators. The SC estimator is computed by the `synth` Stata (see Abadie et al., 2011) and R (see Diamond and Hainmueller, 2023) commands. In Stata, the command's syntax is

```
synth outcome predictors, trunit(#) trperiod(#),
```

where `predictors` is the list of pre-intervention characteristics on which the synthetic control should replicate the treated group. One inputs the identifier of the treated group into `trunit`, and the time period at which the intervention starts into `trperiod`.

4.2.2.2 Synthetic DID

The synthetic DID (SD) estimator. In this section, we no longer assume that there is only one treated group. In settings with several treated groups, Arkhangelsky et al. (2021) propose an SD estimator, that combines features of the SC and DID estimators. To simplify the exposition, let us assume that groups 1 to G_0 are the control groups, while groups $G_0 + 1$ to G are the treated groups. The SD estimator $\hat{\beta}^{\text{sd}}$ is the coefficient on $D_{g,t}$ in a TWFE regression of $Y_{g,t}$ on group and period FEs and $D_{g,t}$, weighted by $\hat{w}_g^{\text{sd}}\hat{\theta}_t^{\text{sd}}$. $\hat{w}_g^{\text{sd}} = 1/G_1$ for all $g \in \{G_0 + 1, \dots, G\}$,

in Step 1: under the null that $Y_{G,T}(1) - Y_{G,T}(0) = \theta$, $Y_{G,T} - \theta = Y_{G,T}(0)$. Finally, assuming that $Y_{G,T}(1) - Y_{G,T}(0)$ is non-random (and thus equal to the ATT), a $1 - \alpha$ level confidence interval for the ATT can be constructed, as the unions of all θ such that the test's p-value is larger than or equal to α .

while $(\hat{w}_0^{\text{sd}}, \hat{w}_1^{\text{sd}}, \dots, \hat{w}_{G_0}^{\text{sd}})$ are the minimizers of

$$\sum_{t=1}^{T_0} \left(\frac{1}{G_1} \sum_{g=G_0+1}^G Y_{g,t} - w_0 - \sum_{g=1}^{G_0} w_g Y_{g,t} \right)^2 + \xi^2 T_0 \sum_{g=1}^{G_0} w_g^2 \quad (4.20)$$

subject to $w_g \geq 0 \quad \forall g \geq 1, \sum_{g=1}^{G_0} w_g = 1$. ξ is a tuning parameter that Arkhangelsky et al. (2021) recommend to set at $(G_1 T_1)^{1/4} \hat{\sigma}$, with $\hat{\sigma}$ the standard deviation of the first differences $Y_{g,t} - Y_{g,t-1}$ across all control groups in the pre-treatment periods. Similarly, $\hat{\theta}_t^{\text{sd}} = 1/T_1$ for all $t \in \{T_0 + 1, \dots, T\}$, and $(\hat{\theta}_0^{\text{sd}}, \hat{\theta}_1^{\text{sd}}, \dots, \hat{\theta}_{T_0}^{\text{sd}})$ are the minimizers of

$$\sum_{g=1}^{G_0} \left(\frac{1}{T_1} \sum_{t=T_0+1}^T Y_{g,t} - \theta_0 - \sum_{t=1}^{T_0} \theta_t Y_{g,t} \right)^2 \quad (4.21)$$

subject to $\theta_t \geq 0 \quad \forall t \geq 1, \sum_{t=1}^{T_0} \theta_t = 1$. To compute SD estimators of the event-study effects ATT_ℓ , one can run a TWFE regression of $Y_{g,t}$ on group and period FEs and the relative time indicators $(1\{t = T_0 + \ell\})_{\ell \in \{1, \dots, T_1\}}$, weighted by $\hat{w}_g^{\text{sd}} \hat{\theta}_t^{\text{sd}}$.

Commonalities and differences with the usual DID estimator and with the SC estimator. What is the commonality between $\hat{\beta}^{\text{sd}}$ and the DID estimator $\hat{\beta}^{\text{fe}}$? What is the difference between those two estimators? What is the commonality between $\hat{\beta}^{\text{sd}}$ and the SC estimator? What is the difference between those two estimators?

$\hat{\beta}^{\text{sd}}$ and $\hat{\beta}^{\text{fe}}$ are computed using the same TWFE regression, except that to obtain the former, we use weights. Accordingly, one can show that $\hat{\beta}^{\text{sd}}$ is a weighted DID estimator:

$$\hat{\beta}^{\text{sd}} = \frac{1}{G_1 T_1} \sum_{g=G_0+1}^G \sum_{t=T_0+1}^T Y_{g,t} - \frac{1}{G_1} \sum_{g=G_0+1}^G \sum_{t=1}^{T_0} \hat{\theta}_t^{\text{sd}} Y_{g,t} - \left(\frac{1}{T_1} \sum_{g=1}^{G_0} \sum_{t=T_0+1}^T \hat{w}_g^{\text{sd}} Y_{g,t} - \sum_{g=1}^{G_0} \sum_{t=1}^{T_0} \hat{w}_g^{\text{sd}} \hat{\theta}_t^{\text{sd}} Y_{g,t} \right).$$

The weights are the product of a group-specific component \hat{w}_g^{sd} and a period-specific component $\hat{\theta}_t^{\text{sd}}$. The group-specific component of the weights, \hat{w}_g^{sd} , is just $1/G_1$ for the treated groups, and for the control groups it is similar to the SC weights in the previous section, up to two differences. First, by allowing for an intercept \hat{w}_0^{sd} , the weighted average of control groups' outcomes may systematically differ from the average of treated group's outcomes, by \hat{w}_0^{sd} , but that difference

has to remain stable from period 1 to T_0 , thus enforcing the parallel-trends assumption in the pre-treatment periods. Second, the objective function includes a regularization penalty $\xi^2 T_0 \sum_{g=1}^{G_0} w_g^2$, to ensure that the minimizers of (4.20) are not too sparse. The period-specific component $\hat{\theta}_t^{\text{sd}}$ is just $1/T_1$ for the treated periods. For the pre-treatment periods, one finds the weighted average of pre-treatment outcomes that replicates best control groups' post-treatment outcomes. In the same way that the SC estimator gives more weight to control groups whose pre-period outcomes "resemble" that of treated groups, the SD estimator gives more weight to pre-treatment periods when control groups' outcomes resemble their post-treatment outcomes.

SC estimator with many treated groups. Arkhangelsky et al. (2021) also consider another estimator, the coefficient on $D_{g,t}$ in a TWFE regression of $Y_{g,t}$ on period FEs and $D_{g,t}$, weighted by \hat{w}_g^{sc} , where the weights \hat{w}_g^{sc} solve almost the same minimization problem as (4.20), without the intercept. That second estimator is a close analogue of the SC estimator applied to the average outcome of the treated groups. The only difference is the penalization in the objective function.

Identifying assumptions underlying the SD estimator. To simplify the discussion, let us assume that the TWFE-IFE model in (4.14) holds, though the results in Arkhangelsky et al. (2021) still hold under a relaxation of that model. Then, the authors show that the SD estimator relies on the following identifying assumption: there exist intercepts w_0^* and θ_0^* , and positive weights w_g^* and θ_t^* summing to one such that

$$\frac{1}{G_1} \sum_{g=G_0+1}^G \lambda_{g,r} - \left(w_0^* + \sum_{g=1}^{G_0} w_g^* \lambda_{g,r} \right)$$

goes to zero faster than $1/(G_1 T_1)^{1/4}$,

$$\frac{1}{T_1} \sum_{t=T_0+1}^T f_{t,r} - \left(\theta_0^* + \sum_{t=1}^{T_0} \theta_t^* f_{t,r} \right)$$

does not diverge "too quickly", and

$$\begin{aligned} & \frac{1}{T_1 G_1} \sum_{g=G_0+1}^G \sum_{t=T_0+1}^T \left(\sum_{r=1}^R \lambda_{g,r} f_{t,r} \right) - \frac{1}{G_1} \sum_{g=G_0+1}^G \sum_{t=1}^{T_0} \theta_t^* \left(\sum_{r=1}^R \lambda_{g,r} f_{t,r} \right) \\ & - \left(\frac{1}{T_1} \sum_{g=1}^{G_0} \sum_{t=T_0+1}^T w_g^* \left(\sum_{r=1}^R \lambda_{g,r} f_{t,r} \right) - \sum_{g=1}^{G_0} \sum_{t=1}^{T_0} w_g^* \theta_t^* \left(\sum_{r=1}^R \lambda_{g,r} f_{t,r} \right) \right) \end{aligned}$$

goes to zero faster than $1/\sqrt{G_1 T_1}$. The first condition requires that the control groups should not be too different from the treated ones: up to an intercept and asymptotically, it should be possible to perfectly replicate the loadings of the treated group using a convex combination of control groups. The second condition imposes a similar but weaker requirement on the similarity of the pre- and post-treatment periods: it should be possible to replicate “not too poorly” the factors of the post-treatment periods using a weighted average of pre-treatment ones. Finally, the third condition requires that with the weights w_g^* and θ_t^* , a weighted DID of $(\sum_{r=1}^R \lambda_{g,r} f_{T,r})$, the factor-model part of $Y_{g,t}(0)$, goes to zero, at a faster rate than in the first condition. This last condition shows that by weighting both the control groups but also the pre-treatment periods, the SD estimator inherits a kind of double-robustness property. It is consistent either if the loadings of treated and control groups are “sufficiently similar”, or if the factors of pre- and post-treatment periods are “sufficiently similar”. Like a related condition underlying the SC estimator, that third condition can still fail if the untreated outcome follows a TWFE model with linear trends that differ for the treated and untreated groups.

Asymptotic theory for the SD estimator. Arkhangelsky et al. (2021) show that the SD estimator is consistent and asymptotically normal. This is a stronger result than those available for the SC estimator, in part because the authors assume that G_0 , T_0 , and $G_1 T_1$ diverge to $+\infty$, while the results on the SC estimator we have discussed so far only assume that G_0 and/or T_0 diverge to $+\infty$. On top of the aforementioned identifying assumptions, their result also relies on other assumptions, restated informally as follows:

1. G_0 and T_0 are of “similar magnitudes”, and G_0 is “much larger” than G_1 and T_1 .
2. The vectors of errors $(\varepsilon_{g,t})_{t \in \{1, \dots, T\}}$ are independent and identically distributed across groups, and follow a normal distribution.
3. Up to a logarithmic term, ξ is of order $(G_1 T_1)^{1/4}$.

Inference. Arkhangelsky et al. (2021) propose to use a block bootstrap, where one draws groups with replacement from the original sample, one recomputes the SD estimator in each bootstrap sample, and one finally uses the variance of the estimator across bootstrap samples to estimate the variance of $\widehat{\beta}^{\text{sd}}$. While valid, this procedure might be computationally costly, so

the authors also propose a less-computationally intensive jackknife procedure, that is valid but can yield a conservative variance estimator.

Placebo tests with the SD estimator. As for the SC estimator, it is important to placebo-test the identifying assumptions of the SD estimator. Otherwise, it could be that the treated group's outcomes diverge from their counterfactual in the post-treatment periods, just because the treated (g, t) cells are precisely those that were not used to construct the counterfactual. For that purpose, just pretending that the treatment took place at period $T_0 + 1 - P$ instead of $T_0 + 1$ will not work, because then periods $T_0 + 1 - P$ to T_0 are still used to estimate the period-specific component of the weights. Instead, one can keep only the first T_0 periods, and compute the SD estimator on this restricted dataset, pretending that treatment took place at $T_0 + 1 - P$. Again, this testing strategy reduces the number of pre-treatment periods one can use in the estimation, which may be an issue if T_0 was not very large to begin with.

Simulation evidence. On top of the identifying assumptions of the SD estimator, the asymptotic result in Arkhangelsky et al. (2021) relies on some strong conditions, like the errors' normality assumption, as well as rate conditions on G_0 , T_0 , and $G_1 T_1$ whose applicability may be hard to gauge in applications. Then, assessing if the resulting confidence intervals have close-to-nominal coverage in realistic settings is important. The authors conduct two simulation studies, calibrated to datasets representative of those typically used for panel data studies. Outcomes are generated from a TWFE-IFE model where the loadings and factors, as well as the errors' variance-covariance matrix, are estimated from some actual outcomes in these datasets, and the treatment assignment mechanism is also inspired from actual policies that took place over the data period (while satisfying the identifying assumptions of the SD estimator). In their first simulation, $G = 50$, $T = 40$, $G_0 \in \{40, 49\}$, and $T_0 \in \{30, 39\}$. In their second simulation, $G = 111$, $T = 48$, $G_0 = 101$, and $T_0 = 38$. In both cases, the authors find that their confidence intervals generally have good coverage. While those results are encouraging, further simulation studies could be useful, in particular to assess whether the SD estimator can reliably be used with less than 30 pre-treatment periods. It is also important to assess the power of placebo tests of the SD identifying assumptions in realistic simulations, and if violations that those tests do not have power to detect could lead to substantial biases. Finally, assessing if the SD confidence

intervals have good coverage when errors are not normally distributed could be valuable.

Computation: Stata and R commands to compute the SD estimator. The SD estimator is computed by the `sdid` Stata (see Pailañir, Clarke and Ciccia, 2022) and `synthdid` R (see Hirshberg, 2023) commands. In Stata, the command's syntax is

```
sdid outcome groupid timeid treatment.
```

The SD event-study estimators are computed by the `sdid_event` Stata command. In Stata, the command's syntax is

```
sdid_event outcome groupid timeid treatment.
```

The SC estimator with many treated groups of Arkhangelsky et al. (2021) is also computed by `sdid`, when the `method(sc)` option is specified. Similarly, SC event-study estimators with many treated groups are computed by `sdid_event`, when the `method("sc")` option is specified.

4.2.2.3 Application to the compulsory licensing example

TWFE-IFE estimators. Execute

```
net install fect, from(https://raw.githubusercontent.com/xuyiqing/fect\_stata/master/)
replace
```

to install the `fect` package, and install `_gwtmean`, an auxiliary package used by `fect`, from SSC.

Use `fect` to determine by cross-validation the optimal number of factors, between 1, 2, 3, and 4, in a TWFE-IFE model estimated on the `moser_voena_didtextbook` dataset.

```
fect patents, treat(twea) unit(subclass) time(year) method("ife") r(4) tol(1e-4)
cv
```

According to the cross-validation, choosing two factors is optimal.¹² Estimate the TWFE-IFE

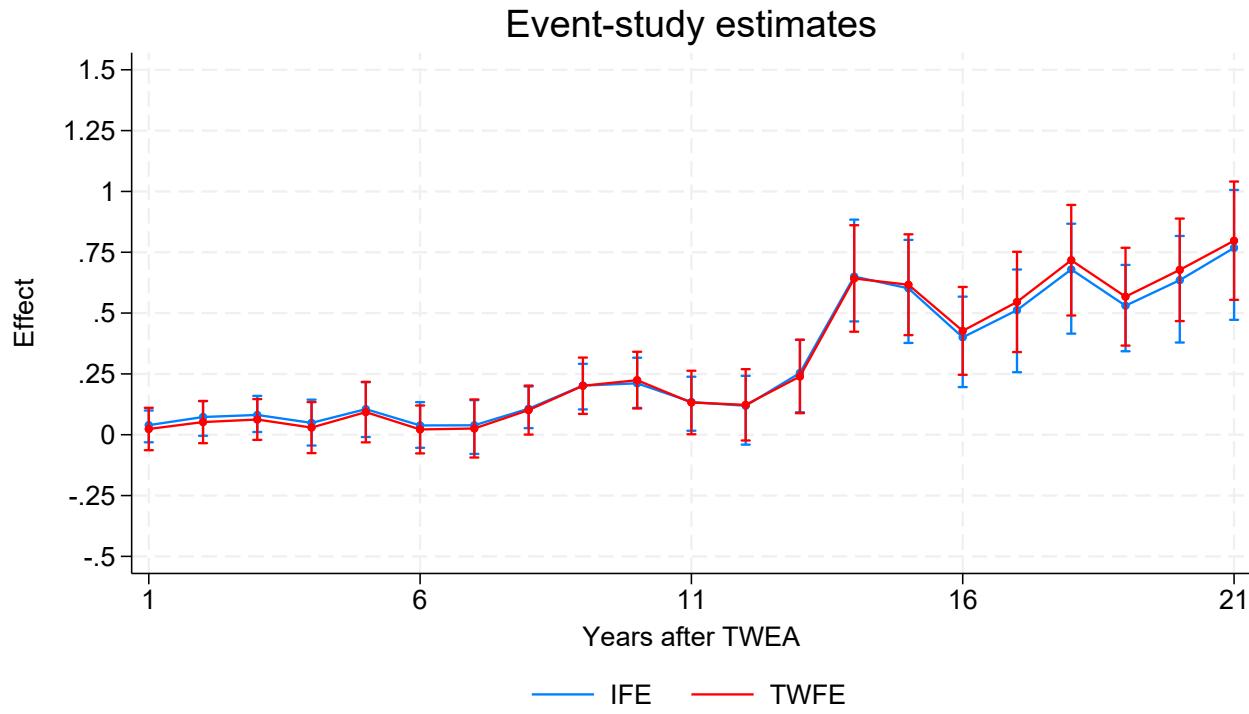
¹²If instead one performs the cross-validation on the treatment groups only (then one has to restrict attention to the pre-periods), the optimal number of factors is equal to one.

model with two factors, using the `se` option to request that standard errors be computed.

```
fект patents, treat(twea) unit(subclass) time(year) method("ife") r(2) tol(1e-4)
se
```

On a Dell Optiplex 7090 desktop computer, with an i7-11700T processor and Stata/MP 19, the command's run time is slightly more than nine minutes. Figure 4.2 below shows that the TWFE-IFE estimators are very close to the TWFE ones. Some event-study TWFE-IFE estimators are less noisy than the corresponding TWFE estimators, but others are more noisy. The TWFE-IFE estimator of the ATT, which is just the simple average of the event-study estimators, is equal 0.297, which is very close to the TWFE estimator of the ATT. The standard error of the TWFE-IFE estimator, 0.049, is 28% larger than that of the TWFE estimator of the ATT.

Figure 4.2: TWFE estimators and TWFE-IFE estimators of Xu (2017), on the data of Moser and Voena (2012)



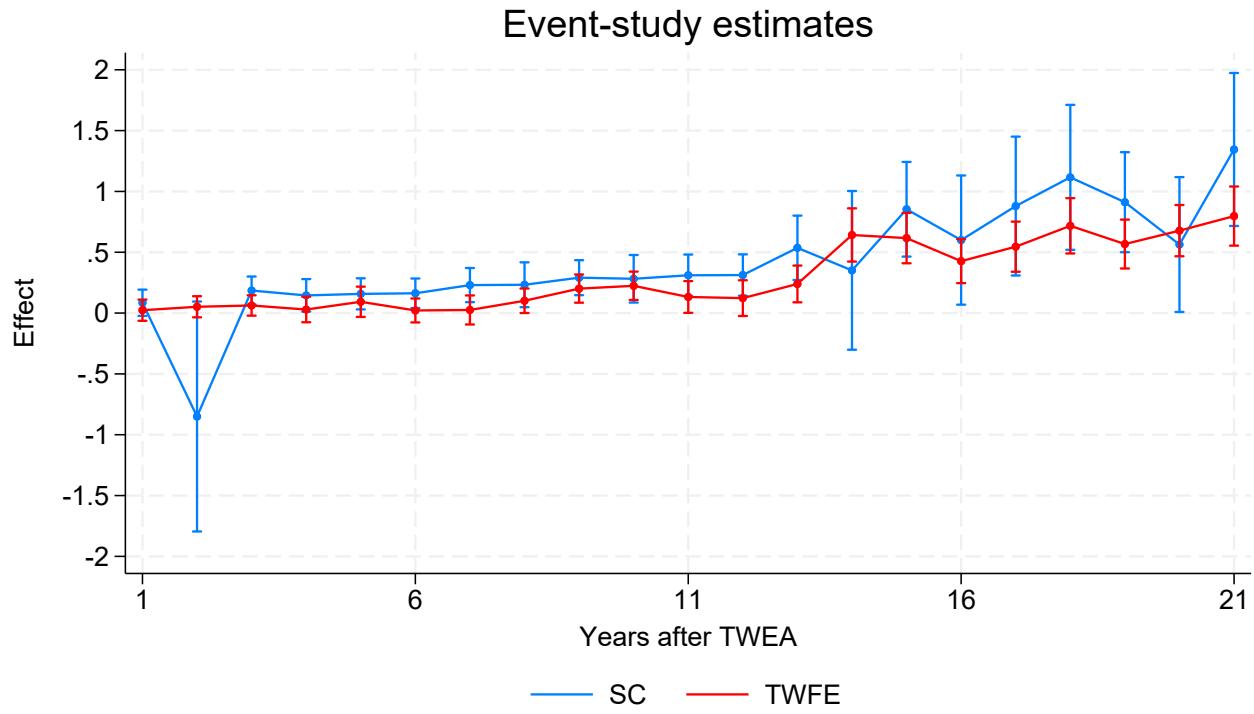
Note: This figure shows the estimated effects of compulsory licensing on patents, using years 1900 to 1939 of the data from Moser and Voena (2012), and two estimators: the TWFE event-study regressions in (3.6) and the TWFE-IFE imputation estimator. Standard errors are clustered at the patent subclass level. The 95% confidence intervals rely on a normal approximation.

SC estimators of Arkhangelsky et al. (2021). Rather than using the original SC estimator of Abadie et al. (2010), applied to the average of treated subclasses, we use the SC estimator of Arkhangelsky et al. (2021). The reason is simply that the confidence intervals attached to the SC estimator of Abadie et al. (2010) rely on the assumption that the treated units were chosen at random, which, as shown in Section 3.2.1.1, is clearly implausible in this application. Instead, the confidence intervals of the SC estimator of Arkhangelsky et al. (2021) rely on the TWFE-IFE model in (4.14) and a large sample approximation. Install the `sdid_event` package from SSC, and use it to compute event-study SC estimators using the `moser_voena_didtextbook` dataset, with 200 bootstrap replications.

```
sdid_event patents subclass year tweas, method("sc") brep(200)
```

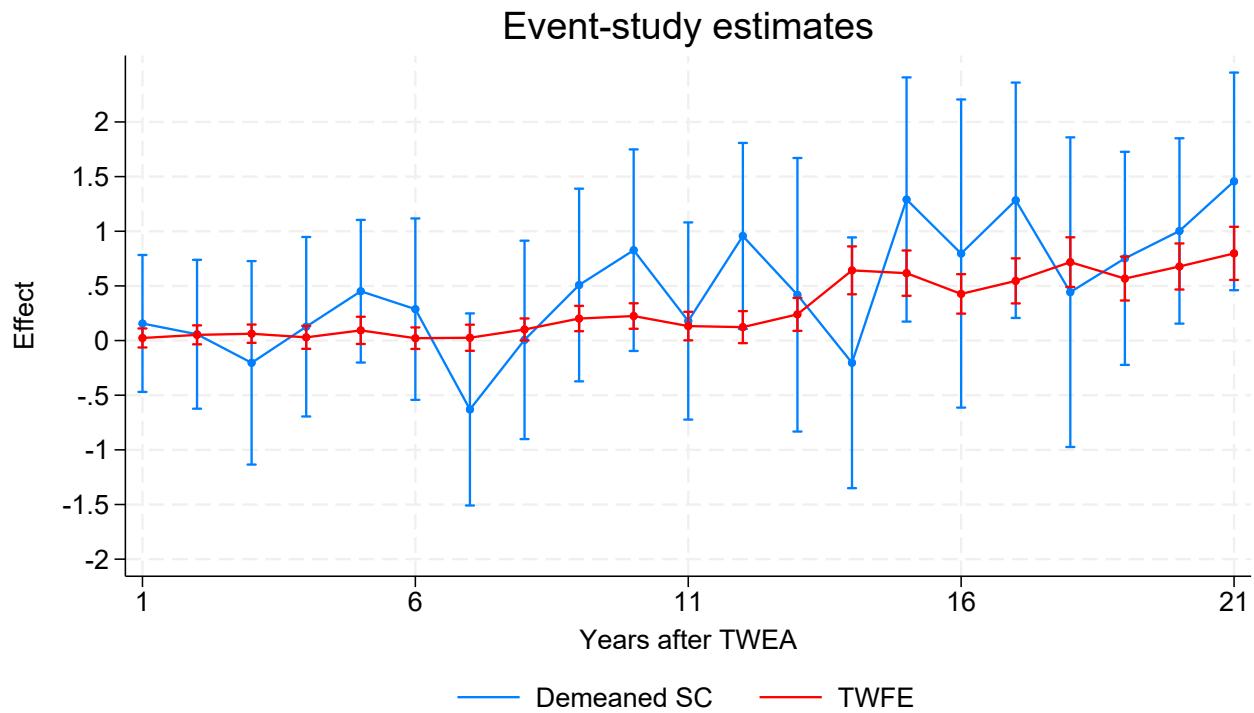
The command's run time is slightly less than 20 minutes. Figure 4.3 below shows that the event-study SC estimators are overall slightly larger and much noisier than the TWFE estimators. The estimated effect of two years of exposure to treatment stands out as implausibly negative and extremely noisy. This is due to the fact that the weights \hat{w}_g^{sd} are extremely sparse (only 1.4% of control groups receive a strictly positive weight) and have a very large outlier (one control group receives 49.5% of the SC weights). Those extreme weights may in part be due to the fact that the level of the outcome is very different in the treatment and control groups during the pre-periods (see Section 3.2.1.1). Then, to match the level of the outcome for treated subclasses, the SC estimator can only give a strictly positive weight to the control subclasses with the lowest outcome levels. However, Figure 4.4 shows that the demeaned SC estimators are also extremely noisy, while those estimators should in principle work well even if the level of the outcome is very different in the treatment and control groups during the pre-periods.

Figure 4.3: TWFE estimators and Synthetic Control estimators of Arkhangelsky et al. (2021), on the data of Moser and Voena (2012)



Note: This figure shows the estimated effects of compulsory licensing on patents, using years 1900 to 1939 of the data from Moser and Voena (2012), and the TWFE event-study regressions in (3.6) and the synthetic control estimator of Arkhangelsky et al. (2021). Standard errors are clustered at the patent subclass level. The 95% confidence intervals rely on a normal approximation.

Figure 4.4: TWFE estimators and Synthetic Control estimators of Arkhangelsky et al. (2021) applied to demeaned outcome, on the data of Moser and Voena (2012)



Note: This figure shows the estimated effects of compulsory licensing on patents, using years 1900 to 1939 of the data from Moser and Voena (2012), and the TWFE event-study regressions in (3.6) and the demeaned synthetic control estimator of Arkhangelsky et al. (2021). Standard errors are clustered at the patent subclass level. The 95% confidence intervals rely on a normal approximation.

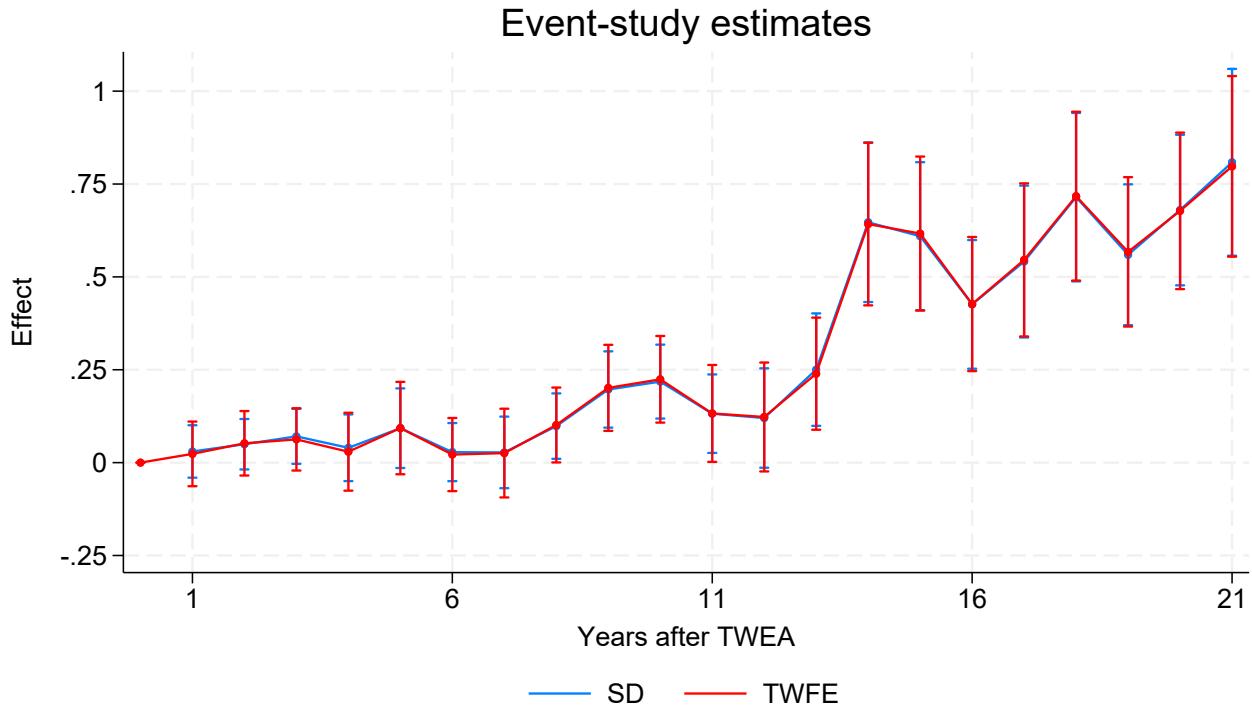
SD estimators. Use the `sdid_event` package to compute event-study SD estimators using the `moser_voena_didtextbook` dataset, with 200 bootstrap replications.

```
sdid_event patents subclass year tweas, brep(200)
```

The command's run time is slightly less than one hour and a half. Figure 4.5 below shows that the event-study SD estimators are undistinguishable from the TWFE ones. Some event-study

SD estimators are less noisy than the corresponding TWFE estimators, but others are more noisy. The standard error of the SD estimator of the ATT, which is just the simple average of the event-study estimators, is equal to 0.043, which is 12% larger than the standard error of the TWFE estimator of the ATT.

Figure 4.5: TWFE estimators and Synthetic Control estimators, on the data of Moser and Voena (2012)



Note: This figure shows the estimated effects of compulsory licensing on patents, using years 1900 to 1939 of the data from Moser and Voena (2012), and the TWFE event-study regressions in (3.6) and the synthetic DID estimator. Standard errors are clustered at the patent subclass level. The 95% confidence intervals rely on a normal approximation.

Conclusion. In this application, the SC estimator does not perform very well, while the TWFE-IFE and SD estimators are similar to, but noisier than, the TWFE estimator. Thus, there is no compelling argument to move away from the TWFE estimator, all the more so as the pre-trend estimators on Figure 3.2 suggest that the parallel-trends assumption is plausible. The TWFE-IFE, SC, and SD estimators remain appealing alternatives in applications where pre-trend tests indicate a violation of the parallel-trends assumption.

4.3 Bounded differential trends

Bounded differential trends. Rambachan and Roth (2023) propose an alternative relaxation of the parallel-trends condition. Let us assume that Assumption ND holds, and $T = 3$, $T_0 = 2$.

Assumption BDT (*Bounded differential trends*) *There is a positive real number M such that*

$$\begin{aligned} & \left| E \left[\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,3}(0) - Y_{g,2}(0)) \right] - E \left[\frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,3}(0) - Y_{g,2}(0)) \right] \right| \\ & \leq M \left| E \left[\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,2}(0) - Y_{g,1}(0)) \right] - E \left[\frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,2}(0) - Y_{g,1}(0)) \right] \right|. \end{aligned} \quad (4.22)$$

Assumption BDT allows treated and control groups to experience differential trends, but requires that their period-two-to-three differential trend be bounded in absolute value by some constant M times their period-one-to-two differential trend. Thus, period-two-to-three and period-one-to-two differential trends can be different but should not be too different, where M indexes how large the difference can be. Note that with $M = 0$, Assumption BDT is equivalent to parallel-trends from period 2 to 3. Similarly, if

$$\left| E \left[\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,2}(0) - Y_{g,1}(0)) \right] - E \left[\frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,2}(0) - Y_{g,1}(0)) \right] \right| = 0,$$

then Assumption BDT implies parallel trends from period 2 to 3, irrespective of the value of M .

Partial identification of the ATT under Assumption BDT.

Theorem 7 *In Design CLA, if $T = 3$ and $T_0 = 2$, under Assumptions NA and BDT,*

$$E(\hat{\beta}_1^{fe}) - M|E(\hat{\beta}_{-1}^{fe})| \leq ATT \leq E(\hat{\beta}_1^{fe}) + M|E(\hat{\beta}_{-1}^{fe})|.$$

We saw in Chapter 3 that $\hat{\beta}_1^{fe}$ in the TWFE ES regression in (3.6) is a DID comparing the outcome evolution of treated and control groups from period two to three, while $\hat{\beta}_{-1}^{fe}$ is a DID comparing their outcome evolutions from period two to one. Then, $E(\hat{\beta}_1^{fe})$ is equal to the ATT plus the differential outcome evolution that treated and controls would have experienced without treatment from period two to three. Under Assumption BDT, that differential evolution is included between $-M|E(\hat{\beta}_{-1}^{fe})|$ and $M|E(\hat{\beta}_{-1}^{fe})|$, hence the bounds for ATT.

Proof of Theorem 7.*

$$\begin{aligned}\hat{\beta}_1^{\text{fe}} &= \frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,3} - Y_{g,2}) - \frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,3} - Y_{g,2}) \\ &= \frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,3}(1) - Y_{g,2}(0)) - \frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,3}(0) - Y_{g,2}(0)) \\ &= \frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,3}(1) - Y_{g,3}(0)) + \frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,3}(0) - Y_{g,2}(0)) - \frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,3}(0) - Y_{g,2}(0)),\end{aligned}$$

where the last equality follows from adding and subtracting $Y_{g,3}(0)$. Taking expectations and rearranging,

$$\text{ATT} = E(\hat{\beta}_1^{\text{fe}}) - E \left[\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,3}(0) - Y_{g,2}(0)) - \frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,3}(0) - Y_{g,2}(0)) \right].$$

Finally, the result follows from Assumption BDT, and the fact that

$$\hat{\beta}_{-1}^{\text{fe}} = \frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,1}(0) - Y_{g,2}(0)) - \frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,1}(0) - Y_{g,2}(0)).$$

QED.

Estimation and inference. Given M , the lower and upper bounds in Theorem 7 can respectively be estimated by $\hat{\beta}_1^{\text{fe}} - M|\hat{\beta}_{-1}^{\text{fe}}|$ and $\hat{\beta}_1^{\text{fe}} + M|\hat{\beta}_{-1}^{\text{fe}}|$. Which condition should hold to have that 0 does not belong to the interval $[\hat{\beta}_1^{\text{fe}} - M|\hat{\beta}_{-1}^{\text{fe}}|, \hat{\beta}_1^{\text{fe}} + M|\hat{\beta}_{-1}^{\text{fe}}|]$?

One should have that $|\hat{\beta}_1^{\text{fe}}| > M|\hat{\beta}_{-1}^{\text{fe}}|$. Whenever $|\hat{\beta}_1^{\text{fe}}| \leq M|\hat{\beta}_{-1}^{\text{fe}}|$, 0 is included between the lower and upper bounds for the ATT, so we cannot reject $\text{ATT} = 0$. Of course, $0 \notin [\hat{\beta}_1^{\text{fe}} - M|\hat{\beta}_{-1}^{\text{fe}}|, \hat{\beta}_1^{\text{fe}} + M|\hat{\beta}_{-1}^{\text{fe}}|]$ is necessary but not sufficient to reject $\text{ATT} = 0$: one also needs to take into account the sampling error in the estimation of the bounds. Rambachan and Roth (2023) leverage results from the moment inequality literature (see Andrews et al., 2023) to construct an asymptotically valid confidence interval for ATT.

Sensitivity analysis. Practitioners may not have a good sense of which value of M they should choose. Rather than recommending a particular value, Rambachan and Roth (2023)

recommend that they conduct the following sensitivity analysis. Assume that $\hat{\beta}_1^{\text{fe}}$ is strictly positive and significantly different from zero. Under parallel trends, researchers would conclude that the treatment has a positive effect. To assess if that conclusion is robust to plausible violations of parallel trends, Rambachan and Roth (2023) propose to compute M^* , the lowest value of M such that 0 belongs to the confidence interval of ATT. $M^* = 5$ means that even under differential trends five times larger from period two to three than from period one to two, one can still conclude that the treatment had a positive effect: the researcher's conclusion is very robust to differential trends. On the other hand, $M^* = 0.2$ means that differential trends five times smaller from period two to three than from period one to two are enough for the researcher's conclusion to break down, thus suggesting that results are not robust to plausible differential trends.

Generalization to multiple time periods. With more than three time periods, if $T_0 \geq 2$ Assumption BDT can be generalized as follows: for all $t > T_0$, there is a positive real number M such that

$$\begin{aligned} & \left| E \left[\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,t}(0) - Y_{g,t-1}(0)) \right] - E \left[\frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,t}(0) - Y_{g,t-1}(0)) \right] \right| \\ & \leq M \max_{t' \in \{2, \dots, T_0\}} \left| E \left[\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,t'}(0) - Y_{g,t'-1}(0)) \right] - E \left[\frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,t'}(0) - Y_{g,t'-1}(0)) \right] \right|. \end{aligned}$$

Computation: Stata and R commands to conduct a sensitivity analysis à la Rambachan and Roth (2023). Confidence intervals for ATT under varying values of M are computed by the `honestdid` Stata (see Bravo et al., 2022) and R (see Rambachan, 2022) commands. In Stata, the command can for instance be run after estimating the TWFE ES regression in (3.6), absorbing the time FEs and the treatment group indicator. Then, one can simply run

```
honestdid, pre(1/T_0-1) post(T_0/T-1),
```

where the first and second option respectively indicate the indices of the pre-trend and event-study coefficients in the regression's vector of coefficients.

Application to the compulsory licensing example. Figure 3.2 in Chapter 3 suggests that the compulsory licensing of German patents in 1919 had a large effect on patenting in the US, in

patent subclasses of organic-chemistry where at least one German patent was licensed in 1919. In 1932, this effect becomes truly very large, but the corresponding estimator is a DID from 1918 to 1932, namely over a 14 years period. Then, one might worry that over such a long period, treated and control subclasses could have experienced different patenting evolutions, even in the absence of compulsory licensing. Therefore, we now conduct the sensitivity analysis proposed by Rambachan and Roth (2023), where $\hat{\beta}_1^{\text{fe}}$ is the 1918 to 1932 event-study estimator, and $\hat{\beta}_{-1}^{\text{fe}}$ is the symmetric 1918 to 1904 pre-trend estimator. Using the `moser_voena_didtextbook` dataset, estimate a TWFE ES regression computing $\hat{\beta}_1^{\text{fe}}$ and $\hat{\beta}_{-1}^{\text{fe}}$, and then run the sensitivity analysis of Rambachan and Roth (2023).

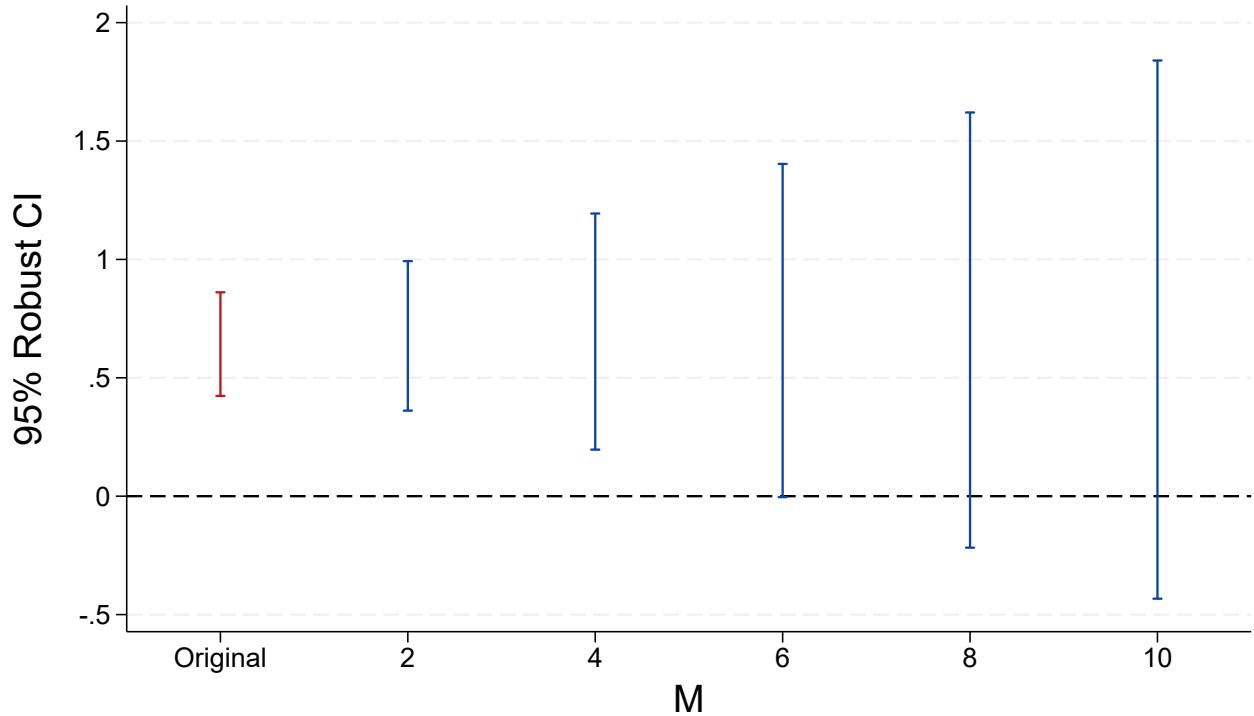
```

preserve
keep if year==1918|year==1904|year==1932
reghdfe patents reltimeminus14 reltimeplus14 ///
, absorb(year treatmentgroup) cluster(subclass)
honestdid, pre(1) post(2) mvec(2(2)10) coefplot ///
xtitle(M, size(large)) ytitle(95% Robust CI, size(large))
restore

```

The `mvec` option is used to indicate the values of M the command should use, and the `coefplot xtitle(M, size(large)) ytitle(95% Robust CI, size(large))` options request that results be put on a graph, shown in Figure 4.6 below. The figure shows that for the 1932 effect of compulsory licensing to become insignificant, one needs to allow for differential trends post treatment almost six times larger than the pre-treatment differential trends. This suggests that this effect is very robust to violations of parallel trends. Intuitively, this is due to the fact that in this application, $\hat{\beta}_1^{\text{fe}}$ is almost 15 times larger than $\hat{\beta}_{-1}^{\text{fe}}$. In applications where event-study estimators are not much larger than pre-trend estimators, the sensitivity analysis of Rambachan and Roth (2023) would indicate a lower robustness to differential pre-trends.

Figure 4.6: Sensitivity analysis à la Rambachan and Roth (2023), on the data of Moser and Voena (2012)



Note: This figure shows the sensitivity of the estimated effect of compulsory licensing on US innovation in 1932, estimated on the data of Moser and Voena (2012), to violations of parallel trends no larger than M times the 1904 to 1918 differential trend between treated and control subclasses. Standard errors are clustered at the patent subclass level.

Bibliographic notes. Assumption BDT is related to another bounded differential trends assumption that had been previously proposed by Manski and Pepper (2018). With $T = 3$ and $T_0 = 2$, their assumption requires that there is a positive real number \tilde{M} such that

$$\left| E \left[\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,3}(0) - Y_{g,2}(0)) \right] - E \left[\frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,3}(0) - Y_{g,2}(0)) \right] \right| \leq \tilde{M}.$$

With

$$\tilde{M} = M \left| E \left[\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,2}(0) - Y_{g,1}(0)) \right] - E \left[\frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,2}(0) - Y_{g,1}(0)) \right] \right|,$$

the previous display is equivalent to Assumption BDT.¹³ However, as

$$\left| E \left[\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,2}(0) - Y_{g,1}(0)) \right] - E \left[\frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,2}(0) - Y_{g,1}(0)) \right] \right|$$

has to be estimated, \tilde{M} is no longer a tuning parameter chosen ex-ante by the researcher but the product of a first-step estimation. Thus, Rambachan and Roth (2023) extend the ideas in Manski and Pepper (2018), by accounting for the estimation error in the differential pre-trends, which are a natural benchmark to calibrate the differential post-trends. Note also that Rambachan and Roth (2023) propose other relaxations of parallel-trends than bounded differential trends, see for instance their so-called “smoothness restrictions”.

4.4 Appendix*

4.4.1 Proof of Theorem 5

Let $\Delta Y_g = Y_{g,2} - Y_{g,1}$. As $T = 2$ and $D_{g,t} = D_g 1\{t = 2\}$, $\hat{\beta}_X^{\text{fe}}$ is numerically equivalent to the coefficient on D_g in the following regression:

$$\Delta Y_g = \hat{\gamma}_2 + X_{g,1}\hat{\theta} + \hat{\beta}_X^{\text{fe}}D_g + \hat{\epsilon}_{g,t}.$$

This regression is saturated in D_g . Therefore, it follows from Angrist (1998) that

$$\hat{\theta} = w_1\hat{\theta}_1 + (1 - w_1)\hat{\theta}_0 = \hat{\theta}_0 + w_1(\hat{\theta}_1 - \hat{\theta}_0),$$

where

$$\begin{aligned} \hat{\theta}_1 &= \frac{\frac{1}{G_1} \sum_{g:D_g=1} (X_{g,1} - \mu_{X,1})\Delta Y_g}{\sigma_{X,1}^2} \\ \hat{\theta}_0 &= \frac{\frac{1}{G_0} \sum_{g:D_g=0} (X_{g,1} - \mu_{X,0})\Delta Y_g}{\sigma_{X,0}^2}. \end{aligned}$$

¹³In their empirical application, Manski and Pepper (2018) let

$$\tilde{M} = \left| E \left[\frac{1}{G_1} \sum_{g:D_g=1} (Y_{g,2}(0) - Y_{g,1}(0)) \right] - E \left[\frac{1}{G_0} \sum_{g:D_g=0} (Y_{g,2}(0) - Y_{g,1}(0)) \right] \right|.$$

Under Assumption LPT,

$$E(\hat{\theta}_0) = \frac{\frac{1}{G_0} \sum_{g:D_g=0} (X_{g,1} - \mu_{X,0}) (\gamma_2 + X_{g,1}\theta)}{\sigma_{X,0}^2} = \theta,$$

and

$$E(\hat{\theta}_1) = \frac{\frac{1}{G_1} \sum_{g:D_g=1} (X_{g,1} - \mu_{X,1}) (\gamma_2 + X_{g,1}\theta + TE_{g,t})}{\sigma_{X,1}^2} = \theta + \beta_X.$$

Then,

$$\begin{aligned} \hat{\beta}_X^{\text{fe}} &= \frac{1}{G_1} \sum_{g:D_g=1} (\Delta Y_g - X_{g,1}\hat{\theta}) - \frac{1}{G_0} \sum_{g:D_g=0} (\Delta Y_g - X_{g,1}\hat{\theta}) \\ &= \frac{1}{G_1} \sum_{g:D_g=1} (\Delta Y_g - X_{g,1}\hat{\theta}_0) - \frac{1}{G_0} \sum_{g:D_g=0} (\Delta Y_g - X_{g,1}\hat{\theta}_0) \\ &\quad + w_1(\mu_{X,1} - \mu_{X,0})(\hat{\theta}_0 - \hat{\theta}_1). \end{aligned}$$

Finally, taking expectations,

$$E(\hat{\beta}_X^{\text{fe}}) = \text{ATT} + w_1(\mu_{X,1} - \mu_{X,0})\beta_X.$$

4.4.2 Comparison of variances of DID estimators with and without covariates

Formal result. Following the discussion in Subsection 4.1.3.5, we compare the asymptotic variances of DID estimators of the ATT with and without controls. To do so, we adopt a superpopulation framework where groups are i.i.d. We thus omit the index g below, and index expectations and variances by u to indicate that they are not conditional on the design. We also let D denote the treatment status at period 2, and we let $\Delta Y(0)$ and ΔY correspond respectively to $Y_2(0) - Y_1(0)$ and $Y_2 - Y_1$. Finally, we make an homoskedasticity assumption: $V_u(\Delta Y(0)|D=0, X) = \sigma^2$.

Assume that

$$E_u(\Delta Y(0)|D=1, X) = E_u(\Delta Y(0)|D=0, X), \quad (4.23)$$

an analogue of Assumption CPT in the super-population framework we consider in this appendix. Sant'Anna and Zhao (2020) show that under (4.23), the asymptotic variance of the most efficient ATT estimators is $V_X = V_u(\psi_X)$, where

$$\psi_X := \left(\frac{D}{p} - \frac{\pi(X)(1-D)}{(1-\pi(X))p} \right) (\Delta Y - m(X)) - \frac{D}{p} \text{ATT},$$

with $p := E_u(D)$, $\pi(X) := P_u(D = 1|X)$ and $m(X) := E_u[\Delta Y(0)|X]$. Now assume that

$$E_u(\Delta Y(0)|D = 1) = E_u(\Delta Y(0)|D = 0), \quad (4.24)$$

an analogue of (2.5). Under (4.24), the asymptotic variance of the standard DID estimator without covariates is $V = V_u(\psi)$, where

$$\psi := \left(\frac{D}{p} - \frac{1-D}{1-p} \right) (\Delta Y - m) - \frac{D}{p} \text{ATT},$$

where $m := E_u[\Delta Y(0)]$.

(4.23) and (4.24) are non-nested, but both conditions hold if

$$E_u(\Delta Y(0)|D = 1, X) = E_u(\Delta Y(0)|D = 0, X) = m, \quad (4.25)$$

meaning that counterfactual outcome trends are mean-independent of D and X . Under (4.25), the DID estimator with and without covariates are both consistent for the ATT. Under (4.25), $m(X) = m$.

We now show that under (4.25), $V_X \geq V$. First, under (4.25), $\psi_X = \psi + \xi$, where

$$\xi := (\Delta Y - m)(1 - D) \left(\frac{1}{1-p} - \frac{\pi(X)}{(1-\pi(X))p} \right).$$

Thus, because $E_u[\psi_X] = E_u[\psi] = 0$,

$$\begin{aligned} V_X - V &= 2E_u[\psi\xi] + E_u[\xi^2] \\ &= E \left[(\Delta Y - m)^2 (1 - D) \left\{ \frac{1}{1-p} - \frac{\pi(X)}{(1-\pi(X))p} \right\} \left\{ -\frac{2}{1-p} + \frac{1}{1-p} - \frac{\pi(X)}{(1-\pi(X))p} \right\} \right] \\ &= \sigma^2 E \left[(1 - D) \left\{ \left(\frac{\pi(X)}{(1-\pi(X))p} \right)^2 - \frac{1}{(1-p)^2} \right\} \right]. \end{aligned} \quad (4.26)$$

In the second line we used $D(1 - D) = 0$ to simplify $2\psi\xi$ while in the third, we used $(a - b)(-a -$

$b) = b^2 - a^2$, $(\Delta Y - m)^2(1 - D) = (\Delta Y(0) - m)^2(1 - D)$ and $V_u(\Delta Y(0)|D = 0, X) = \sigma^2$. Now,

$$\begin{aligned} E_u \left[(1 - D) \left(\frac{\pi(X)}{(1 - \pi(X))p} \right)^2 \right] &= \frac{1 - p}{p^2} E_u \left[\left(\frac{\pi(X)}{(1 - \pi(X))} \right)^2 \middle| D = 0 \right] \\ &\geq \frac{1 - p}{p^2} E_u \left[\frac{\pi(X)}{1 - \pi(X)} \middle| D = 0 \right]^2 \\ &= \frac{1}{p^2(1 - p)} E_u \left[\frac{\pi(X)(1 - D)}{1 - \pi(X)} \right]^2 \\ &= \frac{1}{1 - p} \\ &= E_u \left[\frac{1 - D}{(1 - p)^2} \right], \end{aligned}$$

where the second line follows by Jensen's inequality and the fourth and fifth hold by the law of iterated expectations. Combined with (4.26), this implies that $V_X \geq V$. Moreover, the inequality above is strict unless $V_u(\pi(X)/(1 - \pi(X))|D = 0) = 0$ or, equivalently, $V_u(\pi(X)|D = 0) = 0$.

Intuition. The result above is in contrast with unconditional RCTs, where controlling for covariates is not needed for identification but can improve precision. Instead, in DID estimation, if controlling for covariates is not needed for identification, then controlling for covariates cannot help with precision. Here is some intuition for this difference. To have that controlling for covariates is not needed for identification, (4.23) and (4.24) have to hold. For those two conditions to hold jointly for any joint distribution of (D, X) , (4.25) also has to hold: the outcome evolution without treatment should be mean-independent of the covariates, which then implies that controlling for covariates does not increase precision. Instead, in an unconditional RCT, the analogues of (4.23) and (4.24) in levels rather than in first difference do not have to hold for any joint distribution of (D, X) , because $D \perp\!\!\!\perp X$. Then, those conditions can hold even if X is not mean independent of $Y(0)$, in which case controlling for X may increase precision.

4.4.3 Details on the computation of the TWFE-IFE estimator

We explain here how to compute the $(\hat{f}_{t,r})_{r \in \{1, \dots, R\}, t \in \{1, \dots, T\}}$ and $(\hat{\lambda}_{g,r})_{g: D_g = 0, r \in \{1, \dots, R\}}$ in Step 1.(b) of the algorithm presented in Section 4.2.1. First, let Λ denote the $G_0 \times R$ matrix with typical (g, r) element $\lambda_{g,r}$. Similarly, let F denote the $T \times R$ matrix with typical (t, r) element $f_{t,r}$ and

$\boldsymbol{\varepsilon}$ be the $G_0 \times T$ matrix with typical (g, t) element $\hat{\varepsilon}_{g,t}$. Then, we seek to solve

$$\min_{\tilde{F}, \tilde{\Lambda}} \|\boldsymbol{\varepsilon} - \tilde{\Lambda} \tilde{F}'\|_F^2, \quad (4.27)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\|A\|_F = \text{trace}(A'A)^{1/2}$. To ensure identifiability, we assume hereafter that $R \leq p := \min(G_0, T)$. The solution to (4.27) is still not unique since for any invertible matrix A , $\tilde{\Lambda} \tilde{F}' = (\tilde{\Lambda} A)(\tilde{F} A'^{-1})'$. We follow Bai (2009) by imposing that $\Lambda' \Lambda$ is diagonal and $F' F = T \times I_R$, where I_R is the identity matrix of size R .

We now show how to solve (4.27) under such constraints on \tilde{F} and $\tilde{\Lambda}$. Let

$$\boldsymbol{\varepsilon} = \sum_{i=1}^p \sigma_i u_i v_i'$$

denote a singular value decomposition of $\boldsymbol{\varepsilon}$. Here $\sigma_1 \geq \dots \geq \sigma_p \geq 0$ are the singular values of $\boldsymbol{\varepsilon}$ and the $(u_i)_{i=1,\dots,p}$ (resp. the $(v_i)_{i=1,\dots,p}$) are orthonormal vectors of \mathbb{R}^{G_0} (resp. \mathbb{R}^T). Then, let us define $\hat{\Lambda} = [\sigma_1 u_1, \dots, \sigma_R u_R]/T^{1/2}$ and $\hat{F} := T^{1/2}[v_1, \dots, v_R]$. By construction, $\hat{\Lambda}' \hat{\Lambda}$ is diagonal and $\hat{F}' \hat{F} = T \times I_R$. Moreover, note that for any $G_0 \times R$ and $T \times R$ matrices $\tilde{\Lambda}$ and \tilde{F} , $\text{rank}(\tilde{\Lambda} \tilde{F}') \leq R$. Then, by Eckart–Young–Mirsky theorem,

$$\|\boldsymbol{\varepsilon} - \tilde{\Lambda} \tilde{F}'\|_F \geq \|\boldsymbol{\varepsilon} - \sum_{i=1}^R \sigma_i u_i v_i'\|_F = \|\boldsymbol{\varepsilon} - \hat{\Lambda} \hat{F}'\|_F. \quad (4.28)$$

Therefore, $(\hat{F}, \hat{\Lambda})$ solves (4.27) under the aforementioned constraints.

Part III

Beyond the classical design

Chapter 5

TWFE estimators outside of the classical design

In the classical design, $\hat{\beta}^{\text{fe}}$ is a simple DID estimator, and it is unbiased for the ATT under partly testable no-anticipation and parallel-trends assumption. However, the majority of papers estimating TWFE regressions do so in more complicated designs, with treatments that may be non-absorbing and/or non-binary, and where groups may experience several treatment changes, at different points in time. This chapter investigates what $\hat{\beta}^{\text{fe}}$ estimates in such designs.

Chapter's running example: the effect of newspapers on voters' participation in elections. Our running example in this chapter is Gentzkow, Shapiro and Sinkinson (2011). They use a US panel data set at the county \times presidential-election level, with 1,195 counties and from the 1872 to the 1928 election. They seek to test a conjecture in De Tocqueville (1850), that newspapers encourage citizens to participate more in democratic institutions. For that purpose, they let $Y_{g,t}$ denote the turnout rate in county g and the presidential election that took place in year t , they let $D_{g,t}$ denote the number of newspapers circulating in county g and year t , and they run a regression closely related to the TWFE regression in (3.1), described in details below. The design of that study is much more complicated than Design CLA: the number of newspapers is a non-binary treatment, it can increase or decrease over time, counties can experience several changes in their number of newspapers, at different points in time.

Dataset used in this chapter. To answer the green questions in this chapter, you need to use the `gentzkowetal_didtextbook` dataset, which contains the following variables:

- `cnty90`: a county identifier;
- `st`: a state identifier;
- `year`: an election-year identifier;
- `styr`: a state \times election-year identifier;
- `prestout`: the turnout rate in county g and election-year t ;
- `numdailies`: the number of daily newspapers circulating in county g and election-year t ;
- `changeprestashop`: the change in the turnout rate in county g between election-year $t - 1$ and t ;
- `changedailies`: the change in the number of daily newspapers circulating in county g between election-year $t - 1$ and t ;
- `lag_ishare_urb`: the lagged urbanization rate of county g at t (its urbanization at $t - 1$);
- `first_change`: the year when a county's number of daily newspapers changes for the first time;
- `same_treat_after_first_change`: an indicator for counties that keep the same number of newspapers for at least one period after their first change.

No dynamic effects. Throughout this chapter, we assume that the treatment has no dynamic effects, namely, we maintain Assumption ND. This is consistent with the TWFE regression in (3.1), where the current treatment $D_{g,t}$ is one of the independent variables, but the lagged treatments $D_{g,t-1}$, $D_{g,t-2}$ etc. are not part of the independent variables, thus implicitly ruling out dynamic treatment effects.¹ For instance, in the newspaper example, omitting the lagged

¹In a classical design, omitting the lagged treatments does not implicitly rule out dynamic effects, because $D_{g,t} = D_{g,t-k}$ whenever $D_{g,t-k} \neq 0$ for some $k > 0$. Then, the coefficient on $D_{g,t}$ captures the sum of the effects of the current and past treatment on the outcome.

treatments implicitly assumes that the number of newspapers available in county g in previous elections no longer affects the turnout rate in election-year t . Importantly, we will allow for dynamic effects in the following chapters.

Target parameter. To accommodate potentially non-binary treatments, we redefine the treatment effect $\text{TE}_{g,t}$. Under Assumption ND, for all (g, t) such that $D_{g,t} \neq 0$, we let

$$\text{TE}_{g,t} = \frac{E [Y_{g,t}(D_{g,t}) - Y_{g,t}(0)]}{D_{g,t}}.$$

Note that under Assumption ND and with a binary treatment, the definition of $\text{TE}_{g,t}$ above coincides with the definition we have used so far, which is why we recycle notation. [Interpret \$\text{TE}_{g,t}\$, in general and in the context of the newspaper example.](#)

$\text{TE}_{g,t}$ denotes the expected effect in cell (g, t) of moving the treatment from 0 to $D_{g,t}$, scaled by $D_{g,t}$. In other words, $\text{TE}_{g,t}$ is the slope of (g, t) 's potential outcome function, from 0 to its actual treatment $D_{g,t}$. In the newspaper example, $\text{TE}_{g,t}$ is the difference between the actual turnout rate in county g and year t and its counterfactual turnout rate without any newspaper, divided by its number of newspapers. Thus, $\text{TE}_{g,t}$ can be interpreted as an effect per newspaper. Let N_1 denote the number of (g, t) cells such that $D_{g,t} \neq 0$, namely the number of treated (g, t) cells. A natural target parameter is

$$\text{ATT} = \frac{1}{N_1} \sum_{(g,t): D_{g,t} \neq 0} \text{TE}_{g,t}.$$

ATT is the average, across all treated cells, of the slope of their potential outcome functions, from 0 to their actual treatment. In Design CLA, ATT reduces to the ATT parameter we have considered in previous chapters, which is why we recycle notation.

5.1 A decomposition of $\hat{\beta}^{\text{fe}}$

Let

$$W_{g,t} = \frac{\hat{u}_{g,t} D_{g,t}}{\sum_{(g',t'): D_{g',t'} \neq 0} \hat{u}_{g',t'} D_{g',t'}},$$

where $\hat{u}_{g,t}$ denotes the sample residual from a regression of $D_{g,t}$ on group and period FEs. In the newspaper example, how could you compute the residuals $\hat{u}_{g,t}$?

$\hat{u}_{g,t}$ is the residual from a regression of the number of newspapers in county g and year t on county and year FEs.

Theorem 8 If Assumptions NA, ND, and PT hold,

$$E [\hat{\beta}^{\text{fe}}] = \sum_{(g,t): D_{g,t} \neq 0} W_{g,t} \text{TE}_{g,t}. \quad (5.1)$$

Interpret Theorem 8, in general and in the context of the newspaper example.

Theorem 8 says that $\hat{\beta}^{\text{fe}}$ is unbiased for a weighted sum of the treatment effects $\text{TE}_{g,t}$, across all treated (g, t) cells, and where the treatment effect of cell (g, t) receives a weight equal to $W_{g,t}$. In the newspaper example, $\hat{\beta}^{\text{fe}}$ is unbiased for a weighted sum of the effects of newspapers on turnout across all county \times year cells with at least one newspaper.

What is the value of $\sum_{(g,t): D_{g,t} \neq 0} W_{g,t}$?

It directly follows from the definition of $W_{g,t}$ that $\sum_{(g,t): D_{g,t} \neq 0} W_{g,t} = 1$. Therefore, $\hat{\beta}^{\text{fe}}$ is unbiased

for a weighted sum of the treatment effects $\text{TE}_{g,t}$, across all treated (g, t) cells, with weights summing to one. As $\sum_{(g,t):D_{g,t} \neq 0} W_{g,t} = 1$, the average value of the weights $W_{g,t}$ across the N_1 treated cells is $1/N_1$.

Proof of Theorem 8.* First, we have

$$\begin{aligned} E[Y_{g,t}] &= E[Y_{g,t}(0)] + E[Y_{g,t}(D_{g,t}) - Y_{g,t}(0)] \\ &= E[Y_{g,t}(0)] + D_{g,t}E[Y_{g,t}(D_{g,t}) - Y_{g,t}(0)] / D_{g,t} \\ &= E[Y_{g,t}(0)] + D_{g,t}\text{TE}_{g,t}, \end{aligned} \quad (5.2)$$

with the convention that $0/0 = 0$. Then,

$$\hat{\beta}^{\text{fe}} = \frac{\sum_{g,t} \hat{u}_{g,t} Y_{g,t}}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}}. \quad (5.3)$$

The equality follows from the Frisch-Waugh-Lovell theorem, restated in the appendix of Chapter 3. Moreover, by the first-order conditions attached to an OLS regression, $\hat{u}_{g,t}$ is uncorrelated to all the group and time FEs: for all g' ,

$$\sum_{g,t} \hat{u}_{g,t} 1\{g = g'\} = 0 \Leftrightarrow \sum_{t=1}^T \hat{u}_{g',t} = 0, \quad (5.4)$$

and for all t' ,

$$\sum_{g,t} \hat{u}_{g,t} 1\{t = t'\} = 0 \Leftrightarrow \sum_{g=1}^G \hat{u}_{g,t'} = 0. \quad (5.5)$$

Finally,

$$\begin{aligned} E[\hat{\beta}^{\text{fe}}] &= \frac{\sum_{g,t} \hat{u}_{g,t} E[Y_{g,t}]}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} \\ &= \frac{\sum_{g,t} \hat{u}_{g,t} (E[Y_{g,t}(0)] + D_{g,t} \text{TE}_{g,t})}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} \\ &= \frac{\sum_{g,t} \hat{u}_{g,t} \alpha_g}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} + \frac{\sum_{g,t} \hat{u}_{g,t} \gamma_t}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} + \frac{\sum_{g,t} \hat{u}_{g,t} D_{g,t} \text{TE}_{g,t}}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} \\ &= \frac{\sum_{g=1}^G \alpha_g \sum_{t=1}^T \hat{u}_{g,t}}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} + \frac{\sum_{t=1}^T \gamma_t \sum_{g=1}^G \hat{u}_{g,t}}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} + \frac{\sum_{g,t} \hat{u}_{g,t} D_{g,t} \text{TE}_{g,t}}{\sum_{g,t} \hat{u}_{g,t} D_{g,t}} \\ &= \frac{\sum_{(g,t):D_{g,t} \neq 0} \hat{u}_{g,t} D_{g,t} \text{TE}_{g,t}}{\sum_{(g,t):D_{g,t} \neq 0} \hat{u}_{g,t} D_{g,t}} \\ &= \sum_{(g,t):D_{g,t} \neq 0} W_{g,t} \text{TE}_{g,t}. \end{aligned}$$

Justify each step of this derivation.

The first equality follows from (5.3) and the fact the design is conditioned upon, the second equality follows from (5.2), the third equality follows from (2.2), the fifth equality follows from (5.4) and (5.5), the last equality follows from the definition of $W_{g,t}$ **QED**.

5.2 $\hat{\beta}^{\text{fe}}$ may be biased for the ATT

Outside of the classical design, $\hat{\beta}^{\text{fe}}$ may be biased for the ATT under a no-anticipation and a parallel-trends assumption. Note that $E[\hat{\beta}^{\text{fe}}] = \text{ATT}$ for all possible values of the treatment effects $\text{TE}_{g,t}$ if and only if $W_{g,t} = 1/N_1$ for all treated (g, t) cells. If $W_{g,t}$ varies across treated cells, then $W_{g,t} > 1/N_1$ for some cells, $W_{g,t} < 1/N_1$ for some other cells, and $\hat{\beta}^{\text{fe}}$ may be biased for the ATT if the treatment effects of cells down- and up-weighted by $\hat{\beta}^{\text{fe}}$ differ. Having that $W_{g,t} = 1/N_1$ for all treated (g, t) cells is equivalent to having that $D_{g,t}\hat{u}_{g,t}$ is constant across those cells. Remember that $\hat{u}_{g,t}$ is the residual from a regression of $D_{g,t}$ on group and period FEs. One can show that the coefficient on the FE for group g in the regression is equal to the average treatment of group g across periods, denoted $D_{g,..}$, while the coefficient on the FE for period t is equal to the average treatment at period t across groups, denoted $D_{.,t}$. Then,

$$\hat{u}_{g,t} = D_{g,t} - D_{g,..} - D_{.,t} + D_{..}, \quad (5.6)$$

where $D_{..}$ is the average treatment across groups and periods, and this term ensures that the average of the residuals $\hat{u}_{g,t}$ is equal to zero. Then,

$$D_{g,t}\hat{u}_{g,t} = D_{g,t}(D_{g,t} - D_{g,..} - D_{.,t} + D_{..}).$$

In Design CLA, $D_{g,t}$, $D_{g,..}$, and $D_{.,t}$ are constant across treated cells, so $D_{g,t}\hat{u}_{g,t}$ is also constant across treated cells. $D_{g,t}\hat{u}_{g,t}$ is also constant across treated cells in designs with a non-absorbing binary treatment, without variation in treatment timing, where $D_{g,t} = 1\{t \in \mathcal{T}_1\}D_g$, and \mathcal{T}_1

denotes the set of (non-necessarily consecutive) periods where the treated groups are treated. On the other hand, in designs with variation in treatment timing and/or treatment dose, $D_{g,t}\hat{u}_{g,t}$ generally varies across treated (g, t) cells, because $D_{g,t}$, $D_{g,..}$, and/or $D_{.,t}$ vary.

5.2.1 Additional assumptions under which $\hat{\beta}^{fe}$ is unbiased

Find an assumption which, added to those in Theorem 8, ensures that $\hat{\beta}^{fe}$ is unbiased for the ATT.

$\hat{\beta}^{fe}$ is unbiased for the ATT if the treatment-effect is constant. Assume that the treatment effect is constant, across groups and over time:

$$\text{There exists a real number } \delta \text{ such that } \text{TE}_{g,t} = \delta \text{ for all } (g, t). \quad (5.7)$$

If (5.7) holds, what is the value of $E[\hat{\beta}^{fe}]$?

Plugging (5.7) into Theorem 8, and using the fact that $\sum_{(g,t):D_{g,t}=1} W_{g,t} = 1$, one has that $E[\hat{\beta}^{fe}] = \delta$: $\hat{\beta}^{fe}$ is unbiased for δ , which is also equal to the ATT if the treatment effect is constant. However, (5.7) is often an implausible assumption. For instance, could it be that the effect of newspapers varies across counties?

Yes, it seems difficult to rule out ex-ante that possibility. For instance, the effect of newspapers could be stronger in rural counties, where newspapers may play a larger role in the transmission of political ideas, than in more urban counties where political ideas may have other ways to circulate even in the absence of newspapers.

$\hat{\beta}^{\text{fe}}$ is unbiased for the ATT if the weights $W_{g,t}$ are uncorrelated with the treatment effects $\text{TE}_{g,t}$. Instead of assuming constant treatment effects, let us consider the following assumption:

$$\sum_{(g,t):D_{g,t} \neq 0} \left(W_{g,t} - \frac{1}{N_1} \right) (\text{TE}_{g,t} - \text{ATT}) = 0. \quad (5.8)$$

As the average value of $W_{g,t}$ across treated cells is equal to $\frac{1}{N_1}$, the condition in the previous display requires that $W_{g,t}$ is uncorrelated with $\text{TE}_{g,t}$. If the treatment effect is constant then (5.8) automatically holds, so (5.8) is weaker than (5.7). (5.8) implies that

$$\sum_{(g,t):D_{g,t} \neq 0} W_{g,t} \text{TE}_{g,t} = \sum_{(g,t):D_{g,t} \neq 0} W_{g,t} \text{ATT} + \frac{1}{N_1} \sum_{(g,t):D_{g,t} \neq 0} \text{TE}_{g,t} - \frac{1}{N_1} \sum_{(g,t):D_{g,t} \neq 0} \text{ATT} = \text{ATT},$$

where the second equality follows from $\sum_{(g,t):D_{g,t} \neq 0} W_{g,t} = 1$. Combined with Theorem 8, the previous display implies that

$$E [\hat{\beta}^{\text{fe}}] = \text{ATT}. \quad (5.9)$$

Intuitively, (5.8) ensures that the treatment effects that are up- and down-weighted by $\hat{\beta}^{\text{fe}}$ do not systematically differ, so $\hat{\beta}^{\text{fe}}$ is unbiased for the ATT under (5.8).

To simplify, let us momentarily assume that treatment is binary. Then the numerator of $W_{g,t}$, the part of $W_{g,t}$ that varies across (g, t) cells, is equal to $1 - D_{g,.} - D_{.,t} + D_{.,.}$. Under which economic model of selection into treatment could we have that this quantity is correlated with the treatment effects $\text{TE}_{g,t}$, and thus (5.8) fails?

(5.8) is likely to fail if selection into treatment follows a Roy model. If selection into treatment follows a Roy selection model where (g, t) cells decide to get treated when their benefit from treatment is larger than the cost, (5.8) is likely to fail. To see this, note that $1 - D_{g,.} - D_{.,t} + D_{.,.}$ is decreasing in $D_{g,.}$, meaning that $\hat{\beta}^{\text{fe}}$ downweights the treatment effect of groups with the highest average treatment from period 1 to T . However, in a Roy selection model, groups with the largest average treatment may be those with the largest treatment effect, which could lead to a correlation between $W_{g,t}$ and $\text{TE}_{g,t}$.

Tests of (5.8). While (5.8) may not be very plausible when selection into treatment is based on a Roy model, there may be other instances where (5.8) is more plausible. As we will see in Section 6.4 of the next chapter, (5.8) is sometimes testable. For now, we just note that (5.8) can be suggestively tested, if one observes a proxy variable $P_{g,t}$ likely to be correlated with $\text{TE}_{g,t}$. Then, one can test if $W_{g,t}$ and $P_{g,t}$ are correlated.

5.3 $\hat{\beta}^{\text{fe}}$ may not estimate a convex combination of treatment effects

Are the weights $W_{g,t}$ positive for all (g, t) ?

$\hat{\beta}^{\text{fe}}$ may not estimate a convex combination of treatment effects. (5.6) implies that some of the weights $W_{g,t}$ may be negative, if there are treated (g, t) cells such that

$$D_{g,t} + D_{.,.} < D_{g,.} + D_{.,t}. \quad (5.10)$$

In the newspapers example, with negative weights, $\hat{\beta}^{\text{fe}}$ could be estimating something like 3 times the effect of newspapers on turnout in Santa Clara county, minus 2 times the effect in Wayne county. Then, if adding one more newspaper raises turnout by 1 percentage points in Santa Clara county and by 2 percentage points in Wayne county, one would have $E[\hat{\beta}^{\text{fe}}] = 3 \times 0.01 - (2 \times 0.02) = -0.01$. $E[\hat{\beta}^{\text{fe}}]$ would be negative, while the effect of newspapers is positive in both counties. This example shows that $\hat{\beta}^{\text{fe}}$ may not satisfy the “no-sign reversal property” (Imbens and Angrist, 1994; Small, Tan, Ramsahai, Lorch and Brookhart, 2017): $E[\hat{\beta}^{\text{fe}}]$ could for instance be negative, even if the treatment effect is strictly positive in every (g, t) . This phenomenon can only arise when some of the weights $W_{g,t}$ are negative: when all those weights are positive, $\hat{\beta}^{\text{fe}}$ does satisfy the no-sign reversal property.

No-sign-reversal and Pareto dominance.* Despite its intuitive appeal and its popularity among applied researchers, the no-sign reversal property is not grounded in statistical decision theory, unlike other commonly-used criteria to discriminate estimators such as the mean-squared error. Still, it is connected to the economic concept of Pareto dominance (de Chaisemartin and D'Haultfœuille, 2023b). If an estimator does not satisfy “no-sign-reversal”, its expectation or its probability limit could for instance be positive, even if the treatment is Pareto-dominated by the absence of treatment, meaning that everybody is hurt by the treatment.

A taxonomy of cells whose treatment effect is likely to be weighted negatively. In view of (5.10), treated (g, t) cells that receive a low treatment dose $D_{g,t}$ are more likely to be such that their treatment effect is weighted negatively. Similarly, groups whose average treatment $D_{g, \cdot}$ is the highest are the most likely to be such that at some periods, their treatment effects are weighted negatively. Finally, time periods when the average treatment $D_{\cdot, t}$ is the highest are the most likely to be such that the treatment effects of some groups are weighted negatively at those periods.

$\hat{\beta}^{\text{fe}}$ can be used to test the sharp null of no treatment effect. Finally, while $\hat{\beta}^{\text{fe}}$ may not provide an easily interpretable measure of the treatment's effect, it can be used to test the so-called sharp null of no treatment effect ($Y_{g,t}(d) - Y_{g,t}(0) = 0$ for all (g, t, d)): under that null, it follows from Theorem 8 that $E[\hat{\beta}^{\text{fe}}] = 0$.

Bibliographic notes. Theorem 8 is a restatement of Theorem 1 in de Chaisemartin and D'Haultfœuille (2020), and (5.9) is a restatement of Corollary 2 therein.² Several teams of researchers have been involved in the realization that, outside of the classical design, TWFE estimators are no longer simple DID estimators and may not be unbiased for the ATT or even

²Those two results assume that treatment is binary, but de Chaisemartin and D'Haultfœuille (2020) extend their Theorem 1 to a non-binary treatment, see Theorem S3 in their web appendix.

just a convex combination of effects under the parallel-trends assumption:

1. The equation on p.590 of Blundell and Costa-Dias (2009) is the first decomposition of a DID estimator as a potentially non-convex weighted sum of treatment effects under the parallel-trends assumption.³ They consider designs with two groups and two periods, where exposure to treatment increases more in one group than in the other, and allow for heterogeneous effects across groups, but they assume constant effects over time. de Chaisemartin (2011) independently obtains a similar decomposition without assuming constant effects over time. His decomposition shows that time-varying effects also lead to negative weights. See also Fricke (2017) for a related result, in designs where the two groups receive different treatments or different treatment doses at period two.⁴
2. Theorems S1 and S2 of the Supplementary Material of de Chaisemartin and D'Haultfœuille (2015) are the first decompositions of TWFE regression coefficients as potentially non-convex weighted sums of treatment effects under the parallel-trends assumption. See also Appendix C of Borusyak and Jaravel (2017), that contains a result related to that in (5.1), in designs with an absorbing binary treatment, with variation in treatment timing.⁵

5.4 Decompositions of related estimators

Decomposition of the first-difference estimator. de Chaisemartin and D'Haultfœuille (2020) also consider $\widehat{\beta}^{\text{fd}}$, the treatment's coefficient in a regression of $Y_{g,t} - Y_{g,t-1}$, the outcome's first difference, on $D_{g,t} - D_{g,t-1}$, the treatment's first difference and period FE. With $T = 2$, $\widehat{\beta}^{\text{fe}}$ and $\widehat{\beta}^{\text{fd}}$ are numerically equivalent, but with $T \geq 3$ the two estimators generally differ. They show that $\widehat{\beta}^{\text{fd}}$ can also be decomposed as a weighted sum of $\text{TE}_{g,t}$ under Assumptions NA, ND, and PT, with weights $W_{g,t}^{\text{fd}}$ that differ from those in (5.1) when $T \geq 3$, but that also sum to one

³There is a typo in that equation: $+(p_{01} - p_{00})$ should actually be $-(p_{01} - p_{00})$, leading to negative weights.

⁴Theorem 1 in Chernozhukov et al. (2013) shows that under the assumption that $E(Y_{g,t}(0))$ does not depend on t , one-way FE regressions, with group FE but no period FE, may be biased for the average treatment effect, but unlike TWFE regressions they always estimate a convex combination of effects.

⁵Imai and Kim (2021) also derive another decomposition of TWFE regressions, under a sequential-exogeneity assumption.

and that may also be negative:

$$E \left[\widehat{\beta}^{\text{fd}} \right] = \sum_{(g,t): D_{g,t} \neq 0} W_{g,t}^{\text{fd}} \text{TE}_{g,t}. \quad (5.11)$$

This implies that under constant treatment effects, the expectations of $\widehat{\beta}^{\text{fe}}$ and $\widehat{\beta}^{\text{fd}}$ are equal. Accordingly, if the two coefficients significantly differ, under Assumptions NA, ND, and PT one can reject the null that the treatment effect is constant.

Decomposition of TWFE and first-difference estimators with controls. de Chaisemartin and D'Haultfoeuille (2020) also derive a decomposition similar to (5.1) for TWFE regressions with control variables, see Theorem S4 in their Web Appendix, which we already mentioned in Chapter 4.

5.5 Stata and R commands to compute the weights

The `twowayfeweights` Stata (see de Chaisemartin, D'Haultfoeuille and Deeb, 2019) and R (see Zhang and de Chaisemartin, 2021) commands compute the weights attached to TWFE and FE regressions.

To compute the weights attached to a TWFE regression, the syntax of the Stata command is:

```
twowayfeweights outcome groupid timeid treatment, type(feTR)
```

To compute the weights attached to an FD regression, the syntax is:

```
twowayfeweights fdoutcome groupid timeid fdtreatment treatment, type(fdTR)
```

To compute the weights attached to TWFE or FD regressions with control variables, users can input those variables into the `controls` option. To suggestively test (5.8), users can specify the `test_random_weights` option, inputting variables likely to be correlated with the treatment effects $\text{TE}_{g,t}$ into the option. Then, the command will compute the correlation between the weights $W_{g,t}$ and those variables, and test if those correlations are significant.

5.6 Application

Gentzkow et al. (2011) estimate a regression similar to the TWFE one in (3.1), up to two differences. First, they include state-year FEs as additional control variables. Second, they estimate the regression in first difference. In our replication, we start by estimating the basic TWFE regression in (3.1), we then estimate a TWFE regression with state-year FEs, and we finally replicate the authors' specification.

5.6.1 Basic TWFE regression

Using `gentzkowetal_didtextbook`, regress turnout on number of newspapers and county and year FEs, clustering standard errors at the county level. Interpret the results.

```
areg prestout i.year numdailies, absorb(cnty90) cluster(cnty90)
```

$\hat{\beta}^{fe} = 0.0029$: according to this regression, one more newspaper would increase turnout by 0.29 percentage points. The coefficient is marginally significant (s.e. = 0.0016). Use the `twowayfeweights` Stata package to decompose $\hat{\beta}^{fe}$, and interpret the results.

```
twowayfeweights prestout cnty90 year numdailies, type(feTR)
```

Under the parallel-trends assumption, $\hat{\beta}^{fe}$ estimates a weighted sum of the effects of newspapers on turnout in 10,378 county×election-year cells, where 6,180 effects are weighted positively while 4,198 are weighted negatively, and where negative weights sum to -0.47. Accordingly, $\hat{\beta}^{fe}$ is far from estimating a convex combination of effects.

5.6.2 TWFE regression with state-year FEs

Regress turnout on number of newspapers and county, year, and state-year FEs, clustering standard errors at the county level. Interpret the results.

```
qui areg prestout i.year i.styr numdailies, absorb(cnty90) cluster(cnty90)
di _b[numdailies], _se[numdailies]
```

According to this regression, one more newspaper would reduce turnout by 0.12 percentage points. The coefficient is insignificant (s.e. = 0.0011). Use the `twowayfeweights` Stata package to decompose the newspapers' coefficient in this regression, and interpret the results. As the regression controls for state-year FEs, you need to create those FEs and input them into the command's `controls` option.

```
qui tab styr, gen(styr)
twowayfeweights prestout cnty90 year numdailies, type(feTR) controls(styr1-styr683)
```

The newspapers' coefficient estimates a weighted sum of the effects of newspapers on turnout in 10,342 county×election-year cells, where 6,195 effects are weighted positively while 4,147 are weighted negatively, and where negative weights sum to -0.53.

5.6.3 FD regression with state-year FEs

Regress the change in turnout on the change in the number of newspapers and state-year FEs, clustering standard errors at the county level. Interpret the results.

```
areg change prestout changedailies, absorb(styr) cluster(cnty90)
```

According to this regression, one more newspaper would increase turnout by 0.26 percentage points. The coefficient is significant at all conventional levels (s.e. = 0.0009). Use the `twowayfeweights` Stata package to decompose the newspapers' coefficient in this regression, and interpret the results. You need to use the command's syntax for an FD rather than a TWFE regression.

```
twowayfeweights change prestout cnty90 year changedailies numdailies,  
type(fdTR) controls(styr1-styr683)
```

The newspapers' coefficient estimates a weighted sum of the effects of newspapers on turnout in 9,876 county×election-year cells, where 5,371 effects are weighted positively while 4,505 are weighted negatively, and where negative weights sum to -1.43. Assess if the weights are correlated with the year variable. Interpret the results.

```
twowayfeweights change prestout cnty90 year changedailies numdailies,  
type(fdTR) controls(styr1-styr683) test_random_weights(year)
```

The weights are negatively correlated with the election year, and the correlation is significant: the FD regression is more likely to upweight newspapers' effects in early elections, and to downweight or weight negatively newspapers' effects in late elections. If the effect of newspapers diminishes over time, for instance because new means of communication, like the radio, appear in the end of the period, the negative correlation between the weights and the year variable could lead the FD regression to overestimate newspapers' average effect. If the effect of newspapers increases over time, for instance because readership per newspaper increases as the literacy rate increases over the period, this negative correlation could lead the FD regression to underestimate newspapers'

average effect.

5.7 Next steps

In the next two chapters, we will focus on two seemingly small departures from the classical design: designs with an absorbing, binary treatment and variation in treatment timing, and designs with two periods and variability in the treatment dose received by treated groups at period two. In each case, we will see that TWFE estimators may not estimate the ATT or a convex combination of effects because they leverage DIDs with a control group that is actually treated. Then, we will review alternative estimators, which avoid leveraging DIDs with a treated control group, and are unbiased for averages of (g, t) -specific effects. Finally, in the book's last chapter, we will combine the insights from the two preceding chapters to propose estimators robust to heterogeneous effects in general designs, with non-absorbing and/or non-binary treatments. Importantly, while we have ruled out dynamic effects in this chapter, we will allow for effects of the lagged treatments on the outcome in the following chapters.

Chapter 6

Designs with variation in treatment timing

Binary and staggered designs. Throughout this chapter, we assume that treatment is absorbing and binary, as in Chapters 3 and 4, but we assume that there is variation in treatment timing: treated groups start receiving the treatment at different dates. As a shortcut, we refer to such treatments as binary and staggered.

Design BST (*Binary and staggered design*) $D_{g,t} = 1\{t \geq F_g\}$, with $\min_{g:F_g>1} F_g < \max_g F_g$.

F_g is the first date at which group g becomes treated, and group g remains treated thereafter. If g never becomes treated over the study period, we let $F_g = T + 1$. F_g may be equal to 1, meaning that group g is always treated. $\min_{g:F_g>1} F_g < \max_g F_g$ requires that among groups that are untreated at period 1, not all groups get treated at the same period.

Chapter's running example: the effect of unilateral divorce laws on divorce rates. Between 1968 and 1988, 29 US states adopted a unilateral divorce law (UDL), allowing one spouse to terminate the marriage without the consent of the other. Wolfers (2006), building upon Friedberg (1998), uses a yearly panel of US states (plus the District of Columbia, hereafter incorrectly referred to as a state) from 1956 to 1988, to estimate the effects of those laws on divorce rates. The UDL treatment satisfies Design BST: the treatment is binary, states adopt UDLs at different dates, and they never repeal those laws. Then, F_g denotes the year when state g adopts a UDL.

Dataset used in this chapter. To answer the green questions in this chapter, you need to use the `wolfers_didtextbook` dataset, which contains the following variables:

- `state`: a state identifier;
- `year`: a year identifier;
- `cohort`: the year when state g adopted treatment (F_g in our notation, except that it is equal to zero rather than $T + 1$ for never-treated states);
- `early_late_never`: a variable equal to 1 for states with an adoption year below the median, to 2 for states with an adoption year above the median, and to 3 for never adopters;
- `udl`: a variable equal to 1 if state g has a UDL at year t and to 0 otherwise;
- `exposurelength`: a variable equal to the number of years for which state g has had a UDL at year t for treated (g, t) cells, and to 0 for untreated (g, t) cells;
- `rel_time1` to `rel_time15`: indicators equal to one if $t = F_g - 1 + \ell$, namely if group g has been treated for ℓ years at year t , for $\ell \in \{1, \dots, 15\}$;
- `rel_time16`: an indicator equal to one if $t \geq F_g - 1 + 16$, namely if group g has been treated for at least 16 years at year t ;
- `rel_timeminus1` to `rel_timeminus8`: indicators equal to one if $t = F_g - 1 - \ell$ for $\ell \in \{1, \dots, 8\}$, namely if group g will be treated $\ell + 1$ years after year t ;
- `rel_timeminus9`: an indicator equal to one if $t \leq F_g - 1 - 9$, namely if group g will be treated at least 9 + 1 years after year t ;
- `div_rate`: the number of divorces per 1,000 people in state g and year t ;
- `stpop`: the population of state g in year t ;
- `stpop1968`: the population of state g in year 1968, the last year before states start adopting UDLs;
- `controlgroup`: a variable equal to one for never-treated groups.

6.1 Target parameters

(g, t) -specific effects. Remember that for any integer $k \geq 1$, $\mathbf{1}_k$ denotes a vector of k ones.

For all g such that $F_g \leq T$, and $t \in \{F_g, \dots, T\}$, let

$$\text{TE}_{g,t} = E[Y_{g,t} - Y_{g,t}(\mathbf{0}_t)] = E[Y_{g,t}(\mathbf{0}_{F_g-1}, \mathbf{1}_{t-(F_g-1)}) - Y_{g,t}(\mathbf{0}_t)].$$

Interpret $\text{TE}_{g,t}$.

$\text{TE}_{g,t}$ is the expected effect, in group g and at period t , of having been treated rather than untreated from period F_g to t , namely for $t - (F_g - 1)$ periods. In the UDL example, $\text{TE}_{g,t}$ is the effect, in state g and year t , of having been exposed to a UDL for $t - (F_g - 1)$ years.

Average treatment effect on the treated. Letting N_1 denote the number of treated (g, t) cells, a natural aggregated target parameter is

$$\text{ATT} = \frac{1}{N_1} \sum_{(g,t): D_{g,t}=1} \text{TE}_{g,t},$$

the average effect of having been treated rather than untreated for $t - (F_g - 1)$ periods, across all treated (g, t) cells. This parameter generalizes the ATT parameter introduced in Chapter 3 to the binary-and-staggered designs we consider in this chapter.

(g, ℓ) -specific effects. For all g such that $F_g \leq T$, and for $\ell \in \{1, \dots, T - (F_g - 1)\}$, let

$$\text{TE}_{g,\ell}^r = E[Y_{g,F_g-1+\ell}(\mathbf{0}_{F_g-1}, \mathbf{1}_\ell) - Y_{g,F_g-1+\ell}(\mathbf{0}_{F_g-1+\ell})].$$

Interpret $\text{TE}_{g,\ell}^r$.

$\text{TE}_{g,\ell}^r$ is the expected effect, in group g and at period $F_g - 1 + \ell$, of having been treated rather

than untreated from period F_g to $F_g - 1 + \ell$, namely for ℓ periods. $\text{TE}_{g,\ell}^r = \text{TE}_{g,F_g-1+\ell}$, so $\text{TE}_{g,\ell}^r$ is just a convenient notation to index treatment effects with respect to length of exposure to treatment rather than calendar time.

Average effect of having been treated for ℓ periods. Let $\underline{T} = \max_g F_g - 1$ denote the last period when there is still at least one untreated group. If there are never-treated groups, $\underline{T} = T$. Let $\underline{F} = \min_{g:F_g>1} F_g$ denote the earliest period when a group adopts the treatment. For instance, if all groups are untreated at periods 1 and 2 and one group becomes treated at period 3, $\underline{F} = 3$. For any $\ell \in \{1, \dots, \underline{T} - (\underline{F} - 1)\}$, let G_ℓ denote the number of groups such that $F_g \geq 2$ and $F_g - 1 + \ell \leq \underline{T}$, meaning that those groups are initially untreated and reach their ℓ th period of exposure to treatment at a time period where there is still at least one untreated group. Then, let

$$\text{ATT}_\ell = \frac{1}{G_\ell} \sum_{g: F_g \geq 2, F_g - 1 + \ell \leq \underline{T}} \text{TE}_{g,\ell}^r.$$

ATT_ℓ is the average effect of having been exposed to treatment for ℓ periods, across all initially untreated groups that reach ℓ treatment periods before all groups are treated. We restrict attention to those groups, because we will not be able to propose unbiased DID estimators of $\text{TE}_{g,\ell}^r$ for initially treated groups and for groups reaching ℓ treatment periods after \underline{T} . ATT_ℓ generalizes the parameter ATT_ℓ defined in Chapter 3, to binary-and-staggered designs with variation in treatment timing. If no group is treated at period one and there are never-treated groups, $\underline{T} = T$ and one can show that the ATT is a weighted average of the ATT_ℓ s:

$$\text{ATT} = \sum_\ell \frac{G_\ell}{N_1} \text{ATT}_\ell. \quad (6.1)$$

Caveats in interpreting $\ell \mapsto \text{ATT}_\ell$. In Chapter 3, we have seen that without assuming that the treatment's effect does not change with calendar time, $\ell \mapsto \text{ATT}_\ell$ cannot be used to determine if the treatment effect increases with length of exposure. This remains true here, though the variation in treatment timing implies that now, length of exposure to treatment is not perfectly collinear with calendar time, thus implying that under some assumptions, it may be possible to disentangle how treatment effects vary with length of exposure and with calendar time, an interesting avenue for future research. But in binary-and-staggered designs, a new difficulty arises when it comes to interpreting $\ell \mapsto \text{ATT}_\ell$, which was not present in Chapter 3.

Even if the treatment effect does not vary with length of exposure ℓ and with calendar time t , could we have that $\text{ATT}_\ell \neq \text{ATT}_{\ell'}$ in a binary-and-staggered design?

Yes, because for $\ell \neq \ell'$, ATT_ℓ and $\text{ATT}_{\ell'}$ do not apply to the same groups, as fewer and fewer groups reach ℓ treatment periods before \underline{T} , as ℓ increases. Thus, variations in ATT_ℓ across ℓ can come from treatment effects varying with length of exposure or calendar time, but also from compositional changes if treatment effects vary across groups. Some of the Stata and R packages discussed in this chapter have options to estimate event-study effects that all apply to the same groups, thus avoiding such compositional changes.

6.2 Two-Way Fixed Effects estimators

6.2.1 Static Two-Way Fixed Effects estimator

6.2.1.1 Decomposition of $\hat{\beta}^{\text{fe}}$

Theorem 8 in binary-and-staggered designs. In Design BST, the weights in the decomposition of $\hat{\beta}^{\text{fe}}$ in Theorem 8 simplify to

$$W_{g,t} = \frac{\hat{u}_{g,t}}{\sum_{(g',t'):D_{g',t'}=1} \hat{u}_{g',t'}},$$

and one has that $D_{g,.} = \frac{T-(F_g-1)}{T}$, so

$$\hat{u}_{g,t} = D_{g,t} - \frac{T - (F_g - 1)}{T} - D_{.,t} + D_{.,.},$$

for treated cells. In view of the previous display, which groups are the most likely to be such that some of the weights $W_{g,t}$ are negative for some t ?

Groups that become treated early. In particular, always-treated groups are such that $\hat{u}_{g,t} = D_{.,.} - D_{.,t}$. As $D_{.,t}$ is weakly increasing in t in Design BST, $D_{.,T} > D_{.,.}$, so if there are always-treated groups, their treatment effect at the last period is always weighted negatively by $\hat{\beta}^{\text{fe}}$. Then, to mitigate or eliminate the negative weights, one could drop the always-treated groups from the estimation sample. As those groups are never observed without treatment, their treatment effect cannot be estimated under no-anticipation and parallel-trends assumptions. Dropping them from the estimation sample is thus necessary if one would prefer not imposing other assumptions. Which time periods are the most likely to be such that some of the weights $W_{g,t}$ are negative for some g ?

The last time periods of the panel, because $D_{.,t}$ is weakly increasing in t in Design BST. Overall, the long-run treatment effects of early-treated groups are the most likely to be weighted negatively, something that was first noted by Borusyak and Jaravel (2017). When are all the weights $W_{g,t}$ likely to be positive?

All the weights are positive if and only if $\frac{T-(F_g-1)}{T} + D_{.,t} \leq 1 + D_{.,.}$ for all (g, t) . Accordingly, all the weights are likely to be positive when there is no group that is treated most of the time, and no time period where most groups are treated. For instance, if a large proportion of groups are never treated, it is likely that $\hat{\beta}^{\text{fe}}$ estimates a convex combination of effects, thus implying that $\hat{\beta}^{\text{fe}}$ satisfies the no-sign reversal property. Even then, $\hat{\beta}^{\text{fe}}$ may still be biased for the ATT.

Application to the UDL example. Using the `wolfers_didtextbook` dataset, run the static TWFE regression of the divorce rate on state and year FEs and the UDL treatment, weighting the regression by the state's population and clustering standard errors at the state level. According to this regression, do UDLs have an effect on divorces?

```
reg div_rate udl i.state i.year [w=stpop], vce(cluster state)
```

The coefficient on the UDL treatment is small and insignificant, so according to this regression UDLs do not have an effect on divorces. Execute the following Stata command:

```
twowayfeweights div_rate state year udl, type(feTR) test_random_weights(exposurelength)
weight(stpop),
```

where `test_random_weights(exposurelength)` is used to test whether the weights attached to $\hat{\beta}^{\text{fe}}$ are correlated with the number of years for which a state has been exposed to a UDL, while `weight(stpop)` indicates that the TWFE regression we seek to decompose is weighted by `stpop`. Interpret the results: does $\hat{\beta}^{\text{fe}}$ estimate a convex, or almost convex combination of effects? Could it be the case that $\hat{\beta}^{\text{fe}}$ is biased for ATT?

Under Assumptions NA and PT, $\hat{\beta}^{\text{fe}}$ estimates a weighted sum of 522 TE_{*g,t*}. 490 TE_{*g,t*} receive a positive weight, and 32 receive a negative weight. Negative weights are small and sum to -0.026: $\hat{\beta}^{\text{fe}}$ estimates an “almost convex” combination of effects. Yet, weights are strongly and negatively correlated with length of exposure, meaning that $\hat{\beta}^{\text{fe}}$ downweights effects of longer lengths of exposure. Then, $\hat{\beta}^{\text{fe}}$ could differ from the ATT if treatment effects vary with length of exposure to a UDL. For instance, $\hat{\beta}^{\text{fe}}$ would overestimate the ATT if treatment effects decrease with length of exposure.

Redefining the treatment to make the problem go away? With never-treated groups and no always-treated group, one could redefine the treatment as $\tilde{D}_{g,t} = 1\{t \geq \min_{g'} F_{g'}\}1\{F_g \leq T\}$: cell (*g,t*) is considered as treated if group *g* is eventually treated and *t* is after the first period when a group becomes treated. $\tilde{D}_{g,t}$ is a binary and absorbing treatment, without variation in treatment timing, so it follows from results in the previous chapter that under Assumptions NA and PT, the coefficient $\tilde{\beta}^{\text{fe}}$ in a TWFE regression of $Y_{g,t}$ on group and period FEs and $\tilde{D}_{g,t}$ yields an unbiased estimator of $\tilde{\text{ATT}}$, the ATT of $\tilde{D}_{g,t}$ on the outcome. The issue with this strategy is that there are (*g,t*) cells with $\tilde{D}_{g,t} = 1$ that are actually untreated, such that the effect of $\tilde{D}_{g,t}$ on the outcome is actually equal to zero for those cells. Then, letting \tilde{N}_1 denote the number

of cells such that $\tilde{D}_{g,t} = 1$, one can show that $\tilde{\text{ATT}} = (N_1/\tilde{N}_1)\text{ATT}$: $\tilde{\text{ATT}}$ and ATT are of the same sign, but $\tilde{\text{ATT}}$ is biased towards zero. Bias-correcting $\tilde{\beta}^{\text{fe}}$ is easy: $\tilde{N}_1/N_1\tilde{\beta}^{\text{fe}}$ is unbiased for the ATT. However, $V(\tilde{N}_1/N_1\tilde{\beta}^{\text{fe}}) = (\tilde{N}_1/N_1)^2 V(\tilde{\beta}^{\text{fe}})$: the variance of the bias-corrected estimator is $(\tilde{N}_1/N_1)^2$ larger than that of $\tilde{\beta}^{\text{fe}}$, a TWFE estimator. Then, this redefinition of the treatment may lead to a low statistical precision, especially when \tilde{N}_1/N_1 is large, as is the case when treatment timing varies substantially across groups, meaning that some groups get treated long before some other groups. Another issue with this strategy is that while it can be used to estimate the ATT, it cannot be used to estimate the event-study effects ATT_ℓ , while these target parameters are often of interest.

6.2.1.2 The origin of the negative weights

Decomposing $\hat{\beta}^{\text{fe}}$ as a weighted average of 2×2 DIDs. Goodman-Bacon (2021) shows that in Design BST,

$$\hat{\beta}^{\text{fe}} = \sum_{g \neq g', t < t'} v_{g,g',t,t'} \text{DID}_{g,g',t,t'}, \quad (6.2)$$

where $\text{DID}_{g,g',t,t'}$ is a DID comparing the outcome evolution of two groups g and g' from a pre period t to a post period t' , and where $v_{g,g',t,t'}$ are non-negative weights summing to one, with $v_{g,g',t,t'} > 0$ if and only if g 's treatment changes between t and t' while g' 's treatment does not change.¹ Given that g 's treatment changes between t and t' , what is the only possible value of $(D_{g,t}, D_{g,t'})$? Given that g' 's treatment does not change between t and t' , what are the two possible values of $(D_{g',t}, D_{g',t'})$?

$(D_{g,t}, D_{g,t'}) = (0, 1)$. $(D_{g',t}, D_{g',t'}) = (0, 0)$, or $(D_{g',t}, D_{g',t'}) = (1, 1)$. Thus, some of the $\text{DID}_{g,g',t,t'}$ in (6.2) compare a group switching from untreated to treated between t and t' to a group untreated at both dates, while other $\text{DID}_{g,g',t,t'}$ compare a switching group to a group treated at

¹Goodman-Bacon (2021) actually decomposes $\hat{\beta}^{\text{fe}}$ as a weighted average of DIDs between cohorts of groups becoming treated at the same date, and between periods of time where their treatment remains constant. One can then further decompose his decomposition, as we do here.

both dates. The negative weights in (5.1) originate from this second type of DIDs.

Forbidden comparisons in binary-and-staggered designs: a tale of two patients who had an infection.

Let us consider a simple example, whose design is inspired from an example in Borusyak and Jaravel (2017), who have also coined the “forbidden comparisons” expression we borrow here. Patients e and ℓ are feeling sick, and consult their doctor at $t = 1$. The doctor suspects an infection, and asks for some lab tests. At $t = 2$, the doctor receives the lab results for e : the test confirms that e has an infection, so the doctor prescribes them antibiotics, which they start taking right away. At $t = 3$, the doctor receives the lab results for ℓ : the test again confirms that ℓ has an infection, so the doctor prescribes them antibiotics, which they start taking right away. This gives us a design with two patients and three periods, such that e , the early-treated patient, is untreated at period 1 and treated at periods 2 and 3, while ℓ , the late-treated patient, is untreated at periods 1 and 2 and treated at period 3. A smart econometrician comes by, and immediately sees some research potential in this natural experiment. Accustomed to running TWFE regressions, they regress $Y_{g,t}$, the fever of patient g at t , on patient FEs, period FEs, and the treatment status of patient g at t . In this simple design, (6.2) reduces to

$$\hat{\beta}^{\text{fe}} = (\text{DID}_{e,\ell,1,2} + \text{DID}_{\ell,e,2,3})/2, \quad (6.3)$$

with

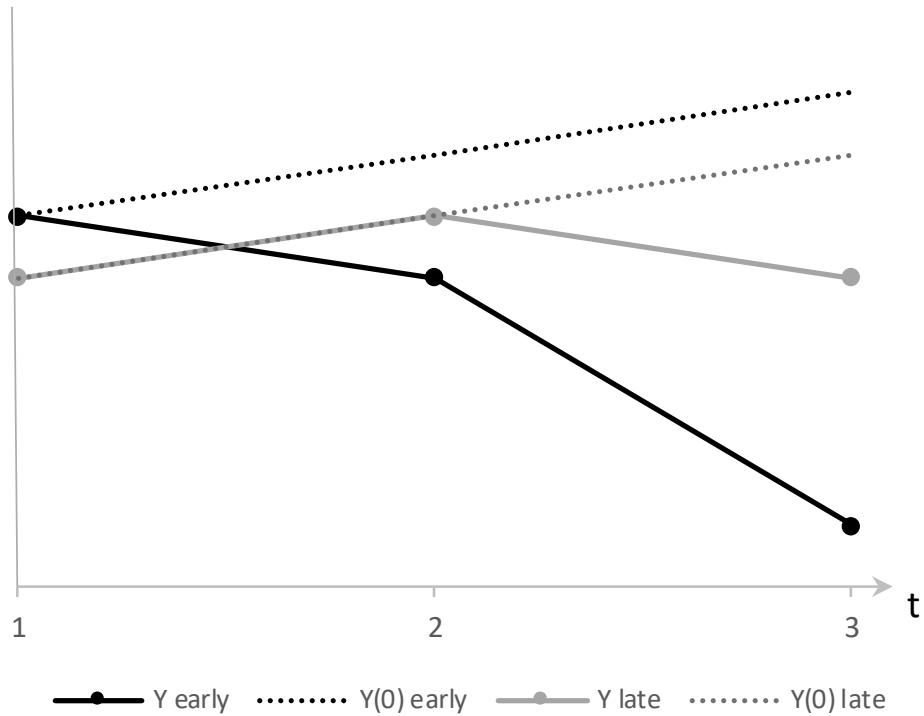
$$\text{DID}_{e,\ell,1,2} = Y_{e,2} - Y_{e,1} - (Y_{\ell,2} - Y_{\ell,1}),$$

$$\text{DID}_{\ell,e,2,3} = Y_{\ell,3} - Y_{\ell,2} - (Y_{e,3} - Y_{e,2}).$$

In words, the econometrician’s TWFE coefficient is just the simple average of $\text{DID}_{e,\ell,1,2}$, which compares the period-1-to-2 fever evolution of patients e and ℓ , and of $\text{DID}_{\ell,e,2,3}$, which compares the period-2-to-3 fever evolution of patients ℓ and e . To his surprise, the econometrician finds that $\hat{\beta}^{\text{fe}} > 0$. Is it correct to conclude that antibiotics increase fever, or could something else explain why $\hat{\beta}^{\text{fe}} > 0$?

Antibiotics are slow-acting drugs, which take a few days to reduce symptoms. Due to that, the fever of the early-treated patient may drop slightly from period one to two, and sharply from period two to three, as at period three that patient has been receiving antibiotics for two periods. Similarly, the fever of the late-treated patient may drop slightly from period two to three. Thus, one may have that $DID_{e,\ell,1,2}$ takes a small negative value, while $DID_{\ell,e,2,3}$ takes a large positive value, and eventually $\hat{\beta}^{fe}$ is positive. Figure 6.1 below shows patients' actual (solid lines) and counterfactual (dashed lines) fever evolutions, in a numerical example where both patients always benefit from the antibiotics treatment, but $\hat{\beta}^{fe}$ is positive.

Figure 6.1: A tale of two patients who had an infection.



Forbidden comparisons in binary-and-staggered designs: formal analysis. More formally, $DID_{e,\ell,1,2}$ is a simple DID comparing a group going from untreated to treated to a group that is always untreated, like the DID estimator in (1.3). Then, under no-anticipation and parallel-trends assumptions it is unbiased for the treatment effect in group e at period 2:

$$E [DID_{e,\ell,1,2}] = TE_{e,2}. \quad (6.4)$$

$\text{DID}_{\ell,e,2,3}$, on the other hand, compares the period-2-to-3 outcome evolution of group ℓ , that switches from untreated to treated from period 2 to 3, to the outcome evolution of group e that is treated at both dates. Accordingly,

$$E [Y_{e,3} - Y_{e,2}] = E [Y_{e,3}(0, \mathbf{1}_2) - Y_{e,2}(0, 1)] = E [Y_{e,3}(\mathbf{0}_3) - Y_{e,2}(\mathbf{0}_2)] + \text{TE}_{e,3} - \text{TE}_{e,2}. \quad (6.5)$$

On the other hand, group ℓ is only treated at period 3, so

$$E [Y_{\ell,3} - Y_{\ell,2}] = E [Y_{\ell,3}(\mathbf{0}_2, 1) - Y_{\ell,2}(\mathbf{0}_2)] = E [Y_{\ell,3}(\mathbf{0}_3) - Y_{\ell,2}(\mathbf{0}_2)] + \text{TE}_{\ell,3}. \quad (6.6)$$

Taking the difference between the two previous equations, and using the fact that $E [Y_{e,3}(\mathbf{0}_3) - Y_{e,2}(\mathbf{0}_2)]$ and $E [Y_{\ell,3}(\mathbf{0}_3) - Y_{\ell,2}(\mathbf{0}_2)]$ cancel each other out under the parallel-trends assumption,

$$E [\text{DID}_{\ell,e,2,3}] = \text{TE}_{\ell,3} - \text{TE}_{e,3} + \text{TE}_{e,2}. \quad (6.7)$$

Finally, it follows from (6.3), (6.4), and (6.7) that

$$E [\widehat{\beta}^{\text{fe}}] = \frac{1}{2} \text{TE}_{\ell,3} + \text{TE}_{e,2} - \frac{1}{2} \text{TE}_{e,3}. \quad (6.8)$$

In this simple example, the decomposition of $\widehat{\beta}^{\text{fe}}$ in Theorem 8 reduces to (6.8). The right-hand-side of the previous display is a weighted sum the effect of antibiotics on ℓ 's fever at period three, on e 's fever at period two, and on e 's fever at period three, with weights summing to one, and where e 's effect at period three is weighted negatively. Intuitively, the negative weight comes from the fact that e is treated at periods two and three, and $\text{DID}_{\ell,e,2,3}$, which uses e as a control group, subtracts its period-three effect out. If $\text{TE}_{\ell,3}$ and $\text{TE}_{e,2}$, and $\text{TE}_{e,3}$ are all negative, but $\text{TE}_{e,3}$ is more than three times larger in absolute value than both $\text{TE}_{\ell,3}$ and $\text{TE}_{e,2}$, for instance because antibiotics take time to act, $E [\widehat{\beta}^{\text{fe}}] > 0$.

Find an assumption on the treatment effects $\text{TE}_{g,t}$ such that under that supplementary assumption, the negative weight in the decomposition of $\widehat{\beta}^{\text{fe}}$ in (6.8) disappears.

$\hat{\beta}^{\text{fe}}$ estimates a convex combination of effects if treatment effects do not change over time. If one is ready to assume that $\text{TE}_{e,3} = \text{TE}_{e,2}$, (6.7) simplifies to

$$E[\text{DID}_{\ell,e,2,3}] = \text{TE}_{\ell,3}. \quad (6.9)$$

Then, the negative weight in (6.7) disappears, and $\hat{\beta}^{\text{fe}}$ estimates a weighted average of treatment effects. This extends beyond this simple example: Theorem S2 of the Web Appendix of de Chaisemartin and D'Haultfœuille (2020) and Equation (16) of Goodman-Bacon (2021) show that in staggered adoption designs with a binary treatment, $\hat{\beta}^{\text{fe}}$ estimates a convex combination of effects, if $\text{TE}_{g,t}$ does not depend on t . This conclusion, however, no longer holds if the treatment is not binary or the design is not staggered. Moreover, assuming that $\text{TE}_{g,t}$ does not depend on t is often implausible: this requires that treatment effects do not vary with length of exposure and with calendar time. The decomposition of $\hat{\beta}^{\text{fe}}$ under parallel trends and the assumption that treatment effects do not change over time in Theorem S2 of the Web Appendix of de Chaisemartin and D'Haultfœuille (2020) can be computed by the `twowayfeweights` Stata command, replacing `type(feTR)` by `type(feS)`.

Leveraging $\text{DID}_{\ell,e,2,3}$ to estimate the treatment's effect could lead to a biased estimator, if treatment effects are heterogeneous. But could there be an argument in favor of leveraging $\text{DID}_{\ell,e,2,3}$, as $\hat{\beta}^{\text{fe}}$ does?

Forbidden or efficient? Assume that Assumption PT holds, and that

$$Y_{g,t} - Y_{g,t}(\mathbf{0}_t) = \text{TE}_{g,t} = \beta \quad (6.10)$$

for some real number β . The first equality in the previous display requires that treatment effects are non-stochastic, while the second one requires that they do not vary across groups and time periods. Then,

$$Y_{g,t} = Y_{g,t}(\mathbf{0}_t) + D_{g,t}(Y_{g,t} - Y_{g,t}(\mathbf{0}_t)) \quad (6.11)$$

$$= \alpha_g + \gamma_t + \beta D_{g,t} + \varepsilon_{g,t}, \quad E[\varepsilon_{g,t}] = 0. \quad (6.12)$$

The second equality follows from (6.10) and from the fact that in Chapter 2, we saw that Assumption PT is equivalent to

$$Y_{g,t}(\mathbf{0}_t) = \alpha_g + \gamma_t + \varepsilon_{g,t}, \quad E[\varepsilon_{g,t}] = 0. \quad (6.13)$$

Thus, under Assumption PT and the constant effect condition in (6.10), $Y_{g,t}$ is generated by a population version of the TWFE regression in (3.1). Under those three assumptions, $\text{DID}_{e,\ell,1,2}$ and $\text{DID}_{\ell,e,2,3}$ are both unbiased for β : $\text{DID}_{\ell,e,2,3}$ is not a forbidden comparison anymore. Now, assume that the errors $\varepsilon_{g,t}$ are homoscedastic and i.i.d., both across g and t . Those are the assumptions under which OLS estimators are the best linear unbiased estimators of population regression coefficients, by the Gauss-Markov theorem. Then, $V(\hat{\beta}^{\text{fe}}) = 0.75 \times V(\text{DID}_{e,\ell,1,2})$, as predicted by the Gauss-Markov theorem. Thus, the reason why $\hat{\beta}^{\text{fe}}$ leverages $\text{DID}_{\ell,e,2,3}$ instead of just leveraging $\text{DID}_{e,\ell,1,2}$ is that doing so may lead to an unbiased estimator with a lower variance, if treatment effects are constant and errors are i.i.d. and homoscedastic.

The decomposition in (6.2) cannot be used to assess if $\hat{\beta}^{\text{fe}}$ estimates a convex combination of effects. Researchers sometimes use the sum of the weights on switchers-versus-always-treated DID in (6.2) as a diagnostic of the robustness of $\hat{\beta}^{\text{fe}}$ to heterogeneous treatment effects. We do not recommend this diagnostic, for the following reason. Let us first consider an example similar to that above, but with a third group n that remains untreated from period 1 to 3. In this second example, (6.2) now indicates that $\hat{\beta}^{\text{fe}}$ assigns a weight equal to 1/6 to DIDs comparing a switcher to a group treated at both periods. On the other hand, all the weights in the decomposition in Theorem 8 are now positive. This phenomenon can also arise in real data sets. In the data of Stevenson and Wolfers (2006) used by Goodman-Bacon (2021) in his empirical application, if one restricts the sample to states that are not always treated and to the first ten years of the panel, all the weights in Theorem 8 are positive, but the sum of the weights in (6.2) on DIDs comparing a switcher to a group treated at both periods is equal to 0.06. Beyond these examples, one can show that having DIDs comparing a switcher to a group treated at both periods in (6.2) is necessary but not sufficient to have negative weights in Theorem 8. Similarly, the sum of the weights on DIDs comparing a switcher to a group treated at both periods in (6.2) is always larger than the absolute value of the sum of the negative weights in Theorem 8. (6.2) “overestimates” the negative weights in Theorem 8, because as soon as there are three distinct

treatment dates, there is not a unique way of decomposing $\hat{\beta}^{\text{fe}}$ as a weighted average of DIDs, and there exists other decompositions than (6.2), putting less weight on DIDs using a group treated at both periods as the control group.²

Computation: Stata and R commands to compute the weights in (6.2). The `bacondecomp` Stata (see Goodman-Bacon et al., 2019) and R (see Flack and Edward, 2020) commands compute the $\text{DID}_{g,g',t,t'}$ entering in (6.2), the weights assigned to them, as well as the sum of the weights on $\text{DID}_{g,g',t,t'}$ using a group treated at both periods as the control group. The syntax of the `bacondecomp` Stata command is:

```
bacondecomp outcome treatment, detail
```

6.2.1.3 Assuming randomized treatment timing rather than parallel trends*

If groups' treatment date is randomly assigned, $\hat{\beta}^{\text{fe}}$ estimates a convex combination of effects. Theorem 8 holds under no-anticipation and parallel-trends assumptions. If instead of parallel of trends, one is ready to assume that treatment timing is as good as randomly assigned, then Athey and Imbens (2022) show that $\hat{\beta}^{\text{fe}}$ is unbiased for a convex combination of treatment effects, even if treatment effects change over time. Let us give some intuition for that result in the simple example we considered earlier, with two groups and three periods, such that one group starts receiving the treatment at period two, while the other starts receiving it at

²To see that, let $t_0 < t_1 < t_2$ be three dates, let e be an early-treated group becoming treated at t_1 , let ℓ be a late-treated group becoming treated at t_2 , and let n be a group untreated yet at t_2 . Let $\underline{v} = \min(v_{\ell,e,t_1,t_2}, v_{e,n,t_0,t_2}) > 0$. One has

$$\text{DID}_{\ell,e,t_1,t_2} = \text{DID}_{\ell,n,t_0,t_2} - \text{DID}_{e,n,t_0,t_2} + \text{DID}_{e,\ell,t_0,t_1}. \quad (6.14)$$

Then, it follows from (6.14) that

$$\begin{aligned} & v_{\ell,e,t_1,t_2} \text{DID}_{\ell,e,t_1,t_2} + v_{e,n,t_0,t_2} \text{DID}_{e,n,t_0,t_2} \\ &= (v_{\ell,e,t_1,t_2} - \underline{v}) \text{DID}_{\ell,e,t_1,t_2} + \underline{v} \text{DID}_{\ell,n,t_0,t_2} + \underline{v} \text{DID}_{e,\ell,t_0,t_1} + (v_{e,n,t_0,t_2} - \underline{v}) \text{DID}_{e,n,t_0,t_2}. \end{aligned} \quad (6.15)$$

Plugging (6.15) into (6.2) will yield a different decomposition of $\hat{\beta}^{\text{fe}}$ as a weighted average of DIDs. But the weight on DIDs using a group treated at both periods as the control group is equal to v_{ℓ,e,t_1,t_2} in the left-hand-side of (6.15), and to $(v_{\ell,e,t_1,t_2} - \underline{v})$ in its right-hand side. Accordingly, this new decomposition puts strictly less weight than (6.2) on DIDs using a group treated at both periods as the control group.

period three. Assume that the early-treated group is chosen at random: with probability 1/2, the early-treated group is group 1, and with probability 1/2 the early-treated group is group 2. Thus, e and ℓ are now random variables, with $P(e = 1) = P(e = 2) = 1/2$, and $\ell = 3 - e$. Consider, as in Athey and Imbens (2022), that potential outcomes are non-stochastic. Then, $\text{TE}_{g,t} = Y_{g,t} - Y_{g,t}(\mathbf{0}_t)$, and

$$\begin{aligned}\widehat{\beta}^{\text{fe}} &= (\text{DID}_{e,\ell,1,2} + \text{DID}_{\ell,e,2,3})/2 \\ &= \frac{1}{2}\text{TE}_{\ell,3} + \text{TE}_{e,2} - \frac{1}{2}\text{TE}_{e,3} \\ &\quad + \frac{1}{2}(Y_{e,2}(\mathbf{0}_2) - Y_{e,1}(0) - (Y_{\ell,2}(\mathbf{0}_2) - Y_{\ell,1}(0))) + \frac{1}{2}(Y_{\ell,3}(\mathbf{0}_3) - Y_{\ell,2}(\mathbf{0}_2) - (Y_{e,3}(\mathbf{0}_3) - Y_{e,2}(\mathbf{0}_2))).\end{aligned}$$

For any random variable X , let $E_e(X)$ denote the expectation of X with respect to the identity of the early-treated group. We have

$$\begin{aligned}E_e\left[\frac{1}{2}\text{TE}_{\ell,3} + \text{TE}_{e,2} - \frac{1}{2}\text{TE}_{e,3}\right] &= P(e = 1)\left[\frac{1}{2}\text{TE}_{2,3} + \text{TE}_{1,2} - \frac{1}{2}\text{TE}_{1,3}\right] + P(e = 2)\left[\frac{1}{2}\text{TE}_{1,3} + \text{TE}_{2,2} - \frac{1}{2}\text{TE}_{2,3}\right] \\ &= \frac{1}{2}\left[\frac{1}{2}\text{TE}_{2,3} + \text{TE}_{1,2} - \frac{1}{2}\text{TE}_{1,3}\right] + \frac{1}{2}\left[\frac{1}{2}\text{TE}_{1,3} + \text{TE}_{2,2} - \frac{1}{2}\text{TE}_{2,3}\right] \\ &= \frac{1}{2}[\text{TE}_{1,2} + \text{TE}_{2,2}].\end{aligned}$$

Using similar steps, one can show that

$$E_e\left[\frac{1}{2}(Y_{e,2}(\mathbf{0}_2) - Y_{e,1}(0) - (Y_{\ell,2}(\mathbf{0}_2) - Y_{\ell,1}(0))) + \frac{1}{2}(Y_{\ell,3}(\mathbf{0}_3) - Y_{\ell,2}(\mathbf{0}_2) - (Y_{e,3}(\mathbf{0}_3) - Y_{e,2}(\mathbf{0}_2)))\right] = 0.$$

Therefore,

$$E_e\left[\widehat{\beta}^{\text{fe}}\right] = \frac{1}{2}[\text{TE}_{1,2} + \text{TE}_{2,2}],$$

so the expectation of $\widehat{\beta}^{\text{fe}}$ with respect to the design and conditional on potential outcomes is equal to the average effect of the treatment at period two.

If groups' treatment date is randomly assigned, cross-sectional comparisons of treated and control groups controlling for their baseline outcomes may be more efficient than TWFE regressions. In the previous chapter, we saw that if $T = 2$, all units are untreated at period 1, and some are randomly assigned to treatment at period 2, $\widehat{\beta}^{\text{fe}}$ can have a higher variance than the treatment coefficient in a regression of $Y_{g,2}$ on an intercept and $D_{g,2}$ controlling for

$Y_{g,1}$ (Frison and Pocock, 1992; McKenzie, 2012). Roth and Sant'Anna (2023) build upon those results, and derive an efficient estimator of the ATT in binary-and-staggered designs with $T \geq 3$ and randomly assigned treatment dates. Again, the efficient estimator is not a DID or TWFE estimator: it is a weighted average of cross-sectional comparisons of treated and controls units controlling for baseline outcomes.

As-good-as-random treatment timing is a strong assumption, which should be thoroughly tested. In natural experiments where researchers do not effectively randomize treatment timing, a randomization claim should be substantiated with thorough and preferably pre-specified balancing checks showing that groups' treatment timing does not predict covariates correlated with the outcome. For instance, if F_g is randomly assigned, one should have

$$F_g \perp\!\!\!\perp (Y_{g,1}(\mathbf{0}_t), \dots, Y_{g,T}(\mathbf{0}_T)).$$

If there are no always-treated groups, all groups are untreated till period $\underline{F} - 1$. Then, for all $t \leq \underline{F} - 1$ $Y_{g,t} = Y_{g,t}(\mathbf{0}_t)$ for all g . Then, the previous display implies that

$$F_g \perp\!\!\!\perp (Y_{g,1}, \dots, Y_{g,\underline{F}-1}), \quad (6.16)$$

an equation that only involves observed variables, and can therefore be tested. To test (6.16), one can for instance run a pooled regression of $Y_{g,t}$ on adoption-cohort FEs for all $t \leq \underline{F} - 1$, or one can regress F_g on $(Y_{g,1}, \dots, Y_{g,\underline{F}-1})$ and run an F-test that all coefficients are equal to zero.

Application to the UDL example. The data starts in 1956. 2 states had already passed a UDL in 1956, but of the remaining 49 states, none passes a UDL before 1969. Thus we observe $Y_{g,t}(\mathbf{0}_t)$, the divorce rate without a UDL, of 49 states for 13 years. If treatment timing is as good as randomly assigned, we should have $Y_{g,t}(\mathbf{0}_t) \perp\!\!\!\perp F_g$, which we can test by regressing $Y_{g,t}$ on FEs for each adoption cohorts, in the subsample of g s untreated in 1956 and $t < 1969$. Unfortunately, many adoption cohorts contain only one state, so running an F-test from a regression of $Y_{g,t}$ on FEs for each value of F_g is infeasible. Instead, using the `wolfers_didtextbook` dataset, regress the divorce rate on FEs for each possible value of the `early_late_never` variable, which groups together states adopting before the median adoption year, those adopting after that year, and the never adopters. Weight the regression by `stpop`, and cluster standard errors at the `state`

level. Can you reject the null that those FEs do not predict the divorce rate? Then, is it the case that treatment timing is as good as randomly assigned?

```
reg div_rate i.early_late_never if cohort!=1956&year<=1968 [w=stpop], vce(cluster state)
```

The p-value of an F-test that all coefficients in the regression are equal to 0 is 0.0015: treatment timing significantly predicts states' divorce rates without a UDL, so treatment timing does not seem to be as good as randomly assigned in this application.

6.2.2 Dynamic TWFE estimators

6.2.2.1 TWFE Event-Study regressions

In Design BST, to estimate dynamic effects and test the no-anticipation and parallel-trends assumptions, researchers have often estimated the following TWFE ES regression, which is similar to that in (3.6) but accommodates groups' heterogeneous treatment dates:

$$Y_{g,t} = \sum_{g'=1}^G \hat{\alpha}_{g'} 1\{g = g'\} + \sum_{t'=1}^T \hat{\gamma}_{t'} 1\{t = t'\} + \sum_{\ell=-K, \ell \neq 0}^L \hat{\beta}_\ell^{\text{fe}} 1\{t = F_g - 1 + \ell\} + \hat{\epsilon}_{g,t}. \quad (6.17)$$

In words, the outcome is regressed on group and period FEs, and relative-time indicators $1\{t = F_g - 1 + \ell\}$ equal to 1 if at t , group g has been exposed to treatment for ℓ periods. For $\ell \geq 1$, $\hat{\beta}_\ell^{\text{fe}}$ is supposed to estimate the cumulative effect of ℓ periods of exposure to treatment. For $\ell \leq -1$, $\hat{\beta}_\ell^{\text{fe}}$ is supposed to be a placebo coefficient testing the parallel-trends assumption, by comparing the outcome trends of groups that will and will not start receiving the treatment in $|\ell - 1|$ periods. Researchers have sometimes estimated a variant of this regression, where the first and last indicators $1\{t = F_g - 1 - K\}$ and $1\{t = F_g - 1 + L\}$ are respectively replaced by an indicator for being at least K periods away from the period before adoption ($1\{t \leq F_g - 1 - K\}$) and an indicator for having been treated for at least L periods ($1\{t \geq F_g - 1 + L\}$). Such endpoint binning is for instance recommended by Schmidheiny and Siegloch (2023): without it,

the regression implicitly assumes that the treatment no longer has any effect after L periods. Instead, with endpoint binning the regression assumes that the treatment effect is constant after L periods, a more plausible assumption.

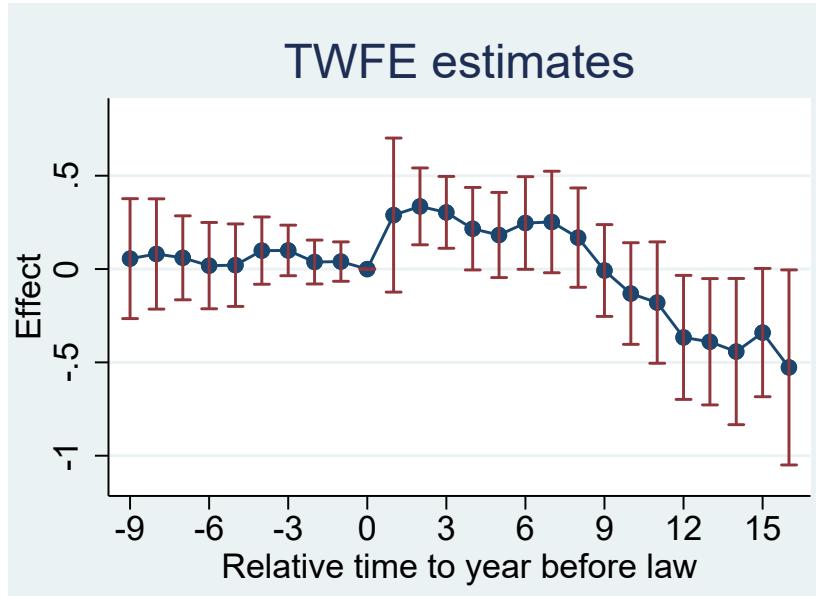
Application to the UDL example. Using the `wolfers_didtextbook` dataset, estimate the event-study TWFE regression in (6.17) with divorce rate as the outcome and relative time indicators to a UDL as the treatments, with $L = 16$, $K = 9$, and endpoint binning, weighting the regression by the state's population and clustering standard errors at the state level. According to this regression, do UDLs have an effect on the divorce rate? Do Assumptions NA and PT seem to hold?

```
reg div_rate rel_time* i.state i.year [w=stpop], vce(cluster state)
```

Figure 6.2 below shows the regression's coefficients. According to this regression, UDLs increase the divorce rate for 8 years. After 12 years, the effect becomes significantly negative. Thus, the small and insignificant coefficient in the static TWFE regression was actually hiding positive short-term effects and negative long-run effects. The pre-trends estimates are small, and individually and jointly insignificant (F-test p-value=0.863). They are also substantially smaller than the estimated effects of UDLs, and the first three pre-trends are precisely estimated: it does not seem that violations of parallel trends can fully account for UDLs' estimated short-run effects. If anything, pre-trends indicate that states that adopt a UDL experience more negative pre-trends than states that do not adopt a UDL, which could bias estimated effects downwards. The positive short-term effects of UDLs are easy to understand: UDLs make it easier for couples to divorce, thus increasing the divorce rate. The negative long-run effects are more puzzling. Perhaps the couples induced to divorce by UDLs would have divorced anyways later, and the treatment just induces them to divorce sooner. It could also be the case that in the long-run, there are differential trends between states adopting and not adopting a UDL, which are not detected by the pre-trend tests: pre-trends coefficients are more noisy in the long than in the short run. The results in Figure 6.2 are consistent with those in Column (1) of Table 2 of

Wolfers (2006). The event-study regression in Figure 6.2 and that in Wolfers (2006) differ on two dimensions: Wolfers (2006) does not include any placebo indicator for pre-adoption periods, and he includes post-adoption indicators for bins of two years (one indicator for the adoption year and the year after that, one indicator for the two following years, etc.). Results seem robust to those specification choices.

Figure 6.2: Effects of Unilateral Divorce Laws, according to a TWFE ES regression



Note: This figure shows the estimated effects of Unilateral Divorce Laws on the divorce rate, as well as placebo pre-trends estimates, using the data of Wolfers (2006), according to the TWFE event-study regression in (6.17), with $L = 16$, $K = 9$, and endpoint binning. The estimation is weighted by states' populations. Standard errors are clustered at the state level. 95% confidence intervals are shown in red.

Without never-treated groups, the “fully-dynamic” event-study regression is not identified. Let

$$\bar{K} = \left(\max_{g: F_g \leq T, F_g > 1} F_g - 1 \right) - 1, \quad \bar{L} = T - \min_{g: F_g > 1} F_g + 1.$$

$\max_{g: F_g \leq T, F_g > 1} F_g - 1$ is the largest number of time periods over which a group untreated at the start of panel is observed prior to getting treated. Thus, \bar{K} is the largest number of pre-trends coefficients one can hope to estimate. Similarly, \bar{L} is the largest number of time periods over

which a group untreated at the start of panel is observed after getting treated. Thus, \bar{L} is the largest exposure to treatment whose effect one can estimate. The TWFE ES regression with $K = \bar{K}$ and $L = \bar{L}$ is called the “fully-dynamic” regression. Below, we restate Proposition 1 in Borusyak et al. (2024).

Proposition 1 *If all groups are eventually treated and $K = \bar{K}$ and $L = \bar{L}$, the regressors in (6.17) are perfectly collinear: if $(\beta_\ell)_{\ell=-\bar{K}, \dots, \bar{L}, \ell \neq 0}$ solves the OLS problem, so does $(\beta_\ell + \kappa\ell)_{\ell=-\bar{K}, \dots, \bar{L}, \ell \neq 0}$, for any $\kappa \in \mathbb{R}$.*

Hence, the fully-dynamic TWFE ES specification requires never-treated groups. Otherwise some pre-trends coefficients should be removed, or dynamic effects should be restricted, for instance by binning at endpoints.

Proof of Proposition 1.* We have:

$$\begin{aligned} \sum_{\substack{\ell=-\bar{K} \\ \ell \neq 0}}^{\bar{L}} \mathbb{1}\{F_g = t - \ell + 1\} \ell &= \sum_{\ell=-\bar{K}}^{\bar{L}} \mathbb{1}\{F_g = t - \ell + 1\} \ell \\ &= (t + 1 - F_g) \sum_{\ell=-\bar{K}}^{\bar{L}} \mathbb{1}\{F_g = t - \ell + 1\} \\ &= \underbrace{t + 1}_{\text{enter in time FE}} \quad \underbrace{-F_g}_{\text{enter in group FE}} \end{aligned}$$

QED.

6.2.2.2 Decomposition of TWFE ES regressions

TWFE ES regressions are not robust to heterogeneous effects, and may suffer from a contamination bias. The result below essentially follows from Proposition 3 in Sun and Abraham (2021):³

³(6.18) follows from Proposition 3 in Sun and Abraham (2021), assuming no binning. A slight difference is that the decomposition in Sun and Abraham (2021) gathers groups that started receiving the treatment at the same period into cohorts. Their decomposition can be further decomposed, as in Theorem 9.

Theorem 9 In Design BST, under Assumptions NA and PT, and if $L = \bar{L}$ and there are never-treated groups, then for $\ell \in \{1, \dots, \bar{L}\}$,

$$E[\hat{\beta}_\ell^{fe}] = \sum_{g:F_g-1+\ell \leq T} W_{g,\ell}^\ell TE_{g,\ell}^r + \sum_{\substack{\ell'=1 \\ \ell' \neq \ell}}^{\bar{L}} \sum_{g:F_g-1+\ell' \leq T} W_{g,\ell'}^\ell TE_{g,\ell'}^r, \quad (6.18)$$

where $W_{g,\ell}^\ell$ and $W_{g,\ell'}^\ell$ are weights such that $\sum_{g:F_g-1+\ell \leq T} W_{g,\ell}^\ell = 1$ and $\sum_{g:F_g-1+\ell' \leq T} W_{g,\ell'}^\ell = 0$ for every $\ell' \in \{1, \dots, \bar{L}\}$, $\ell' \neq \ell$.

The first summation in the right-hand side of (6.18) is a weighted sum across groups of the cumulative effect of ℓ treatment periods, with weights summing to 1 but that may be negative. This first summation resembles that in the decomposition of $\hat{\beta}^{fe}$ in Theorem 8, and it implies that $\hat{\beta}_\ell^{fe}$ may be biased for ATT_ℓ if the cumulative effect of ℓ treatment periods varies across groups. Interpret the second summation in the right-hand side of (6.18).

This second summation is a weighted sum, across $\ell' \neq \ell$ and groups, of the cumulative effect of ℓ' treatment periods in group g , with weights summing to 0. This second summation was not present in the decomposition of $\hat{\beta}^{fe}$. In view of this second summation, is $\hat{\beta}_\ell^{fe}$ estimating the cumulative effect of ℓ treatment periods, or is this coefficient contaminated by other effects?

The presence of this second summation implies that $\hat{\beta}_\ell^{fe}$, which is supposed to estimate the cumulative effect of ℓ treatment periods, may in fact be contaminated by the effects of ℓ' treatment periods. As $\sum_{g:F_g-1+\ell' \leq T} W_{g,\ell'}^\ell = 0$ for every ℓ' , this second summation disappears if $TE_{g,\ell'}^r$ does not vary across groups, but this is often an implausible assumption: this rules out treatment effect heterogeneity across groups, but also over time as groups reach their ℓ th treatment period at different points in time. Importantly, Sun and Abraham (2021) show that the negative result

in Theorem 9 is not specific to the fully-dynamic TWFE ES specification: similar negative results also apply to less flexible ES regressions. This negative result is also not specific to TWFE ES regressions with never-treated groups: similar negative results also apply to regressions with no never-treated groups.

Proof of Theorem 9.* By the Frisch-Waugh-Lovell theorem,

$$\hat{\beta}_\ell^{\text{fe}} = \frac{\sum_{g,t} \eta_{g,t}^\ell Y_{g,t}}{\sum_{g,t} \eta_{g,t}^\ell \mathbb{1}\{t = F_g - 1 + \ell\}}. \quad (6.19)$$

where $\eta_{g,t}^\ell$ is the residual of the regression of $\mathbb{1}\{t = F_g - 1 + \ell\}$ on group and time fixed effects and the indicators $(\mathbb{1}\{t = F_g - 1 + \ell'\})_{-\bar{K} \leq \ell' \leq \bar{L}, \ell' \notin \{0, \ell\}}$. Besides, from (6.11) and (6.13), we obtain

$$\begin{aligned} Y_{g,t} &= \alpha_g + \gamma_t + (Y_{g,t} - Y_{g,t}(\mathbf{0}_t))D_{g,t} + \varepsilon_{g,t} \\ &= \alpha_g + \gamma_t + \sum_{\ell'=1}^{\bar{L}} [Y_{g,t}(\mathbf{0}_{t-\ell'}, \mathbf{1}_{\ell'}) - Y_{g,t}(\mathbf{0}_t)] \mathbb{1}\{t = F_g - 1 + \ell'\} + \varepsilon_{g,t}. \end{aligned}$$

where we recall that $E[\varepsilon_{g,t}] = 0$. By definition of $\eta_{g,t}^\ell$, we have $\sum_g \eta_{g,t}^\ell = 0$ for all t and $\sum_t \eta_{g,t}^\ell = 0$ for all g . Then,

$$\begin{aligned} \sum_{g,t} \eta_{g,t}^\ell Y_{g,t} &= \sum_{g,t} \sum_{\ell'=1}^{\bar{L}} \eta_{g,t}^\ell [Y_{g,t}(\mathbf{0}_{t-\ell'}, \mathbf{1}_{\ell'}) - Y_{g,t}(\mathbf{0}_t)] \mathbb{1}\{t = F_g - 1 + \ell'\} + \sum_{g,t} \eta_{g,t}^\ell \varepsilon_{g,t} \\ &= \sum_{\ell'=1}^{\bar{L}} \sum_{g:F_g-1+\ell' \leq T} \eta_{g,F_g-1+\ell'}^\ell [Y_{g,F_g-1+\ell'}(\mathbf{0}_{F_g-1}, \mathbf{1}_{\ell'}) - Y_{g,t}(\mathbf{0}_{F_g-1+\ell'})] + \sum_{g,t} \eta_{g,t}^\ell \varepsilon_{g,t}. \end{aligned}$$

Hence,

$$E \left[\sum_{g,t} \eta_{g,t}^\ell Y_{g,t} \right] = \sum_{g:F_g-1+\ell \leq T} \eta_{g,F_g-1+\ell}^\ell \text{TE}_{g,\ell}^r + \sum_{\substack{\ell'=1 \\ \ell' \neq \ell}}^{\bar{L}} \sum_{g:F_g-1+\ell' \leq T} \eta_{g,F_g-1+\ell'}^\ell \text{TE}_{g,\ell'}^r. \quad (6.20)$$

Next, observe that

$$\sum_{g,t} \eta_{g,t}^\ell \mathbb{1}\{t = F_g - 1 + \ell\} = \sum_{g:F_g-1+\ell \leq T} \eta_{g,F_g-1+\ell}^\ell.$$

Let $W_{g,\ell}^\ell = \eta_{g,F_g-1+\ell}^\ell / \sum_{g:F_g-1+\ell \leq T} \eta_{g,F_g-1+\ell}^\ell$ for $(\ell, \ell') \in \{1, \dots, \bar{L}\}^2$. Then, by (6.19) and (6.20),

$$E \left[\hat{\beta}_\ell^{\text{fe}} \right] = \sum_{g:F_g-1+\ell \leq T} W_{g,F_g-1+\ell}^\ell \text{TE}_{g,\ell}^r + \sum_{\ell' \neq \ell, \ell' > 0} \sum_{g:F_g-1+\ell' \leq T} W_{g,F_g-1+\ell'}^\ell \text{TE}_{g,\ell'}^r.$$

By definition of $W_{g,\ell}^\ell$, $\sum_g W_{g,\ell}^\ell = 1$. Moreover, by construction of the residual $\eta_{g,t}^\ell$,

$$\sum_{g:F_g-1+\ell' \leq T} \eta_{g,F_g-1+\ell'}^\ell = \sum_{g,t} \eta_{g,t}^\ell \mathbb{1}\{t = F_g - 1 + \ell'\} = 0.$$

This implies that $\sum_{g:F_g-1+\ell' \leq T} W_{g,\ell'}^\ell = 0$ **QED.**

If treatment effects are heterogeneous, TWFE ES regressions cannot be used to test the no-anticipation and parallel-trends assumptions. For $\ell \leq -1$, researchers typically use $\hat{\beta}_\ell^{\text{fe}}$ to test the no anticipation and parallel-trends assumptions, Assumptions NA and PT. Sun and Abraham (2021) show that without making any assumption, those pre-trends TWFE ES coefficients are unbiased for the sum of two terms, hereafter referred to as Terms A and B. As intended, Term A measures differential outcome trends prior to treatment, between groups getting treated at different dates. Under Assumptions NA and PT, Term A is equal to zero. But Term B is similar to the second summation in the right-hand side of (6.18): a weighted sum, across $\ell' \geq 1$ and groups, of the cumulative effect of ℓ' treatment periods in group g , with weights summing to zero. Even if Assumptions NA and PT hold, the expectation of Term B may differ from zero. Thus, due to the presence of Term B, the expectation of $\hat{\beta}_\ell^{\text{fe}}$ may differ from zero even if Assumptions NA and PT hold. Conversely, it could be that the expectation of Term A differs from zero, meaning that Assumptions NA and PT fail, but is exactly offset by the expectation of Term B. Then, the expectation of $\hat{\beta}_\ell^{\text{fe}}$ may be equal to zero even if Assumptions NA and PT fail. Thus, an important consequence of the results in Sun and Abraham (2021) is that in the presence of heterogeneous treatment effects, the TWFE ES regression in (6.17) cannot be used to test the no-anticipation and parallel-trends assumptions, because pre-trend coefficients are contaminated by actual treatment effects.

Computation: Stata commands to compute the weights attached to any TWFE ES regression. The `eventstudyweights` Stata command (see Sun, 2020) computes the weights attached to TWFE ES regressions. Its syntax is:

```
eventstudyweights {rel_time_list}, absorb(i.groupid i.timeid)
cohort(first_treatment) rel_time(ry),
```

where `rel_time_list` is the list of relative-time indicators $1\{t = F_g - 1 + \ell\}$ included in (6.17), `first_treatment` is a variable equal to the period when group g got treated for the first time, and `ry` is a variable equal to `timeid` minus `first_treatment`, the number of periods elapsed since group g started receiving the treatment. Another option to decompose TWFE ES regressions is to use the `twowayfeweights` Stata command. The syntax is similar to that given above to decompose the static TWFE regression, except that one needs to account for the fact that

unlike the static regression, the TWFE ES regressions does not have one but several treatment variables, all the relative time indicators. To decompose, say, $\hat{\beta}_1^{\text{fe}}$, one needs to input the first relative time indicator as the treatment variable, while all the other relative time indicators need to be inputted in the `other_treatments` option. If the TWFE ES regression has relative time indicators for time periods before treatment adoption, those need to be inputted to the `controls` option.

6.2.2.3 The origin of the contamination weights

Intuition for the contamination of event-study estimators: endpoint binning. Consider a simple design with $G = 2$, $T = 4$, an early-treated group that gets treated at $t = 2$, and a late-treated group that gets treated at $t = 3$. We estimate an ES TWFE regression with $L = 2$, endpoint binning, and no pre-trends. One can show that

$$\hat{\beta}_1^{\text{fe}} = \frac{1}{2}\text{DID}_{e,\ell,1,2} + \frac{1}{2}\text{DID}_{\ell,e,2,4}. \quad (6.21)$$

Let us momentarily assume that the effect of being exposed to treatment for three periods is the same as the effect of being exposed for two periods ($\text{TE}_{g,3}^r = \text{TE}_{g,2}^r$), as is implicitly assumed by the endpoint binning in the regression. From period two to four, group ℓ goes from zero to two periods of exposure to treatment, while group e goes from one to three periods of exposure. Then, one can show that

$$E(\text{DID}_{\ell,e,2,4}) = \text{TE}_{\ell,2}^r - (\text{TE}_{e,3}^r - \text{TE}_{e,1}^r) = \text{TE}_{e,1}^r + (\text{TE}_{\ell,2}^r - \text{TE}_{e,2}^r), \quad (6.22)$$

where the second equality follows from our assumption that $\text{TE}_{g,3}^r = \text{TE}_{g,2}^r$. If effects of two periods of exposure do not vary across groups, the previous display reduces to

$$E(\text{DID}_{\ell,e,2,4}) = \text{TE}_{e,1}^r.$$

Thus, $\text{DID}_{\ell,e,2,4}$ is a valid estimator of the effect of one period of exposure to treatment under the assumptions implicitly made by the TWFE ES regression. Then, assume that the errors in the population version of the TWFE ES regression are homoscedastic and i.i.d., both across g and t . Then one can show that $V(\hat{\beta}_1) = 0.75 \times V(\text{DID}_{e,\ell,1,2})$, as predicted by the Gauss-Markov theorem. Thus, the reason why $\hat{\beta}_1$ leverages $\text{DID}_{\ell,e,2,4}$ instead of just leveraging $\text{DID}_{e,\ell,1,2}$ is that

doing so may lead to an unbiased estimator with a lower variance. But leveraging $\text{DID}_{\ell,e,2,4}$ might lead to a bias if being exposed to treatment for three periods does not have the same effect as being exposed for two periods, or if treatment effects vary across groups.

Intuition for the contamination of event-study estimators: no endpoint binning.* In the previous example, we consider a TWFE ES regression with endpoint binning. But contamination weights can also arise without endpoint binning. Consider a simple design with $G = 3$, $T = 4$, an early-treated group that gets treated at $t = 2$, an on-time group o that gets treated at $t = 3$, and a late-treated group ℓ that gets treated at $t = 4$. We estimate an ES TWFE regression with $L = \bar{L} = 3$, and no pre-trends. There is no simple decomposition of $\hat{\beta}_1^{\text{fe}}$ as a weighted average of 2×2 DIDs in this design. Still, one can show that $\hat{\beta}_1^{\text{fe}}$ leverages

$$\text{DID}_{\ell,o,3,4} - \text{DID}_{\ell,e,2,3},$$

the difference between a DID comparing the late and on-time groups from period three to four, and a DID comparing the late and early groups from period two to three. From period three to four, group ℓ goes from zero to one period of exposure to treatment, while group o goes from one to two periods of exposure. Similarly, ℓ is untreated at periods two and three, while e goes from one to two periods of exposure. Then, one can show that under Assumptions NA and PT, $\text{DID}_{\ell,o,3,4} - \text{DID}_{\ell,e,2,3}$ is unbiased for

$$\text{TE}_{\ell,1}^r - (\text{TE}_{o,2}^r - \text{TE}_{o,1}^r - (\text{TE}_{e,2}^r - \text{TE}_{e,1}^r)).$$

If effects of one and two periods of exposure do not vary across groups, the previous display reduces to $\text{TE}_{\ell,1}^r$. Again, $\hat{\beta}_1^{\text{fe}}$ leverages a comparison that is valid if treatment effects are homogeneous between groups, but that lead it to be contaminated by effects of having been exposed to treatment for two periods if effects are heterogeneous.

Intuition for the contamination of pre-trend estimators. Consider a simple design with $G = 3$, $T = 3$, an early-treated group that gets treated at $t = 2$, a late-treated group that gets treated at $t = 3$, and a never-treated group. Assume that one estimates a TWFE ES regression with $K = \bar{K} = 1$ and $L = \bar{L} = 2$, the fully-dynamic specification in this example. Then, one

can show that

$$\begin{aligned}\hat{\beta}_{-1} &= \text{DID}_{n,\ell,1,2} + \frac{2}{5} [\text{DID}_{e,n,1,2} - \text{DID}_{\ell,n,2,3}] \\ &= Y_{n,2}(\mathbf{0}_2) - Y_{n,1}(0) - (Y_{\ell,2}(\mathbf{0}_2) - Y_{\ell,1}(0)) \\ &\quad + \frac{2}{5} \left[Y_{e,2}(0, 1) - Y_{e,1}(0) - (Y_{n,2}(\mathbf{0}_2) - Y_{n,1}(0)) \right. \\ &\quad \left. - (Y_{\ell,3}(\mathbf{0}_2, 1) - Y_{\ell,2}(\mathbf{0}_2) - (Y_{n,3}(\mathbf{0}_3) - Y_{n,2}(\mathbf{0}_2))) \right]\end{aligned}$$

Therefore, under Assumptions NA and PT,

$$E[\hat{\beta}_{-1}] = \frac{2}{5} [\text{TE}_{e,1}^r - \text{TE}_{\ell,1}^r] :$$

$\hat{\beta}_{-1}$ is contaminated by a weighted sum of the effects of one period of exposure to treatment in the early and late treated groups, with weights that sum to zero. This contamination term disappears if $\text{TE}_{e,1}^r = \text{TE}_{\ell,1}^r$, as implicitly assumed by the TWFE ES regression. If the population TWFE ES regression is correctly specified, and if its errors are homoscedastic and i.i.d. across g and t , $\text{DID}_{n,\ell,1,2}$ and $\hat{\beta}_{-1}$ are both unbiased for β_{-1} , the coefficient comparing the outcome evolutions of groups that will be treated in one period and groups that will not be treated in one period, and $V(\hat{\beta}_{-1}) = 0.6 \times V(\text{DID}_{n,\ell,1,2})$. Thus, leveraging $2/5[\text{DID}_{e,n,1,2} - \text{DID}_{\ell,n,2,3}]$ instead of just leveraging $\text{DID}_{n,\ell,1,2}$, the “natural” pre-trends estimator in this example, may lead to an unbiased estimator with a lower variance. But if the TWFE ES regression is misspecified because of heterogeneous treatment effects, the expectation of $2/5[\text{DID}_{e,n,1,2} - \text{DID}_{\ell,n,2,3}]$ is no longer equal to 0, so $\hat{\beta}_{-1}$ is no longer unbiased for the differential trend of groups that will be treated in one period and groups that will not be treated in one period.

6.2.2.4 Application to the UDL example

Use `twowayfeweights` to decompose $\hat{\beta}_1^{\text{fe}}$, and compute the correlation between the weights and the year variable.⁴ Interpret the results.

⁴We use the `twowayfeweights` Stata command, because it has an option to compute the correlation between the weights and other variables.

```
twoWayFeweights div_rate state year rel_time1, type(feTR) test_random_weights(year)
weight(stpop) other_treatments(rel_time2-rel_time16)
controls(rel_timeminus1-rel_timeminus9)
```

As shown in (6.18), $\hat{\beta}_1^{\text{fe}}$ can be decomposed as the sum of two terms. The first term is a weighted sum of effects of having been exposed to a UDL for one year, across 27 states, where all effects receive a positive weight.⁵ The weights are negatively correlated with the year variable (correlation=−0.232), so this first term upweights effects in states passing a law early, and downweights effects in states passing a law late. Accordingly, this first term estimates a convex combination of effects, but it may still differ from ATT_1 , the average effect of having been exposed to a UDL for one year across all states, if the effect of having been exposed to a UDL for one year varies between early- and late-adopting states. The second term is a weighted sum of being exposed to a UDL for more than one year. 29 effects of having been exposed to a UDL for two years enter in that second term. 16 enter with a positive weight, and 13 enter with a negative weight. The positive and negative weights respectively sum to 0.012 and −0.012. 28 effects of having been exposed to a UDL for three years enter in that second term. 10 effects enter with a positive weight, and 18 enter with a negative weight. The positive and negative weights respectively sum to 0.010 and −0.010. Effects of having been exposed to a UDL for four, five, ..., 15, and more than 16 years also enter in that second term. In total, the positive and negative contamination weights respectively sum to 0.064 and −0.064. If UDLs' effects vary across states, that second term may not be equal to zero, thus further biasing $\hat{\beta}_1^{\text{fe}}$ with respect to its target parameter ATT_1 . However, those contamination weights are not very large, so this bias is likely to be small. Overall, $\hat{\beta}_1^{\text{fe}}$ does not have very large negative weights and contamination weights attached to it, presumably because there is a large share of never-treated groups in this application. Results from the decompositions of the other event-study coefficients $\hat{\beta}_\ell^{\text{fe}}$ for $\ell \geq 2$ are similar.

⁵Two of the 29 states that pass a UDL have missing divorce rates the year when they pass the law, which is why effects in 27 rather than 29 states enter in the decomposition.

6.2.2.5 Local projection regressions*

Local-projection panel regressions. In binary-and-staggered designs, another common method to estimate dynamic effects are regressions of leads of the outcome on group and period FEs and the treatment. Such regressions have sometimes been described as a panel-data version of the local-projection method originally proposed by Jordà (2005) for time-series data. For every $\ell \in \{1, \dots, T-1\}$ let $\hat{\beta}_\ell^{\text{lp}}$ denote the coefficient on $D_{g,t}$ in a regression of $Y_{g,t-1+\ell}$ on group and period FEs and $D_{g,t}$, in the subsample such that $1 \leq t \leq T-\ell+1$:

$$Y_{g,t-1+\ell} = \sum_{g'=1}^G \hat{\alpha}_{g'} 1\{g = g'\} + \sum_{t'=1}^T \hat{\gamma}_{t'} 1\{t = t'\} + \hat{\beta}_\ell^{\text{lp}} D_{g,t} + \hat{\epsilon}_{g,t}. \quad (6.23)$$

Intuitively, researchers hope that $\hat{\beta}_\ell^{\text{lp}}$ estimates the average effect of being treated at period t on groups' period- $t-1+\ell$ outcome, thus allowing them to use $\ell \mapsto \hat{\beta}_\ell^{\text{lp}}$ to estimate $\ell \mapsto \text{ATT}_\ell$.

A decomposition of the coefficients in local-projection panel regressions.

Theorem 10 *In Design BST, if $D_{g,1} = 0$ for all g , then under Assumptions NA and PT, $\forall \ell \in \{1, \dots, T-1\}$ such that $\hat{\beta}_\ell^{\text{lp}}$ is well defined,*

$$E \left[\hat{\beta}_\ell^{\text{lp}} \right] = \sum_{\ell'=1}^{T-(F-1)} \sum_{g:\ell \leq F_g - 1 + \ell' \leq T} W_{g,\ell'}^{\text{lp},\ell} TE_{g,\ell'},$$

where $\min_{g,\ell'} W_{g,\ell'}^{\text{lp},\ell} < 0$ for all $\ell \geq 2$, and

$$\sum_{\ell'=1}^{T-F+1} \sum_{g:\ell \leq F_g - 1 + \ell' \leq T} W_{g,\ell'}^{\text{lp},\ell} < 1$$

for all $\ell \in \{2, \dots, F\}$.

The proof, omitted here, can be found in de Chaisemartin and D'Haultfœuille (2025). Theorem 10 shows that $\hat{\beta}_\ell^{\text{lp}}$ estimates a linear combination, across ℓ' and g , of the effect of ℓ' period of exposure to treatment in group g . Accordingly, $\hat{\beta}_\ell^{\text{lp}}$ does not estimate an average across groups of the effect of ℓ periods of exposure to treatment: like the TWFE ES coefficient $\hat{\beta}_\ell^{\text{lp}}$, $\hat{\beta}_\ell^{\text{lp}}$ is contaminated by effects of other lengths of exposure. A further issue is that some of the weights may be negative, and for $\ell \geq 2$, some weights are always negative. A last and perhaps even more concerning issue is that for $\ell \in \{2, \dots, F\}$, the weights sum to strictly less than one. This

implies that even if there is a, say, positive real number δ such that $\text{TE}_{g,\ell}^r = \delta$, meaning that the treatment effect does not vary with length of exposure or across groups, $E(\hat{\beta}_\ell^{\text{lp}}) < \delta$: the local-projection regression is downwards biased. This is because the regression is misspecified: it considers groups with $D_{g,t} = 0$ as untreated, whereas some of them may actually have become treated at some point between $t+1$ and $t-1+\ell$. In their empirical application, de Chaisemartin and D'Haultfœuille (2025) exhibit an example where some local-projection regression coefficients estimate weighted sums of effects where the sum of the weights is negative. Then, even if the treatment effect is constant across length of exposure and groups, $E(\hat{\beta}_\ell^{\text{lp}})$ could be of a different sign than the treatment effect.

Bibliographic notes. Theorem 10 was shown by de Chaisemartin and D'Haultfœuille (2025). This result is specific to (6.23), an extension of local projection regressions to panel data. Those issues are absent in the time-series context considered by Jordà (2005), under the assumptions imposed therein. Those issues are also absent in the careful extension of local projection regressions to panel data with a binary and staggered treatment recently proposed by Dube, Girardi, Jorda and Taylor (2023).

6.3 Heterogeneity-robust estimators

6.3.1 Target parameters

In Design BST, groups can be aggregated into cohorts that start receiving the treatment at the same period. Let $\mathcal{C} = \{c \in \{2, \dots, T\} : \exists g : F_g = c\}$ denote the set of dates at which at least one group adopts the treatment. \mathcal{C} is the set of all adoption cohorts. $\underline{F} = \min \mathcal{C}$ denotes the earliest adoption cohort. For all $c \in \mathcal{C}$ and $t \in \{1, \dots, T\}$, let $\bar{Y}_{c,t}$ denote the average outcome at period t across groups belonging to cohort c . At period T , cohort c has been treated for $T - (c - 1)$ periods. Callaway and Sant'Anna (2021) and Sun and Abraham (2021) define a first set of parameters of interest as

$$\text{TE}_{c,\ell}^r = E \left[\bar{Y}_{c,c-1+\ell}(\mathbf{0}_{c-1}, \mathbf{1}_\ell) - \bar{Y}_{c,c-1+\ell}(\mathbf{0}_{c-1+\ell}) \right],$$

for all $c \in \mathcal{C}$ and $\ell \in \{1, \dots, T - (c - 1)\}$. $\text{TE}_{c,\ell}^r$ are cohort-specific event-study effects. Letting G_c denote the number of groups in adoption-cohort c , we have that for all $\ell \leq T - (\underline{F} - 1)$,

$$\text{ATT}_\ell = \sum_{c:c-1+\ell \leq T} \frac{G_c}{G_\ell} \text{TE}_{c,\ell}^r. \quad (6.24)$$

6.3.2 DID estimators

6.3.2.1 Estimators

Unbiased estimator of $\text{TE}_{c,\ell}^r$. Let $\bar{Y}_{n,t}$ denote the average outcome at period t across groups that remain untreated from period 1 to T , hereafter referred to as the never-treated groups, assuming for now that such groups exist. To estimate $\text{TE}_{c,\ell}^r$, Callaway and Sant'Anna (2021) and Sun and Abraham (2021) propose

$$\widehat{\text{TE}}_{c,\ell}^{\text{did}} = \bar{Y}_{c,c-1+\ell} - \bar{Y}_{c,c-1} - (\bar{Y}_{n,c-1+\ell} - \bar{Y}_{n,c-1}),$$

a DID estimator comparing the period- $c - 1$ -to- $c - 1 + \ell$ outcome evolution in cohort c and in the never-treated groups n .

Theorem 11 *In Design BST, under Assumptions NA and PT, for $c \in \mathcal{C}$ and $\ell \in \{1, \dots, T - (c - 1)\}$,*

$$E \left[\widehat{\text{TE}}_{c,\ell}^{\text{did}} \right] = \text{TE}_{c,\ell}^r. \quad (6.25)$$

Theorem 11 shows that $\widehat{\text{TE}}_{c,\ell}^{\text{did}}$ is unbiased for a well-defined treatment effect parameter, $\text{TE}_{c,\ell}^r$, under Assumptions NA and PT alone, even if the treatment effect is heterogeneous, across groups or over time. Intuitively, why is it that unlike $\widehat{\beta}^{\text{fe}}$, $\widehat{\text{TE}}_{c,\ell}^{\text{did}}$ is robust to heterogeneous treatment effects?

We saw that $\widehat{\beta}^{\text{fe}}$ is not robust to heterogeneous treatment effects, because it leverages DIDs comparing a group going from untreated to treated to a group treated at both periods. $\widehat{\text{TE}}_{c,\ell}^{\text{did}}$

does not leverage such comparisons, as it compares groups going from untreated to treated to groups untreated at both periods.

Based on (6.24) and Theorem 11, propose an unbiased estimator of ATT_ℓ .

Unbiased estimator of ATT_ℓ and ATT . It directly follows from (6.24) and Theorem 11 that

$$\widehat{\text{ATT}}_\ell^{\text{did}} := \sum_{c:c-1+\ell \leq T} \frac{G_c}{G_\ell} \widehat{\text{TE}}_{c,\ell}^{\text{did}}$$

is unbiased for ATT_ℓ . Then, it follows from (6.1) and Theorem 11 that if no group is treated at period one and there are never-treated groups,

$$\widehat{\text{ATT}}^{\text{did}} := \sum_\ell \frac{G_\ell}{N_1} \widehat{\text{ATT}}_\ell^{\text{did}}$$

is unbiased for the ATT .

Proof of Theorem 11.

$$\begin{aligned} & E \left[\widehat{\text{TE}}_{c,\ell}^{\text{did}} \right] \\ &= E \left[\bar{Y}_{c,c-1+\ell}(\mathbf{0}_{c-1}, \mathbf{1}_\ell) - \bar{Y}_{c,c-1}(\mathbf{0}_{c-1}) - (\bar{Y}_{n,c-1+\ell}(\mathbf{0}_{c-1+\ell}) - \bar{Y}_{n,c-1}(\mathbf{0}_{c-1})) \right] \\ &= E \left[\bar{Y}_{c,c-1+\ell}(\mathbf{0}_{c-1}, \mathbf{1}_\ell) - \bar{Y}_{c,c-1+\ell}(\mathbf{0}_{c-1+\ell}) \right] \\ &\quad + E \left[\bar{Y}_{c,c-1+\ell}(\mathbf{0}_{c-1+\ell}) - \bar{Y}_{c,c-1}(\mathbf{0}_{c-1}) - (\bar{Y}_{n,c-1+\ell}(\mathbf{0}_{c-1+\ell}) - \bar{Y}_{n,c-1}(\mathbf{0}_{c-1})) \right] \\ &= \text{TE}_{c,\ell}^r. \end{aligned}$$

Justify each step of this derivation.

The first equality follows from the definition of $\widehat{\text{TE}}_{c,\ell}^{\text{did}}$, Design BST, and Assumption NA. The

second equality follows from adding and subtracting $\bar{Y}_{c,c-1+\ell}(\mathbf{0}_{c-1+\ell})$. The third equality follows from Assumption PT QED.

6.3.2.2 Extensions

Pre-trend tests of Assumptions NA and PT. For $c \in \{3, \dots, T\}$ and $\ell \in \{1, \dots, c-2\}$, let

$$\widehat{\text{TE}}_{c,-\ell}^{\text{did}} = \bar{Y}_{c,c-1-\ell} - \bar{Y}_{c,c-1} - (\bar{Y}_{n,c-1-\ell} - \bar{Y}_{n,c-1})$$

be a placebo DID estimator comparing the outcome evolution in cohort c and in the never-treated groups, from period $c-1$ to $c-1-\ell$, namely over ℓ periods before cohort c got treated. $\widehat{\text{TE}}_{c,-\ell}^{\text{did}}$ exactly mimicks $\widehat{\text{TE}}_{c,\ell}^{\text{did}}$, the estimator of the effect of having been exposed to treatment for ℓ periods in cohort c . To mimick $\widehat{\text{ATT}}_{\ell}^{\text{did}}$, one may then use

$$\widehat{\text{ATT}}_{-\ell}^{\text{did}} := \sum_{c:c-1-\ell \geq 1} \frac{G_c}{G_{-\ell}^{\text{pl}}} \widehat{\text{TE}}_{c,-\ell}^{\text{did}},$$

where $G_{-\ell}^{\text{pl}} = \sum_{c:c-1-\ell \geq 1} G_c$, for any $\ell \geq 1$ such that $G_{-\ell}^{\text{pl}} > 0$. Even if $\widehat{\text{ATT}}_{-\ell}^{\text{did}}$ perfectly mimicks $\widehat{\text{ATT}}_{\ell}^{\text{did}}$, two differences are worth mentioning. First, groups such that $c-1+\ell > T$ are included in $\widehat{\text{ATT}}_{-\ell}^{\text{did}}$ but not in $\widehat{\text{ATT}}_{\ell}^{\text{did}}$ because they have not reached ℓ periods of exposure to treatment yet at period T . Second, groups such that $c-1-\ell < 1$ are included in $\widehat{\text{ATT}}_{\ell}^{\text{did}}$ but not in $\widehat{\text{ATT}}_{-\ell}^{\text{did}}$, because we do not observe their outcome evolution over ℓ periods before they adopt the treatment. One can show that under Assumptions NA and PT, for $\ell \geq 1$ $E[\widehat{\text{ATT}}_{-\ell}^{\text{did}}] = 0$, so one can reject Assumptions NA and PT if $\widehat{\text{ATT}}_{-\ell}^{\text{did}}$ is significantly different from zero. Note that this test of Assumptions NA and PT is robust to heterogeneous treatment effects, unlike the test based on the pre-trends coefficients from a TWFE ES regression.

Using the not-yet-treated instead of the never-treated as controls. Callaway and Sant'Anna (2021) propose estimators similar to those above, but that use the not-yet-treated instead of the never-treated as controls. For instance, all groups not yet treated at period c can be used as control groups in the definition of $\widehat{\text{TE}}_{c,1}^{\text{did}}$. This extension is very useful, for at least three reasons. First, when there is no never-treated group, the effects $\text{TE}_{c,\ell}^r$ can still be estimated, for every $c \geq 2$ and $\ell \geq 1$ such that $c-1+\ell \leq \underline{T}$, where \underline{T} is the last period when at least one group is still untreated. Without never-treated groups, Sun and Abraham (2021) propose to use the last

treated cohort as the control group, but this may result in imprecise estimators when that cohort is small. Second, even when there are never-treated groups, one may worry that such groups are less comparable to groups that get treated at some point, and researchers sometimes prefer to discard them and use only not-yet-but-eventually-treated groups as controls. While doing so can increase the plausibility of the parallel-trends assumption, it can also reduce statistical precision if the never-treated account for a large proportion of the sample. Therefore, we recommend that practice only when dropping the never-treated substantially reduces pre-trend estimates. Third, even if there are never-treated groups and one is fine with keeping them in the analysis, the not-yet-treated is a larger control group, and may lead to more precise estimators. When there are never-treated groups, the only argument against using not-yet-treated as controls is that not-yet-treated might be more subject to anticipation effects.

Estimators with covariates. Callaway and Sant'Anna (2021) also propose estimators relying on a conditional parallel-trends assumption, which extend $\text{DID}_{X,\text{dr}}$, the parametric doubly-robust estimator with covariates reviewed in Section 4.1, to binary-and-staggered designs. See Section 5 of Ahrens, Chernozhukov, Hansen, Kozbur, Schaffer and Wiemann (2025) for a debiased-machine-learning estimator that extends $\text{DID}_{X,\text{dr-ml}}$ to binary-and-staggered designs.

The estimators proposed by de Chaisemartin and D'Haultfœuille (2020) and de Chaisemartin and D'Haultfœuille (2025). In staggered adoption designs with a binary treatment, the DID_M estimator proposed by de Chaisemartin and D'Haultfœuille (2020) also uses the not-yet-treated as controls, and is identical to the estimator of ATT_1 using the not-yet-treated as controls in Callaway and Sant'Anna (2021). As we will discuss in Chapter 8, de Chaisemartin and D'Haultfœuille (2025) build upon Callaway and Sant'Anna (2021) and de Chaisemartin and D'Haultfœuille (2020) to propose event-study estimators that can be used in designs with a non-binary and/or non-staggered treatment. Their estimators can of course also be used in designs with a binary and staggered treatment. Then, without covariates or weighting and a balanced panel, their event-study estimators are numerically equivalent to the estimators of Callaway and Sant'Anna (2021) using the not-yet-treated as controls. Their placebo estimators are similar, except that groups such that $c - 1 + \ell > T$ are not included in their placebo ℓ , to make the samples on which the placebo and event-study estimators are computed more comparable.

Sensitivity analysis under bounded differential trends. The approach discussed in Section 4.3, to bound ATT_ℓ and derive a confidence interval for it under a bounded differential trends assumption, may also be used in Design BST, using the estimated effects and placebos $\widehat{\text{ATT}}_\ell^{\text{did}}$, and their variance-covariance matrix as inputs to the procedure. The only caveat is that because $\widehat{\text{ATT}}_\ell^{\text{did}}$ does not apply to the exact same groups as the placebo $\widehat{\text{ATT}}_{-\ell}^{\text{did}}$, the bounded differential trends assumption underlying the procedure is a little less appealing here: the differential trends experienced by groups included in placebo $\widehat{\text{ATT}}_{-\ell}^{\text{did}}$ may differ from the differential trends experienced by groups included in the actual estimator $\widehat{\text{ATT}}_\ell^{\text{did}}$. One could restrict the estimation sample to groups included in $\widehat{\text{ATT}}_{-\ell}^{\text{did}}$ and $\widehat{\text{ATT}}_\ell^{\text{did}}$, though that might lead to power losses.

Testing (5.8).* Under Assumptions NA and PT, one can show that for all $g : 2 \leq F_g \leq T$ and $t \in \{F_g, F_g + 1, \dots, T\}$,

$$\widehat{\text{TE}}_{g,t} = Y_{g,t} - Y_{g,F_g-1} - (\bar{Y}_{n,t} - \bar{Y}_{n,F_g-1})$$

is unbiased for $\text{TE}_{g,t}$, group g 's treatment effect at period t . The estimators $\widehat{\text{TE}}_{g,t}$ can be used to test (5.8), a condition, discussed in the previous chapter, which ensures that the TWFE estimator is unbiased for the ATT, by requiring that the weights attached to the TWFE estimator are uncorrelated to treatment effects. If no group is treated at period one, then $\text{TE}_{g,t}$ can be unbiasedly estimated for all treated (g, t) cells. Then, the covariance between the weights and the treatment effects can be unbiasedly estimated using

$$\sum_{(g,t): D_{g,t} \neq 0} \left(W_{g,t} - \frac{1}{N_1} \right) \left(\widehat{\text{TE}}_{g,t} - \widehat{\text{ATT}}^{\text{did}} \right).$$

Then, one can reject (5.8) when this estimator significantly differs from zero.

6.3.2.3 Inference

Callaway and Sant'Anna (2021) propose bootstrap confidence intervals (CIs), which are asymptotically valid when the number of groups goes to infinity, under the assumption that groups are independent. de Chaisemartin and D'Haultfœuille (2020) and de Chaisemartin and D'Haultfœuille (2025) propose analytic CIs, which are asymptotically valid, conditional on the design, when the number of groups goes to infinity, also under the assumption that groups are independent. Those

CIs can be conservative (overly wide) when some adoption cohorts contain only one group: with a cohort of size one the variance of $\bar{Y}_{c,c-1+\ell} - \bar{Y}_{c,c-1}$ cannot be unbiasedly estimated and is instead conservatively estimated by $(\bar{Y}_{c,c-1+\ell} - \bar{Y}_{c,c-1})^2$ (de Chaisemartin, Ciccia, D'Haultfoeuille, Knau, Malézieux and Sow, 2024). To avoid this issue, researchers with adoption cohorts containing only one group can coarsen their time variable (e.g. aggregate a daily panel at the weekly level), to ensure that most adoption cohorts have at least two groups, if they are ready to slightly mismeasure groups' exact adoption dates and length of exposure to treatment. Alternatively, they can exclude adoption cohorts with only one group from the estimation sample, if those cohorts account for a small fraction of the sample size. Finally, they may also use bootstrap instead of analytic CIs: in simulations, de Chaisemartin, Ciccia, D'Haultfoeuille, Knau, Malézieux and Sow (2024) find that bootstrap CIs have close-to-nominal coverage in a DGP where all adoption cohorts contain only one group.

6.3.2.4 Numerical equivalences with regression coefficients

Obtaining the estimators $\widehat{\text{TE}}_{c,\ell}^{\text{did}}$ and $\widehat{\text{TE}}_{c,\ell}^{\text{did}}$ from a TWFE ES regression. Sun and Abraham (2021) show that the estimators $\widehat{\text{TE}}_{c,\ell}^{\text{did}}$ can be obtained from the following TWFE ES regression:

$$Y_{g,t} = \sum_{c \in \mathcal{C}} \widehat{\alpha}_c 1\{g \in c\} + \sum_{t'=1}^T \widehat{\gamma}_{t'} 1\{t = t'\} + \sum_{c \in \mathcal{C}} \sum_{\ell=-(c-2), \ell \neq 0}^{T-(c-1)} \widehat{\text{TE}}_{c,\ell}^{\text{did}} 1\{F_g = c, t = c-1+\ell\} + \widehat{\epsilon}_{g,t}. \quad (6.26)$$

With respect to (6.17), (6.26) has cohort instead of group fixed effects, and the relative time indicators are interacted with cohort FEs, thus allowing for heterogeneous effects across adoption cohorts. Note that when control variables are included in this regression, it is no longer guaranteed to estimate a convex combination of the cohort-specific effects $\text{TE}_{c,\ell}^r$.

“Local-projection” regressions.* Dube et al. (2023) propose estimators related to those of Callaway and Sant’Anna (2021) with not-yet-treated groups as controls, that can be obtained by ordinary-least-squares regressions. Their estimator of the effect of having been exposed to treatment for ℓ periods is the coefficient on $D_{g,t} - D_{g,t-1}$ in a so-called local-projection regression (Jordà, 2005) of $Y_{g,t-1+\ell} - Y_{g,t-1}$ on period FEs and $D_{g,t} - D_{g,t-1}$ in the subsample of (g, t) such

that $F_g = t$ or $F_g > t - 1 + \ell$. This estimator is not unbiased for ATT_ℓ , but it is unbiased for a so-called “variance-weighted” convex combination of the cohort-specific effects $\text{TE}_{c,\ell}^r$, with weights proportional to the number of groups such that $F_g = c$ or $F_g > c - 1 + \ell$, multiplied by the variance of $D_{g,c} - D_{g,c-1}$ in the subsample such that $F_g = c$ or $F_g > c - 1 + \ell$ (Angrist, 1998). Dube et al. (2023) also propose a weighted version of their local-projection regressions, with weights that undo the implicit “variance-weighting” in their regression, thus yielding estimators of event-study effects numerically equivalent to those of Callaway and Sant’Anna (2021) with not-yet-treated groups as controls.

6.3.2.5 Computation: Stata and R commands to compute heterogeneity-robust DID estimators

Estimators proposed by Sun and Abraham (2021). The estimators proposed by Sun and Abraham (2021) are computed by the `eventstudyinteract` Stata command (see Sun, 2021). Its syntax is

```
eventstudyinteract outcome {rel_time_list}, absorb(i.groupid i.timeid)
cohort(first_treatment) control_cohort(controlgroup)
```

where `rel_time_list` is the list of relative-time indicators one would include in the event-study regression in (6.17), `first_treatment` is a variable equal to the period when group g got treated for the first time, and `controlgroup` is an indicator for the control group observations (e.g.: the never treated). The command has a an option to include covariates, by including them in the regression in (6.26). The resulting estimators are not guaranteed to estimate a convex combination of the cohort-specific effects $\text{TE}_{c,\ell}^r$.

Estimators proposed by Callaway and Sant’Anna (2021). The estimators proposed by Callaway and Sant’Anna (2021) are computed by the `csdid` Stata command (see Rios-Avila, Sant’Anna and Callaway, 2021), and by the `did` R command (see Sant’Anna and Callaway, 2021). The syntax of the Stata command is

```
csdid outcome, time(timeid) gvar(cohort)
```

where `cohort` is equal to the period when a group starts receiving the treatment. The estimators of Callaway and Sant’Anna (2021) are also implemented in the Stata functions `xthdidregress` and `hdidregress`.

Estimators proposed by de Chaisemartin and D’Haultfœuille (2025). The estimators proposed by de Chaisemartin and D’Haultfœuille (2025) are computed by the `did_multiplegt_dyn` Stata (see de Chaisemartin, Ciccia, D’Haultfœuille, Knau, Malézieux and Sow, 2024b) and `didmultiplegt dyn` R (see de Chaisemartin, Ciccia, D’Haultfœuille, Knau, Malézieux and Sow, 2024a) commands. The syntax of the Stata command is

```
did_multiplegt_dyn outcome groupid timeid treatment, effects(#) placebo(#)
```

where the number inputted to the `effects` option is the number of effects ATT_ℓ the user would like to estimate, while the number inputted to the `placebo` option is the number of pre-trend coefficients the user would like to estimate.

Estimators proposed by Dube et al. (2023).* The estimators proposed by Dube et al. (2023) are computed by the `1pdid` Stata command (Busch and Girardi, 2023). Without control variables, those estimators are very fast to compute, because they rely on OLS regressions. With control variables, the command can compute two types of estimators. Without the `rw` option, estimators with covariates are very fast to compute, but they assume that treatment effects do not vary with the covariates, and if that assumption fails they may estimate a non-convex combination of cohort-specific effects. With the `rw` option, estimators with covariates remain valid if treatment effects vary with the covariates, but the command then relies on the bootstrap for inference, and may therefore be slower. Note that with the `rw` option, the command also estimates ATT_ℓ rather than a “variance-weighted” average of cohort-specific effects.

6.3.2.6 Application: the effect of UDL laws on the divorce rate

Estimators of Sun and Abraham (2021). Using the `wolfers_didtextbook` dataset, compute the same event-study and pre-trend estimators as in the TWFE ES regression, but using the estimators of Sun and Abraham (2021), and using never-treated groups as the control cohort. Pay attention to the fact that with this command, the cohort variable has to be missing for the never-treated groups. Are the results you get very different from those you obtained with the TWFE ES regression?

```
replace cohort =. if cohort==0
eventstudy interact div_rate rel_time* [aweight=stpop], absorb(i.state i.year)
cohort(cohort) control_cohort(controlgroup) vce(cluster state)
```

The top-right panel of Figure 6.3 shows the estimators proposed by Sun and Abraham (2021), computed using the `eventstudy interact` Stata command. The estimated effects are very similar to the TWFE ES coefficients in the top-left panel. This could either be due to the fact that UDLs' effects are not very heterogeneous, or to the fact that the event-study regression does not have very large negative weights and contamination weights attached to it, as shown above. Interestingly, the confidence intervals are, if anything, slightly wider in the top-left than in the top-right panel of Figure 6.3, thus showing that heterogeneity-robust DID estimators are not always less precise than TWFE estimators. The placebos are individually insignificant.

Estimators of Callaway and Sant'Anna (2021). Using the `wolfers_didtextbook` dataset, compute the same event-study and pre-trend estimators with this command, using not-yet-treated groups as the control groups. Pay attention to the fact that with this command, the cohort variable has to be equal to 0 for the never-treated groups. Interpret the results.

```
replace cohort=0 if cohort==.
csdid div_rate [weight=stpop], ivar(state) time(year) gvar(cohort) notyet agg(event)
```

The bottom-left panel of Figure 6.3 shows the estimators proposed by Callaway and Sant'Anna (2021), computed using the `csdid` Stata command, using not-yet-treated states as the control group. The estimated effects are very similar to those obtained with the `eventstudy interact` command. 19 states never adopt a UDL over the period under consideration, so the group of never-treated states used as controls by `eventstudy interact` is quite large, and accounts for a relatively large fraction of the not-yet-treated states used as controls by `csdid`. This may explain why in this application, the two commands yield very similar estimates. Using the larger control group of not-yet-treated states also does not lead to markedly more precise estimates: the widths of the confidence intervals are similar in the two panels. The placebos produced by `csdid` are

small and individually insignificant. The placebos are smaller in the bottom-left than in the top-right panel. This is because by default, `csdid` computes first-difference placebos, comparing the outcome evolution of treated and not-yet treated states, before the treated start receiving the treatment, and between pairs of consecutive periods. On the other hand, `eventstudyinteract` computes long-difference placebos. `csdid` computes long-difference placebos when the `long` option is specified.

Estimators of de Chaisemartin and D'Haultfœuille (2025). Using the `wolfers_didtextbook` dataset, compute the same event-study and pre-trend estimators with this command.

```
did_multiplegt_dyn div_rate state year udl, effects(16) placebo(9) weight(stpop)
```

The estimators we obtain are very close to those of Callaway and Sant'Anna (2021), so we do not report them on Figure 6.3.

6.3.3 Imputation estimators

Borusyak et al. (2024), Gardner (2021), and Liu et al. (2024) have proposed imputation estimators, that differ from the DID estimators discussed in the previous section.

6.3.3.1 Estimators

The estimators in Borusyak et al. (2024) can be obtained by running a TWFE regression of the outcome on group and time FEs, and FEs for every treated (g, t) cell:

$$Y_{g,t} = \sum_{g'=1}^G \widehat{\alpha}_{g'} 1\{g = g'\} + \sum_{t'=1}^T \widehat{\gamma}_{t'} 1\{t = t'\} + \sum_{g',t'} \widehat{\text{TE}}_{g',t'}^{\text{imp}} 1\{g = g', t = t'\} D_{g,t} + \widehat{\epsilon}_{g,t}.$$

To estimate $\text{TE}_{g,t}$, we use $\widehat{\text{TE}}_{g,t}^{\text{imp}}$, the FE for treated cell (g, t) in this regression. Then, to estimate $\text{TE}_{c,\ell}^r$, one can use the average of all the $\widehat{\text{TE}}_{g,t}^{\text{imp}}$ such that group g started receiving the treatment at period c and $t = c - 1 + \ell$. One can show that the resulting estimators are unbiased

under Assumptions NA and PT, and are therefore robust to heterogeneous treatment effects.

Intuitively, why is it that those estimators are robust to heterogeneous treatment effects?

Because the TWFE regression in the previous display does not impose any restriction on treatment effect heterogeneity, as it has one coefficient per treated (g, t) cell.

6.3.3.2 Some numerical equivalences

TWFE ES regression with cohort and cohort \times time-since-adoption FEs. Wooldridge (2021) considers a TWFE regression of the outcome on cohort, period, and cohort \times time-since-adoption FEs, namely

$$Y_{g,t} = \sum_{c \in \mathcal{C}} \hat{\alpha}_c 1\{F_g = c\} + \sum_{t'=1}^T \hat{\gamma}_{t'} 1\{t = t'\} + \sum_{c \in \mathcal{C}} \sum_{\ell=1}^{T-(c-1)} \widehat{\text{TE}}_{c,\ell}^{\text{imp}} 1\{F_g = c, t = c - 1 + \ell\} + \hat{\epsilon}_{g,t}.$$

One can show that the coefficients on $1\{F_g = c, t = c - 1 + \ell\}$ in this regression are numerically equivalent to the estimators of $\text{TE}_{c,\ell}^r$ proposed by Borusyak et al. (2024). Without variation in treatment timing, the equation in the previous display reduces to that in (3.29). Thus, without variation in treatment timing imputation event-study estimators are equivalent to the event-study coefficients in the TWFE ES regression in (3.28), which, instead of using the period just before treatment as the baseline period, use the average of all pre-treatment periods. This numerical equivalence also shows that the only difference between the imputation estimators and the TWFE ES regression of Sun and Abraham (2021) is that the one above does not have cohort \times time-to-adoption FEs to test the parallel trends and no-anticipation assumptions.

Imputation. As in Chapter 3, another numerically equivalent way of computing the estimators in Borusyak et al. (2024) amounts to fitting a TWFE regression of the outcome on group and time FEs in the sample of untreated (g, t) cells, and using that regression to predict the counterfactual untreated outcome of treated cells. Estimates of the treatment effect of those cells are then merely obtained by subtracting their counterfactual to their actual outcome.

6.3.3.3 Extensions

Estimators with covariates. To control for covariates, one can just include them in the imputation first-step. As only untreated observations are used in that first step, this yields an estimator robust to heterogeneous treatment effects, unlike what happens when one simultaneously estimates the coefficients on the covariates and the treatment effect. Similarly, the imputation procedure can easily be extended to triple-differences, or models with group-specific linear trends.

Pre-trend tests of Assumptions NA and PT. To test Assumptions NA and PT, Borusyak et al. (2024) propose to estimate a TWFE regression among all the untreated (g, t) cells, with K leads of the treatment, and use the leads' coefficients as placebos. An issue with this strategy is that changing the number of leads from K to $K + 1$ will change the coefficients on the first K leads: results of the pre-trends test will be sensitive to the researcher's choice of K . Instead, with the pre-trend estimators of Sun and Abraham (2021), Callaway and Sant'Anna (2021), and de Chaisemartin and D'Haultfoeuille (2025), the first K pre-trends estimators and their standard errors are not affected by the addition of one more pre-trend estimator.

Bibliographic notes. Before Borusyak et al. (2024), Liu et al. (2024) and Gardner (2021) have also proposed numerically-equivalent imputation estimators.⁶ The result showing that imputation estimators are efficient with independent and identically distributed (i.i.d.) errors, which we discuss below, is shown in Borusyak et al. (2024).

6.3.3.4 Inference

Borusyak et al. (2024) propose variance estimators and confidence intervals (CIs) based on an asymptotic approximation where the number of groups goes to infinity, under the assumption that groups are independent. Their CIs are valid conditional on the design. Gardner, Thakral, Tô and Yap (2024) show that imputation estimators can be cast as generalized-method-of-moments (GMM) estimators, which gives rise to different variance estimators and CIs. In sim-

⁶We saw in Chapter 4 that even before that, Hsiao et al. (2012), Gobillon and Magnac (2016), and Xu (2017) have proposed a similar strategy to estimate treatment effects under a TWFE model with interactive fixed effects.

ulations, Gardner et al. (2024) find that when cohorts have a small number of groups, the CIs of Borusyak et al. (2024) undercover while their CIs have closer-to-nominal size.

6.3.3.5 Computation: Stata and R commands to compute heterogeneity-robust imputation estimators

The `did_imputation` Stata command (see Borusyak, 2021) and the `didimputation` R command (see Butts, 2021b) compute the estimators proposed by Borusyak et al. (2024). The syntax of the Stata `did_imputation` command is:

```
did_imputation outcome groupid timeid cohort,
```

where `cohort` is a variable equal to the period when group g first got treated. The `did2s` Stata (Butts and Gardner, 2021) and R (Butts, 2021a) commands compute the estimators proposed by Gardner (2021). The `fetc` Stata (Liu et al., 2022b) and R (Liu et al., 2022a) command compute the estimators proposed by Liu et al. (2024).

6.3.3.6 Application: the effect of UDL laws on the divorce rate

Using the `wolfers_didtextbook` dataset, compute the same number of event-study and pre-trend estimators as before, with the `did_imputation` command. Pay attention to the fact that with this command, the cohort variable has to be missing for the never-treated groups. Interpret the results.

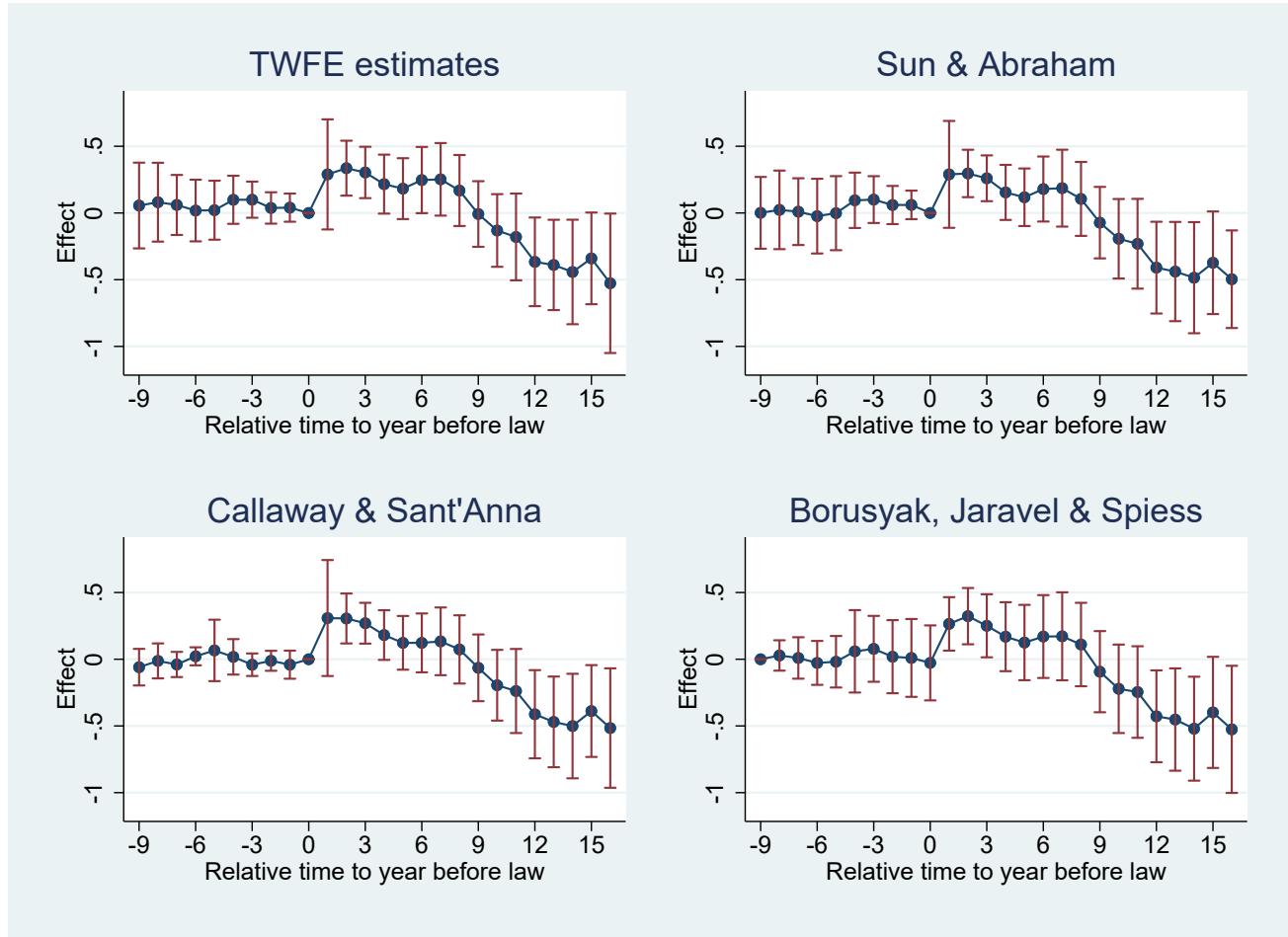
```
replace cohort=. if cohort==0
did_imputation div_rate state year cohort [aweight=stpop], horizons(0/15) autosample
minn(0) pre(9)
```

The bottom-right panel of Figure 6.3 shows the estimators proposed by Borusyak et al. (2024), computed using the `did_imputation` command. The effects are very similar to those found with the previous estimators. The placebos produced by `did_imputation` are small, individually in-

significant, and jointly insignificant (F-test p-value = 0.541).⁷ As explained earlier, the placebos computed by `did_imputation` are different from those computed by the other commands: the command estimates a TWFE regression among all the untreated (g, t) cells, with K leads of the treatment, and uses the leads' coefficients as the placebos. To be consistent with the other estimations, we run the command with 9 leads. Then, everything is relative to 10 periods prior to treatment, which is why the reference period is set at $t = -10$ in the bottom-right panel.

⁷We did not report a joint test that all placebos are equal to 0 based on `eventstudyinteract`: this command does not readily allow to compute this test, as it does not return the covariances between the estimators. Similarly, `csdid` does not allow to jointly test if the placebos in Figure 6.3 are significant: it computes a joint nullity test, but for more disaggregated, cohort-specific placebos.

Figure 6.3: Effects of Unilateral Divorce Laws, using four estimation methods



Note: This figure shows the estimated effects of Unilateral Divorce Laws on the divorce rate, as well as placebo pre-trends estimates, using the data of Wolfers (2006) and four estimation methods. In the top-left panel, we show estimated effects per the event-study regression in (6.17), with $L = 16$, $K = 9$, and endpoint binning. In the top-right (resp. bottom-left, bottom-right) panel, we show estimated effects per the `eventstudyinteract` (resp. `csdid`, `did_imputation`) Stata commands. All estimations are weighted by states' populations. Standard errors are clustered at the state level. 95% confidence intervals are shown in red.

6.3.4 Comparing the properties of those estimators, and their software implementation

Based on their number of downloads from the SSC repository as of June 2024, it seems that `eventstudyinteract`, `csdid`, `did_multiplegt_dyn`, and `did_imputation` currently are the most commonly used commands for heterogeneity-robust DID estimation. In what follows, we compare the estimators, variance estimators, and confidence intervals produced by those four commands, as well as their functionalities.

6.3.4.1 Variance

Comparing the variances of the heterogeneity-robust estimators. If the errors $\varepsilon_{g,t}$ in (2.4) are independent and identically distributed across g and t , Borusyak et al. (2024) show that their estimator is the best linear unbiased estimator (BLUE) of $\text{TE}_{c,\ell}^r$, thus implying that it is more efficient than the estimators of Callaway and Sant'Anna (2021) and Sun and Abraham (2021). As discussed in Chapter 3, in binary designs without variation in treatment timing, the estimator of ATT_ℓ proposed by Borusyak et al. (2024) reduces to $\widehat{\beta}_\ell^{\text{imp}}$, while that proposed by Callaway and Sant'Anna (2021) and Sun and Abraham (2021) reduces to the event-study coefficient $\widehat{\beta}_\ell^{\text{fe}}$. In Chapter 3, we have seen that the efficiency ranking of the two estimators reverses if one assumes that the errors in (2.4) follow a random walk instead of being independent over time. Harmon (2022) shows that this remains for the most part true with variation in treatment timing. He shows that the estimator of ATT_1 proposed by de Chaisemartin and D'Haultfœuille (2020) and Callaway and Sant'Anna (2021), using the not-yet-treated as controls, is BLUE under (2.4), if errors follow a random walk. For $\ell > 1$, he shows that neither the estimators of Callaway and Sant'Anna (2021) nor those of Borusyak et al. (2024) are BLUE under random-walk-errors. The BLUE of $\text{TE}_{c,\ell}^r$ is

$$\widehat{\text{TE}}_{c,\ell}^{\text{chain}} = \sum_{k=1}^{\ell} \left(\bar{Y}_{c,c-1+k} - \bar{Y}_{c,c-1+k-1} - (\bar{Y}_{nyt,c-1+k,c-1+k} - \bar{Y}_{nyt,c-1+k,c-1+k-1}) \right),$$

where for all (t', t) , $\bar{Y}_{nyt,t,t'}$ denotes the average outcome, at period t' , of groups not-yet-treated at t . With the not-yet treated as the control group, the estimator of Callaway and Sant'Anna (2021) compares the $c-1$ to $c-1+\ell$ outcome evolution of cohort c and groups not-yet-treated

at $c - 1 + \ell$. Instead, $\widehat{\text{TE}}_{c,\ell}^{\text{chain}}$ is a “chained” DID estimator, comparing the $c - 1$ to c outcome evolution of cohort c and groups not-yet-treated at c , and adding to it a comparison of the c to $c + 1$ outcome evolution of cohort c and groups not-yet-treated at $c + 1$ etc., until period $c - 1 + \ell$. This chained DID estimator has also been proposed by Bellégo, Benatia and Doret-Bernardet (forthc.), as a way to estimate long-run treatment effects with an imbalanced panel. The chained DID estimator is computed by the `did_stepwise` Stata (Harmon, 2024) and `cdid` R (Benatia, Bellégo, Cuerrier and Doret-Bernadet, 2025) packages.

Looking at Figure 6.3, is it the case that one heterogeneity-robust estimation method leads to sizeably more precise estimates than the others?

Application to the UDL example. The confidence interval of the effect of having been exposed to a UDL for one year is much tighter in the bottom-right panel of Figure 6.3 than in all other panels: for that effect, the estimator proposed by Borusyak et al. (2024) does lead to a large precision gain. However, the opposite can hold when one considers other effects. For instance, the confidence interval of the effect of having been exposed to a UDL for three years is more than 50% larger per `did_imputation` than per `csdid`. Accordingly, the estimators proposed by Borusyak et al. (2024) do not always lead to precision gains, relative to those proposed by Sun and Abraham (2021) or Callaway and Sant’Anna (2021).

Precision loss with respect to TWFE estimators. TWFE ES estimators of the event-study effects ATT_ℓ are BLUE if $\text{TE}_{g,\ell}^r$ is constant across g and the errors $\varepsilon_{g,t}$ in (6.12) are homoscedastic and i.i.d., both across g and t . As discussed earlier in this chapter, the forbidden comparisons leveraged by those estimators can lead to a bias if effects are heterogeneous, but can improve precision if effects are homogeneous and errors are homoscedastic and i.i.d.. Then, there might be a bias-variance trade-off between TWFE and heterogeneity-robust estimators. Unfortunately, we do not yet have a meta-analysis comparing the variance of TWFE and heterogeneity-robust estimators in a large number of binary and staggered designs. Chiu

et al. (2023) compare TWFE and heterogeneity-robust estimators in 49 political science articles with a binary treatment, 11 of which have a binary and staggered design, and 38 of which have a binary and non-absorbing design. Their Figure 3 reports the distribution of the ratio of estimated standard errors of the imputation estimators of Liu et al. (2024) and of TWFE estimators across the 49 articles. It shows that the median ratio is equal to 1.1: in the median application, using an heterogeneity-robust estimator leads to moderate loss of precision, equivalent to a 10% sample-size reduction. Of course, there is heterogeneity across articles: in some of them, heterogeneity-robust estimators are more precise than TWFE estimators, but in others their standard errors are up to three times larger than that of TWFE estimators.

6.3.4.2 Confidence intervals coverage

Egerod and Hollenbach (2024) compute, in simulations, the finite-sample coverage rate of the confidence intervals (CIs) computed by `eventstudyinteract`, `csdid`, `did_multiplegt_dyn`, and `did_imputation`.

Simulation design. The simulations are based on a real data set with the 50 US states, with three cohorts of treated states (early, middle, and late) and some never-treated states, varying the number of states per cohort and the size of the treatment effect. The design is not fixed: in each draw, different states are randomly assigned to the early, middle, and late treated cohorts. Instead, the asymptotic results underlying the CIs computed by `did_multiplegt_dyn` and `did_imputation` are conditional on the design. Thus, the simulations evaluate the performance of those CIs outside of the framework where they have proven guarantees.⁸

Results. Figure 2 in Egerod and Hollenbach (2024) shows that when each cohort has only two states, thus meaning that there are only six treated states in total, all CIs tend to undercover, to different degrees. Specifically, when the treatment effect is low, 95% CIs respectively have effective coverage rates of slightly less than 80% with `eventstudyinteract` and `did_imputation`, of around 85% with `did_multiplegt_dyn`, and of slightly less than 90% with `csdid`. Increasing the

⁸In their Table 1, de Chaisemartin, Ciccia, D'Haultfœuille, Knau, Malézieux and Sow (2024) find that the CI of `did_multiplegt_dyn` has close to nominal coverage in simulations based on a real data set where the design is fixed.

treatment effect worsens the coverage of the CIs of `eventstudyinteract` and `did_imputation`. For `did_imputation`, this could reflect the fact that this CI is valid conditional on the design, not unconditionally, and the design contributes more to the estimator's variance when the treatment effect is large. Increasing the treatment effect does not affect the coverage of the CIs of `did_multiplegt_dyn`, and improves the coverage of the CIs of `csdid`, which actually reaches nominal coverage when the treatment effect is very large. When each cohort has four states, thus meaning that there are twelve treated states in total, all CIs have close to nominal coverage when the treatment effect is low, and the coverage of the CI of `did_imputation` (and to a lesser extent `eventstudyinteract` and `did_multiplegt_dyn`) deteriorates slightly when the treatment effect increases. When each cohort has six states, thus meaning that there are twenty-four treated states in total, all CIs have close to nominal coverage, irrespective of the magnitude of the treatment effect.

Recommendations. We recommend, as in Section 3.3.2, that researchers using those confidence intervals with less than 40 treated groups or with less than 40 control groups perform simulations tailored to their data to assess their coverage rate.⁹ Unfortunately, inference methods with few treated and control groups that we discussed when we studied the classical design have not been extended yet to designs with variation in treatment timing, though we conjecture that this extension should be feasible. For now, when simulations show that the aforementioned asymptotically-valid confidence intervals are unreliable, confidence intervals based on the bootstrap may have better coverage (Weiss, 2024). These issues may be especially relevant for the estimation of the cohort-specific event-study effects $\text{TE}_{c,\ell}^r$. When cohorts are too small to reliably draw inference on $\text{TE}_{c,\ell}^r$, it is more reasonable to focus on aggregated effects, such as ATT_ℓ .

⁹When estimating $\text{TE}_{c,\ell}^r$, the number of treated groups is just G_c , the number of groups in cohort c , while the number of control groups is just the number of never-treated or not-yet-treated groups. When estimating ATT_ℓ , the number of treated groups is the sum of G_c across the cohorts for which $\text{TE}_{c,\ell}^r$ can be estimated. The number of control groups is the number of never-treated groups, or the average number of not-yet-treated groups across the estimators of $\text{TE}_{c,\ell}^r$.

6.3.4.3 Bias

Sensitivity to violations of the no-anticipation and parallel-trends assumptions. In designs without variation in treatment timing, we have seen that the estimators of Borusyak et al. (2024) are more biased than those of Callaway and Sant’Anna (2021) and Sun and Abraham (2021) if Assumption PT is violated with differential trends that widen over time, and less biased if Assumption NA is violated due to anticipation effects arising a few periods before the treatment onset. We have also argued that violations of Assumptions PT and NA may not be equally problematic, as estimators can often be immunized against anticipation effects. The estimators of Borusyak et al. (2024) and those of Callaway and Sant’Anna (2021) and Sun and Abraham (2021) probably still have the same pros and cons in designs with variation in treatment timing, though the fact that the estimators of Borusyak et al. (2024) do not have a simple closed-form expression in those designs makes it hard to ascertain. Simulations that would compare the bias of those estimators under violations of Assumptions PT and NA in binary-and-staggered designs would be useful.

Sensitivity analysis under bounded differential trends. As discussed above, one can construct placebo estimators that closely mimick the actual treatment effect estimators of Callaway and Sant’Anna (2021) and Sun and Abraham (2021), by comparing the outcome evolutions of (almost) the same groups over the same number of periods, before the treatment onset. This makes those estimators amenable to the estimation approach under bounded differential trends proposed by Rambachan and Roth (2023): this approach requires that placebos be informative of the actual estimators’ bias, an easily-rationalizable assumption when the placebos and actual estimators are constructed symmetrically. Building a placebo that would similarly mimick the estimator proposed by Borusyak et al. (2024) is not feasible, because that estimator leverages all pre-treatment periods to estimate the treatment’s effect.

6.3.4.4 Software implementation

In Table 6.1 below, we compare `eventstudyinteract`, `csdid`, `did_multiplegt_dyn`, and `did_imputation` on a number of dimensions. Panel A shows that while all commands are applicable with an absorbing and binary treatment, only `did_multiplegt_dyn` can be used with a non-absorbing

and/or non-binary treatment. Panel B shows that `did_imputation` is the fastest of the four commands, while `csdid` is the slowest. On a relatively large dataset with 5100 groups and 33 periods, constructed by duplicating 100 times each state in the dataset of Wolfers (2006), the run time of `csdid` is almost 6 times larger than that of `did_imputation`, that of `eventstudyinteract` is 1.71 times larger, and that of `did_multiplegt_dyn` is 1.40 times larger. Panel C shows that all commands can be used to estimate the ATT, the event-study effects ATT_ℓ , and cohort-specific event-study effects. `csdid` and `did_imputation` can be used to obtain other, potentially more disaggregated effects. The estimation of several treatment effects could lead to a multiple-hypothesis-testing problem: to address it, `csdid` produces jointly valid confidence intervals, and `did_multiplegt_dyn` produces a joint test that all event-study effects are zero. Panel D shows that `eventstudyinteract` only uses never treated (or the last treated) as controls, `did_imputation` only uses not-yet treated, and `csdid` and `did_multiplegt_dyn` can use both groups as controls depending on the options specified. Panel E shows that `eventstudyinteract` and `did_multiplegt_dyn` produce long-difference pre-trend estimators by default, `csdid` produces long-difference pre-trends if the `long` or `long2` options are specified, and `did_imputation` does not produce long-difference pre-trends. `csdid` produces a joint test that cohort-specific event-study pre-trends are all equal to zero. `did_multiplegt_dyn` and `did_imputation` produce a joint test that event-study pre-trends are all equal to zero. Panel F shows that `eventstudyinteract` does not compute heterogeneity-robust estimators that control for covariates: the estimators with covariates computed by that command assume constant treatment effects. All other commands can control for covariates linearly. Only `csdid` can control for any type of time-invariant covariates non-parametrically, while `did_multiplegt_dyn` and `did_imputation` can only do so for discrete covariates. On the other hand, `csdid` cannot control for time-varying covariates or allow for group-specific linear trends, while `did_multiplegt_dyn` and `did_imputation` can. Finally, Panel G shows that `did_multiplegt_dyn` and `did_imputation` have built-in options to ensure that all event-study effects apply to the same groups, and to investigate heterogeneous treatment effects along some covariates.

Table 6.1: Comparing heterogeneity-robust DID estimators for binary-and-staggered designs, and their Stata implementation.

	eventstudyinteract	csdid	did_multiplegt_dyn	did_imputation
Panel A: Applicability				
Absorbing and binary D	Yes	Yes	Yes	Yes
Non-absorbing and/or non-binary D	No	No	Yes	No
Panel B: Run Time, UDL example				
Original data: G=51,T=33	1.45 sec	10.12 sec	3.58 sec	0.86 sec
Duplicated data: G=5100,T=33	100.56 sec	334.79 sec	81.88 sec	58.69 sec
Panel C: Effects estimated				
ATT	Yes	Yes	Yes	Yes
Event-study (ES) effects ATT_ℓ	Yes	Yes	Yes	Yes
Cohort-specific ES effects	Yes	Yes	With <code>if</code> condition	Yes
Effects by calendar time period	No	Yes	No	Yes
Multiple hypothesis	Yes	Yes	Yes	No
Panel D: Effects estimators				
Control group: never treated	Yes	Yes	Yes	No
Control group: not-yet treated	No	Yes	Yes	Yes
Baseline period	$F_g - 1$	$F_g - 1$	$F_g - 1$	$1, 2, \dots, F_g - 1$
Panel E: Pre-trend Estimators				
Long-difference?	Yes	<code>long</code> option	Yes	No
Joint-test that ES pre-trends=0	No	No	Yes	Yes
Panel F: Heterogeneity-robust estimators, controlling for covariates				
Linearly	No	Yes	Yes	Yes
Non-param, discrete time-invariant X s	No	Yes	Yes	Yes
Non-param, continuous time-invariant X s	No	Yes	No	No
Time-varying X s	No	No	Yes	Yes
Group-specific linear trends	No	No	Yes	Yes
Panel G: Built-in options				
No compositional changes across ES effects	No	No	<code>same_switchers</code>	<code>hbalance</code>
Effects' heterogeneity along covariates	No	No	<code>predict_het, by</code>	<code>project, hetby</code>

6.4 Estimating heterogeneous treatment effects

6.4.1 Estimating the correlation between treatment effects and some covariates

Assume that one wants to assess if the group-specific effects of ℓ periods of exposure to treatment $\text{TE}_{g,\ell}^r$ are correlated with a $K \times 1$ vector of time-invariant covariates X_g , whose first coordinate is a constant. In this section, we propose a generalization of the method described in Section 3.6.1 to designs with variation in treatment timing.

Target parameter. Let $\beta_{\ell,X}$ be the coefficient on X_g , in an infeasible regression of $\text{TE}_{g,\ell}^r$ on X_g and indicators for all possible treatment cohorts $(1\{F_g = c\})_{c \in \mathcal{C}}$, in the sample of groups such that $F_g - 1 + \ell \leq \underline{T}$. $X_g^T \beta_{\ell,X}$ is the part of the best linear predictor of $\text{TE}_{g,\ell}^r$ given $(X_g, (1\{F_g = c\})_{c \in \mathcal{C}})$ associated to X_g . Without variation in treatment timing, $\beta_{\ell,X}$ reduces to the coefficient we have considered in Section 3.6.1, which is why we recycle notation. If $K = 2$ and $X_{g,2}$ is binary, it follows from Angrist (1998) that $\beta_{\ell,X,2}$ is a weighted average, across adoption cohorts, of comparisons of the average of $\text{TE}_{g,\ell}^r$ across groups with $X_g = 1$ and $X_g = 0$ belonging to the same cohort, where cohorts in which the variance of X_g is the largest receive more weight. Still, why not consider instead $\tilde{\beta}_{\ell,X}$, the coefficient in a regression of $\text{TE}_{g,\ell}^r$ on X_g , without controlling for the cohort indicators? If $K = 2$ and $X_{g,2}$ is binary, $\tilde{\beta}_{\ell,X,2}$ is just the difference between the average of $\text{TE}_{g,\ell}^r$ across groups with $X_g = 1$ and $X_g = 0$, which may be a more natural target than $\beta_{\ell,X,2}$. Our (not very good) reason to focus on $\beta_{\ell,X}$ rather than $\tilde{\beta}_{\ell,X}$ is that the former is easier to estimate. To estimate $\beta_{\ell,X}$, we show below that one can use a simple one-step OLS estimator. To estimate $\tilde{\beta}_{\ell,X}$, one could regress $\widehat{\text{TE}}_{g,\ell}^r$ on X_g , but this yields a two-step regression estimator, and we are not aware of a Stata or R command that estimates its asymptotic variance.

Estimator. To estimate $\beta_{\ell,X}$, de Chaisemartin and D'Haultfœuille (2025) propose to use $\widehat{\beta}_{\ell,X}$, the coefficients on X_g in OLS regressions of $Y_{g,F_g-1+\ell} - Y_{g,F_g-1}$ on $(X_g, (1\{F_g = c\})_{c \in \mathcal{C}})$. Let $\beta_{\ell,X}^{\Delta Y(\mathbf{0})}$ be the coefficient in a regression of $E[Y_{g,F_g-1+\ell}(\mathbf{0}_{F_g-1+\ell}) - Y_{g,F_g-1}(\mathbf{0}_{F_g-1})]$ on $(X_g, (1\{F_g = c\})_{c \in \mathcal{C}})$, in the sample of groups such that $F_g - 1 + \ell \leq \underline{T}$. If the covariates are non stochastic or conditioned upon, and if for all $k \in \{2, \dots, K\}$ $\beta_{k,\ell,X}^{\Delta Y(\mathbf{0})} = 0$, meaning that the untreated outcome

evolutions of groups reaching ℓ periods of exposure to treatment before \underline{T} are uncorrelated with the non-constant variables in X_g , then de Chaisemartin and D'Haultfœuille (2025) show that $\hat{\beta}_{\ell,X,k}$ is unbiased for $\beta_{\ell,X,k}$. $\beta_{k,\ell,X}^{\Delta Y(\mathbf{0})} = 0$ is placebo testable, by regressing $Y_{g,F_g-1-\ell} - Y_{g,F_g-1}$ on $(X_g, (1\{F_g = c\})_{c \in \mathcal{C}})$, in the sample of groups such that $F_g - 1 + \ell \leq \underline{T}, F_g - 1 - \ell \geq 2$, and testing if the coefficient on X_g is equal to zero. The `did_multiplegt_dyn` Stata and R packages have a `predict_het` option that can be used to compute $\hat{\beta}_{\ell,X}$ and placebo-test $\beta_{k,\ell,X}^{\Delta Y(\mathbf{0})} = 0$.

Comparison with current practice: fully-interacted TWFE regressions. To estimate heterogeneous treatment effects alongside covariates taking a small number of values, applied researchers sometimes fully interact the standard TWFE regression with indicators for all the values the covariates can take, hereafter referred to as strata. As this is equivalent to estimating the TWFE regression separately in each stratum, in Section 3.6.1 we saw that without variation in treatment timing those regressions yield unbiased estimators of the ATT in each stratum. With variation in treatment timing, under Assumptions NA and PT each stratum-specific TWFE regression estimates a weighted sum of effects among the stratum's treated groups, with weights that sum to one but may be negative. If some weights are negative, one could have, say, that all groups in stratum 1 have a larger treatment effect than all groups in stratum 2, and yet the expectation of the TWFE coefficient in stratum 1 is smaller than that in stratum 2.¹⁰

Comparison with current practice: non-fully-interacted TWFE regressions. Sometimes, researchers do not fully interact the standard TWFE regression with the covariates. For instance, with a binary covariate $X_{g,2}$, one may regress $Y_{g,t}$ on group and period FEs, $D_{g,t}X_{g,2}$, and $(1 - D_{g,t})X_{g,2}$, without interacting the period FEs with $X_{g,2}$. Then, de Chaisemartin and D'Haultfœuille (2023a) show that on top of not being robust to heterogeneous treatment effects within stratum, as was already the case in the fully interacted TWFE regression, the coefficient on $D_{g,t}X_{g,2}$ in the non-fully interacted TWFE regression may also be contaminated by the treat-

¹⁰Therefore, those regressions cannot be used to estimate the $\beta_{\ell,X}$ coefficients introduced above, or an average across ℓ of those coefficients. For instance, if $K = 2$ and $X_{g,2}$ is binary, the difference between the TWFE coefficients in the two strata estimates the difference between weighted sums, across g and ℓ , of $\text{TE}_{g,\ell}^r$ across groups with $X_g = 1$ and $X_g = 0$, while the $\beta_{\ell,X,2}$ coefficients are weighted averages, across adoption cohorts, of the difference between the average of $\text{TE}_{g,\ell}^r$ across groups with $X_g = 1$ and $X_g = 0$.

ment effects of groups with $X_{g,2} = 0$. Specifically, that coefficient estimates a weighted sum of effects among treated groups with $X_{g,2} = 1$, with weights summing to one, plus a weighted sum of effects among treated groups with $X_{g,2} = 0$, with weights summing to zero. The same holds for the coefficient on $D_{g,t}(1 - X_{g,2})$: it may be contaminated by the treatment effects of treated groups with $X_{g,2} = 1$.

Estimating the conditional ATT function? Instead of estimating the best linear predictor of $\text{TE}_{g,\ell}^r$, one may be interested in estimating the function mapping groups' covariates to the average of $\text{TE}_{g,\ell}^r$. Hatamyar, Kreif, Rocha and Huber (2023) combine insights from Callaway and Sant'Anna (2021) and Lu et al. (2019) to form an estimator of that function under a conditional parallel-trends assumption, in designs with variation in treatment timing. However, their estimators are not implemented yet in a Stata or R command.

6.4.2 Estimating the variance and the distribution of group-specific effects.*

Estimating the variance of group-specific effects. To estimate the variance of group-specific treatment effects in binary-and-staggered designs, we can extend the method described in Section 3.6. Let

$$\hat{\sigma}_{\ell,c}^2 = \frac{1}{G_c - 1} \sum_{g:F_g=c} \left(Y_{g,c-1+\ell} - Y_{g,c-1} - \frac{1}{G_c - 1} \sum_{g':F_{g'}=c} Y_{g',c-1+\ell} - Y_{g',c-1} \right)^2$$

denote the sample variance of $Y_{g,c-1+\ell} - Y_{g,c-1}$ in cohort c , let $G_{>t}$ denote the number of groups not yet treated at t , and let

$$\hat{\sigma}_{\ell,nyt,c-1+\ell}^2 = \frac{1}{G_{>c-1+\ell} - 1} \sum_{g:F_g>c-1+\ell} \left(Y_{g,c-1+\ell} - Y_{g,c-1} - (\bar{Y}_{nyt,c-1+\ell,c-1+\ell} - \bar{Y}_{nyt,c-1+\ell,c-1}) \right)^2$$

denote the sample variance of $Y_{g,c-1+\ell} - Y_{g,c-1}$ among groups not-yet-treated at $c-1+\ell$. Under similar assumptions as in Section 3.6, one can show that

$$\frac{G_c - 1}{G_c} (\hat{\sigma}_{\ell,c}^2 - \hat{\sigma}_{\ell,nyt,c-1+\ell}^2)$$

is unbiased for the variance of the group-specific effects of ℓ periods of exposure to treatment in cohort c . Then, one can aggregate those estimators across cohorts to estimate the variance of

the group-specific effects of ℓ periods of exposure to treatment. We are not aware of a Stata or R package computing those variance estimators in designs with variation in treatment timing.

Estimating the distribution of group-specific effects. Earlier, we introduced unbiased estimators $\widehat{\text{TE}}_{g,t}$ of the effects $\text{TE}_{g,t}$. Using the distribution of the estimators $\widehat{\text{TE}}_{g,t}$ to estimate the distribution of those effects would be misleading: one first needs to apply a deconvolution to these noisy measures. However, deconvolution techniques rely on strong assumptions, and yield estimators that often converge at a slow rate.

6.5 Non-linear DID

6.5.1 Limited dependent variables

A parallel-trends assumption for limited dependent variables. In this section, we replace Assumption PT by the following condition: for all $c \in \mathcal{C}$ and $t \geq 2$,

$$L^{-1} \left(E \left[\bar{Y}_{c,t}(\mathbf{0}_t) \right] \right) - L^{-1} \left(E \left[\bar{Y}_{c,t-1}(\mathbf{0}_{t-1}) \right] \right) \quad (6.27)$$

does not depend on c , for a known, strictly increasing function L taking values in $[0, 1]$. (6.27) is a parallel-trends assumption on a monotone transformation of the average untreated outcome. It generalizes (3.34) to designs with variation in treatment timing. Let us assume for now that there exists never-treated groups. Then,

$$\begin{aligned} & E \left[\bar{Y}_{c,t}(\mathbf{0}_t) \right] \\ &= L \left(L^{-1} \left(E \left[\bar{Y}_{c,t}(\mathbf{0}_t) \right] \right) \right) \\ &= L \left(L^{-1} \left(E \left[\bar{Y}_{c,c-1}(\mathbf{0}_{c-1}) \right] \right) + L^{-1} \left(E \left[\bar{Y}_{c,t}(\mathbf{0}_t) \right] \right) - L^{-1} \left(E \left[\bar{Y}_{c,c-1}(\mathbf{0}_{c-1}) \right] \right) \right) \\ &= L \left(L^{-1} \left(E \left[\bar{Y}_{c,c-1}(\mathbf{0}_{c-1}) \right] \right) + L^{-1} \left(E \left[\bar{Y}_{n,t}(\mathbf{0}_t) \right] \right) - L^{-1} \left(E \left[\bar{Y}_{n,c-1}(\mathbf{0}_{c-1}) \right] \right) \right) \\ &= L \left(L^{-1} \left(E \left[\bar{Y}_{c,c-1} \right] \right) + L^{-1} \left(E \left[\bar{Y}_{n,t} \right] \right) - L^{-1} \left(E \left[\bar{Y}_{n,c-1} \right] \right) \right) \end{aligned} \quad (6.28)$$

where the third equality follows from (6.27). Then, for $c \in \mathcal{C}$ and $\ell \in \{1, \dots, T - (c - 1)\}$ a natural estimator of $\text{TE}_{c,\ell}^r$ is

$$\widehat{\text{TE}}_{c,\ell}^{\text{ldv}} := \bar{Y}_{c,c-1+\ell} - L \left(L^{-1} \left(\bar{Y}_{c,c-1} \right) + L^{-1} \left(\bar{Y}_{n,c-1+\ell} \right) - L^{-1} \left(\bar{Y}_{n,c-1} \right) \right).$$

Heterogeneity-robust probit or logit DID estimators. When $Y_{g,t}$ is binary, it follows from Theorem 12 in the chapter's starred appendix that

$$\widehat{\text{TE}}_{c,\ell}^{\text{ldv}} = L\left(\widehat{\alpha} + \widehat{\alpha}_c + \widehat{\gamma}_{c-1+\ell} + \widehat{\beta}_{c,\ell}\right) - L\left(\widehat{\alpha} + \widehat{\alpha}_c + \widehat{\gamma}_{c-1+\ell}\right), \quad (6.29)$$

where $(\widehat{\alpha}, (\widehat{\alpha}_c)_{c \in \mathcal{C}}, (\widehat{\gamma}_t)_{t \in \{2, \dots, T\}}, (\widehat{\beta}_{c,\ell})_{c \in \mathcal{C}, \ell \in \{-c+2, \dots, -1, 1, T-(c-1)\}})$ is the maximum likelihood estimator of a binary choice model assuming

$$\begin{aligned} P(Y_{g,t} = 1) = & L\left(\alpha + \sum_{c \in \mathcal{C}} \alpha_c 1\{g \in c\} + \sum_{t'=2}^T \gamma_{t'} 1\{t = t'\}\right. \\ & \left. + \sum_{c \in \mathcal{C}} \sum_{\ell=-c+2, \ell \neq 0}^{T-(c-1)} \beta_{c,\ell} 1\{F_g = c, t = c-1+\ell\}\right). \end{aligned} \quad (6.30)$$

If L is the logistic (resp. normal) cdf, for instance, those coefficients can be obtained by a logit (resp. probit) version of the linear regression proposed by Sun and Abraham (2021). For $\ell < -1$, the coefficients $\widehat{\beta}_{c,\ell}$ can be used to test (6.28).

Heterogeneity-robust logit imputation estimators. Instead of the estimator in (6.29), Wooldridge (2023) proposes a similar estimator, based on a logit regression of the outcome on period and adoption-cohort FEs and a full set of interactions between cohort and time-since-adoption FEs, estimated in the full sample. This specification is the same as (6.30), except that in the double summation, ℓ only runs from 1 to $T - (c - 1)$ and does not take negative values. Wooldridge (2023) shows that his estimators of $\text{TE}_{c,\ell}^r$ are numerically equivalent to those one would obtain from a logit-based imputation strategy, where one first estimates a logit regression of the outcome on period and adoption-cohort FEs in the untreated sample, and then one compares the average outcome of cohort c at period $c - 1 + \ell$ to its predicted untreated outcome according to the regression in the untreated sample. One can show that the logit-based imputation estimators only rely on Assumption NA and on (6.27), so the logit estimators in Wooldridge (2023) also only rely on those two assumptions. Note that a probit with the specification proposed by Wooldridge (2023) is no longer numerically equivalent to the corresponding probit-based imputation estimator. Then, when this specification is used, it is preferable to use a logit. With respect to the estimators in (6.29), an advantage of those of Wooldridge (2023) is that they use not-yet-treated as controls, so they can be used even if there are no never-treated, and they may have a lower variance. An advantage of the estimators in

(6.29) is that the corresponding regression simultaneously computes event-study and pre-trend estimators.

Heterogeneity-robust Poisson estimators. The heterogeneity-robust estimators in (6.29) and of Wooldridge (2023) generalize to non-negative dependent variables, using Poisson instead of logit or probit regressions.

Incidental parameters? In non-linear models, including cohort rather than group FEs is important: including group FEs could lead to an incidental parameter problem (Neyman and Scott, 1948), which could severely bias the estimator if T , the number of time periods of the panel, is low. If the number of groups is large relative to the number of cohorts, the estimators described above will not be subject to that problem.

6.5.1.1 Application: the effect of regional trade agreements on trade

TWFE regressions have often been used to estimate the effect of regional trade agreements on trade. The effect of regional trade agreements (RTA) on trade is a question that has been extensively studied by trade economists. The outcome of interest is $Y_{i,j,t}$, the trade flow from country i to j at t , and the treatment $D_{i,j,t}$ is an indicator for whether there is an RTA between i and j at t . With respect to our previous notation, groups g now correspond to origin-destination country pairs (i, j) . Researchers sometimes estimate TWFE regressions of $\ln(Y_{i,j,t})$ on origin-destination-pair FEs, year FEs, and $D_{i,j,t}$. But following the influential work of Silva and Tenreyro (2006), they more often estimate a TWFE Poisson regression of $Y_{i,j,t}$ on origin-destination-pair FEs, year FEs, and $D_{i,j,t}$, to account for heteroscedasticity, and also because $Y_{i,j,t}$ can be equal to zero when there is no trade from i to j at t . Motivated by gravity models of international trade, researchers also often include control variables in their specification, such as importer-year and exporter-year FEs, and year FEs interacted with an indicator equal to one when $i = j$, namely when $Y_{i,j,t}$ actually represents a domestic trade flow. The RTA treatment is binary and staggered: there is variation in the timing when different country pairs start having an RTA. Thus, TWFE Poisson regressions of the effect of RTAs on trade could be biased, if the effect of RTAs is heterogeneous across country pairs or over time.

Heterogeneity-robust TWFE estimators are twice larger than standard TWFE estimators. Nagengast and Yotov (2025) investigate whether findings from TWFE Poisson regressions are robust to allowing for heterogeneous effects. They use the Structural Gravity Database of the World Trade Organization to measure trade flows between countries, and the Dynamic Gravity Dataset of the US International Trade Commission to measure RTAs between countries, and their dates of adoption. They start by estimating the standard TWFE Poisson regression in the trade literature. Then, they estimate an heterogeneity-robust version of that regression, where instead of having just one treatment indicator $D_{i,j,t}$, the regression has cohort \times time-since-RTA-adoption FEs, following Wooldridge (2023). Then, they average the coefficients on the cohort \times time-since-RTA-adoption FEs, to obtain an estimator comparable to the coefficient on $D_{i,j,t}$ in the standard regression. They also estimate the standard TWFE OLS regression, and an heterogeneity-robust TWFE OLS regression following Wooldridge (2021). The first (resp. second, third, fourth) line of Table 6.2 below replicates Column (1) of Table 3 (resp. Column (2) of Table 3, Column (1) of Table 7, Column (2) of Table 7) of their paper. Allowing for heterogeneous effects doubles the estimated effect of RTAs on trade, both in the Poisson and in the OLS regression, and the differences between the estimators is statistically significant at all conventional levels. On the other hand, using a Poisson or an OLS regression does not change results much.

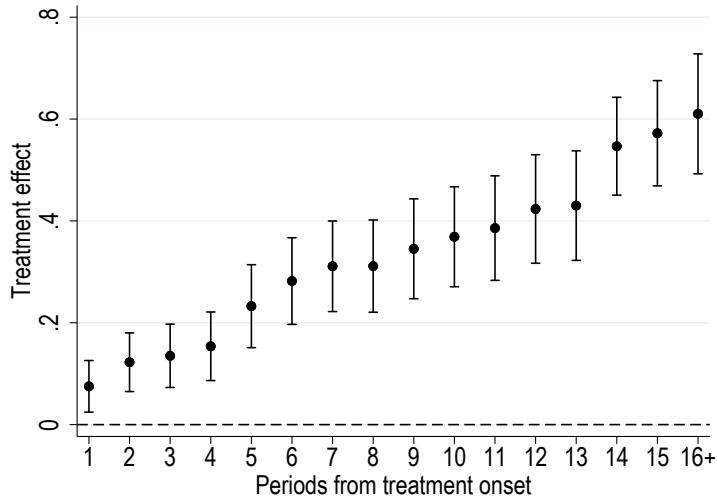
Table 6.2: The Effect of Regional Trade Agreements on Trade

Effect of RTAs	
TWFE Poisson	0.166
	(0.050)
Heterogeneity-robust TWFE Poisson	0.381
	(0.041)
TWFE OLS	0.172
	(0.037)
Heterogeneity-robust TWFE OLS	0.347
	(0.051)

Notes: The table shows estimates of the effect of RTAs on trade, taken from Nagengast and Yotov (2025). The TWFE Poisson estimate is from their Table 3 Column (1), the heterogeneity-robust TWFE Poisson estimate is from their Table 3 Column (2), the TWFE OLS estimate is from their Table 7 Column (1), and the heterogeneity-robust TWFE OLS estimate is from their Table 7 Column (2). Standard errors are shown below the estimates, between parentheses.

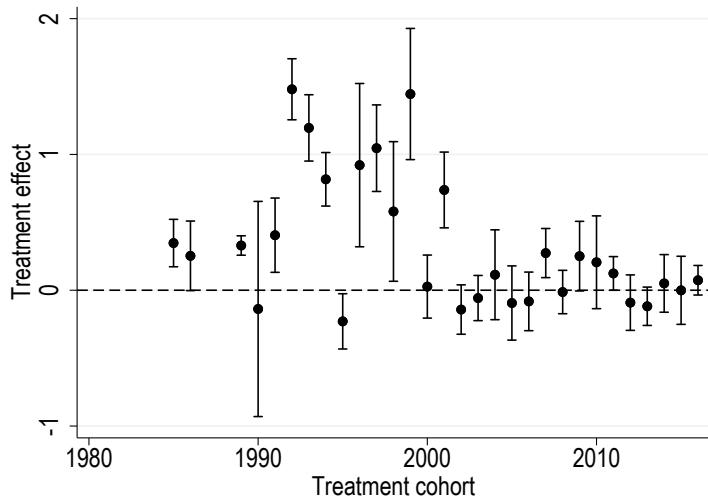
Standard TWFE regressions are downward biased, because RTAs' effects increase with length of exposure, and because RTAs' effects are larger for country pairs that adopt an RTA early. In Panel (a) of their Figure 3, reproduced in Figure 6.4 below, Nagengast and Yotov (2025) show heterogeneity-robust estimates of ATT_ℓ , the average effect of having had an RTA for ℓ years. Effects are increasing with length of exposure.

Figure 6.4: Estimates of RTA effects, by length of exposure.



In Panel (e) of their Figure 3, reproduced in Figure 6.5 below, Nagengast and Yotov (2025) show heterogeneity-robust estimates of RTA effects by adoption cohort. Effects are much larger for cohorts that adopt an RTA earlier. In fact RTAs do not increase trade flows between countries adopting an RTA after 2000.

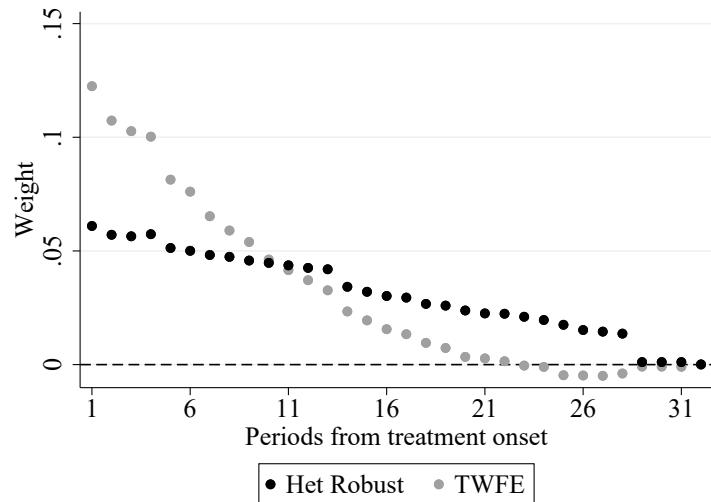
Figure 6.5: Estimates of RTA effects, by cohort of adoption.



Finally, the authors follow de Chaisemartin and D'Haultfœuille (2020), and compute the weights

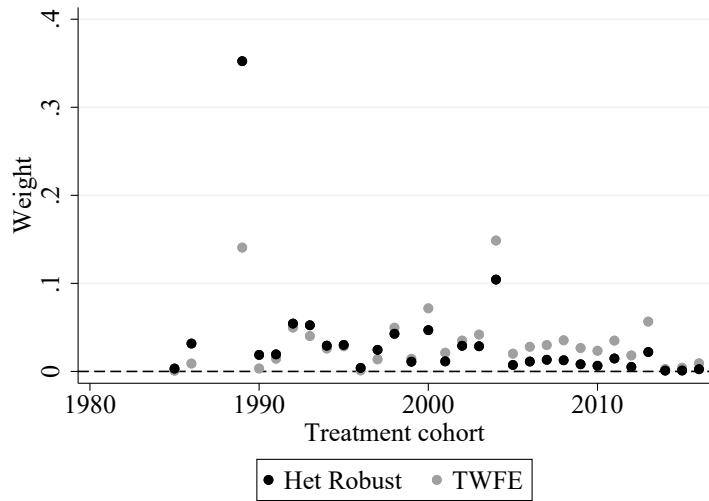
attached to the TWFE OLS regression.¹¹ Then, in Panel (a) (resp. (b)) of their Figure 4, reproduced in Figure 6.6 (resp. 6.7) below, they aggregate those weights by length of exposure to an RTA (resp. by cohort of adoption), and compare them to the weights attached to the heterogeneity-robust TWFE OLS estimator. Consistent with the results discussed in Section 6.2.1, the TWFE OLS estimator gives too much weight to short-run effects, and to the effects of late adopters. As short-run effects are smaller than long-run effects, and late adopters have smaller effects than early adopters, the TWFE OLS estimator is downward biased. Some weights attached to the TWFE OLS estimator are negative, but negative weights are small and sum to -0.028 only.

Figure 6.6: Weights of TWFE regression, by length of exposure



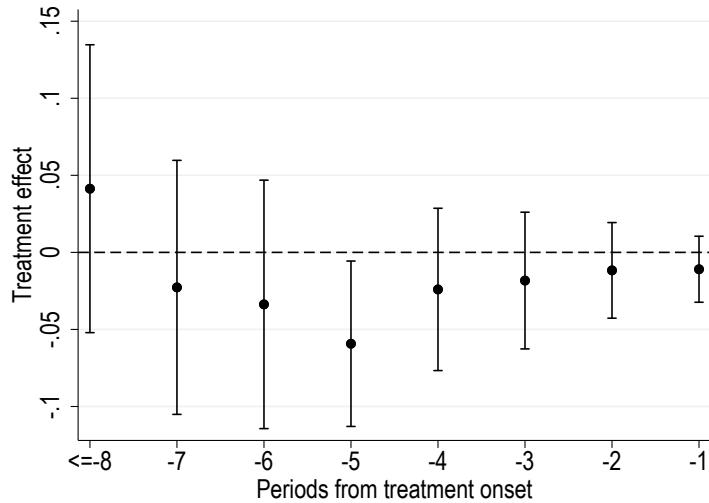
¹¹Unfortunately, a similar decomposition is not available for the TWFE Poisson regression.

Figure 6.7: Weights of TWFE regression, by cohort of adoption



Pre-trend tests suggest that the no-anticipation and parallel-trends assumptions are plausible in this application. In their Figure 2, reproduced in Figure 6.8 below, the authors conduct pre-trend tests, in the spirit of those proposed by Borusyak et al. (2024). Pre-trends estimates are much smaller than the event-study estimates in Figure 6.4, most of them are individually insignificant, and they are jointly insignificant (F-test p-value= 0.298). This suggests that the no-anticipation and parallel-trends assumptions are plausible in this application.

Figure 6.8: Pre-trend tests



6.5.2 Estimating quantile treatment effects*

Athey and Imbens (2006) extend their CIC estimator of QTEs to absorbing and binary treatments with more than two periods and variation in treatment timing. CIC estimators of QTEs at the time period when treated groups start receiving the treatment are computed by the `fuzzydid` Stata package. We are not aware of a Stata or R package that can be used to compute CIC estimators of QTEs at later periods.¹² To our knowledge, the QTE estimators of Kim and Wooldridge (2024) and the quantile-DID estimators have not been extended to designs with more than two periods and variation in treatment timing.

¹²The `cic` (Kraker, 2019) Stata and `qte` (Callaway, 2023) R package can only be used with two time periods.

6.6 Further topics*

6.6.1 You shall never use always-treated as the control group, except maybe if they have been treated for a long time

Sometimes, the only control group available to the researcher consists of always-treated groups: some groups are treated at period one, and the groups untreated at period one all become treated at the same date (see, e.g., Field, 2007). Then, if one wants to compute a DID, the only option is to compare the outcome evolutions of switchers and always-treated groups. In other cases, there is a variation in the timing at which switchers get treated, but a large proportion of groups are treated at period one (see, e.g., Ujhelyi, 2014). Then, not leveraging DIDs comparing switchers and always-treated may lead to a large loss of statistical precision. In this chapter, we have seen that such DIDs rely on Assumptions NA and PT, and on the assumption that, by the time when the switchers get treated, the treatment effect of the always-treated groups is constant over time. This assumption may be implausible if the always-treated started receiving the treatment a few periods before the switchers get treated. If, on the other hand, the always-treated groups started receiving the treatment a long time before the switchers get treated, it may be reasonable to assume that their treatment effect is no longer evolving by that time, and pre-trend tests can be useful to assess the plausibility of this assumption. All the Stata and R commands reviewed in this chapter can easily be tweaked to ensure that all or some always-treated groups are used as controls. For instance, with `did_multiplegt_dyn` one just needs to redefine the treatment variable of those groups as equal to zero throughout the panel.

6.6.2 Imbalanced panels and weighting

Adapting our general recommendation to designs with variation in treatment timing.

Our recommendation for imbalanced panels in Section 3.9.1 can easily be extended to designs with variation in treatment timing: then, to estimate the effect of ℓ periods of exposure to treatment, we recommend restricting the sample to groups observed at period $F_g - 1$, or at least at one period before F_g , and at $F_g - 1 + \ell$. Such sample restrictions are implemented by default

in the `did_imputation`, `did_multiplegt_dyn`, and `event_study_interact` packages, and are implemented when an option is specified in the `csdid` and `did` packages, see the help files for further details.

In an imbalanced panel with variation in treatment timing, there may be groups whose treatment adoption date is unknown. Assume that there are (g, t) cells for which $D_{g,t}$ is missing. Without variation in treatment timing, one has $D_{g,t} = D_g 1\{t \geq F\}$, so one can easily recover $D_{g,t}$, unless D_g is also missing for group g but that is usually not the case as D_g is typically a simple function of groups' observables. With a binary and staggered treatment, there may be cases where some values of $D_{g,t}$ remain missing, even after a thorough search. Then, we can leverage the monotonicity of groups' treatment with respect to time in binary-and-staggered designs to impute some missing treatments. If $D_{g,t}$ is unobserved but there is a $t' \leq t$ such that $D_{g,t'} = 1$, $D_{g,t} = 1$. Similarly, if $D_{g,t}$ is unobserved but there is a $t' \geq t$ such that $D_{g,t'} = 0$, $D_{g,t} = 0$. This leaves us with two cases where missing treatments cannot be imputed. The first case is when i) $D_{g,t}$ is unobserved, ii) $D_{g,t'} = 0$ for all $t' \leq t$ such that $D_{g,t'}$ is observed, and iii) $D_{g,t'} = 1$ for all $t' \geq t$ such that $D_{g,t'}$ is observed. Then, g 's treatment adoption date is unknown. One could let F_g be the lowest t' such that $D_{g,t'} = 1$, but this could lead to measurement error if actually g started receiving the treatment before F_g . For instance, one would mistakenly consider g 's effect at F_g as an effect of one period of exposure to treatment, while this is actually an effect of several periods of exposure. Instead, one can drop g from the estimation sample, or at least drop all observations of g after the last t' such that $D_{g,t'} = 0$. The latter is the convention used by the `did_multiplegt_dyn` Stata and R commands, as explained in Appendix A of the commands' companion paper (see de Chaisemartin, Ciccia, D'Haultfoeuille, Knau, Malézieux and Sow, 2024). Alternatively, the `eventstudyinteract`, `csdid`, `did`, `did_imputation` Stata and/or R commands let the user define F_g , the `cohort` variable used as an argument by all those commands. The second case where missing treatments cannot be imputed is when $D_{g,t}$ is unobserved, $D_{g,t'} = 0$ for all $t' \leq t$ such that $D_{g,t'}$ is observed, and $D_{g,t'}$ is unobserved for all $t' \geq t$. Then, g 's treatment sequence after the last date where its treatment is observed is unknown. In this case, `did_multiplegt_dyn` drops all observations of g after the last t' such that $D_{g,t'} = 0$.

Instances where the outcome is observed less frequently than the treatment. There are circumstances where the outcome is observed less frequently than the treatment. For instance, electoral outcomes are observed only during election years, while treatment may be observed every year. Then, one could restrict the estimation sample to years where the outcome is observed, and naively apply some of the event-study estimators reviewed in this chapter. However, doing so may yield hard to interpret event-study effects, that conflate effects of different exposure lengths. To see this, assume that elections take place every other year, on even years. In the sample restricted to even years, two groups g and g' such that g adopted treatment in year $2k$ while g' adopted treatment in year $2k - 1$ will be considered as having both adopted in $2k$. Then, estimators of ATT_1 will actually average effects of one year of exposure for groups that adopted during an even year, and of two years of exposure for groups that adopted during an odd year. A simple solution amounts to estimating effects separately for groups adopting during an even year, and for groups adopting during an odd year. Then, one may combine in a single event-study graph effects of odds numbers of years of exposure for groups adopting during an even year, and effects of even numbers of years of exposure for groups adopting during an odd year. A tutorial showing how this can be done with the `did_multiplegt_dyn` Stata and R commands is available [here](#).

6.6.3 Weighting

With variation in treatment timing, weighting does not lead to further issues beyond those already mentioned in Section 3.9.2 of Chapter 3.

6.6.4 Accounting for and estimating spillover effects

In applications with variation in treatment timing and where the treatment effect might spill over onto untreated groups geographically close to a treated group, Butts (2021c) proposes a method to estimate the average total effect of the treatment across all treated groups, and the indirect effect of the treatment across all affected untreated groups. Assume that the researcher is ready to assume that groups located more than x kilometers away from a treated group cannot

be affected by its treatment. The choice of x should be based on context-specific knowledge, and researchers will typically present sensitivity analyses, where they show that results are robust to changes in x . Then, let $A_{g,t}$ be an indicator equal to one if group g is untreated at t but indirectly affected because a group located less than x kilometers away from g is treated at t . Then, Butts (2021c) proposes an extension of the imputation estimator of Borusyak et al. (2024), Gardner (2021), and Liu et al. (2024). First, one fits a TWFE regression of the outcome on group and time FEs in the sample of untreated and unaffected (g, t) cells, with $D_{g,t} = 0$ and $A_{g,t} = 0$. Then, one uses that regression to predict the counterfactual untreated outcome of treated cells and of untreated but affected cells. Estimates of the total effect of treated cells are obtained by subtracting their counterfactual to their actual outcome. Similarly, estimates of the indirect effect of untreated but affected cells are obtained by subtracting their counterfactual to their actual outcome. Instead of imputation estimators, one can also use the DID estimators of Sun and Abraham (2021), Callaway and Sant'Anna (2021), or de Chaisemartin and D'Haultfœuille (2025). To estimate the total effect of the treatment, one just needs to compute those estimators in the subsample such that $A_{g,t} = 0$ or $D_{g,t} = 1$. To estimate the indirect effect of the treatment, one just needs to compute those estimators in the subsample such that $A_{g,t} = 0$ or $A_{g,t} = 1$. If one is interested in estimating the incremental effect of becoming treated after having already been indirectly affected by the treatment, one can follow the estimation procedure described later in Section 8.3.4.6, treating $A_{g,t}$ and $D_{g,t}$ as two different treatments.

6.7 Appendix*

We show here that the estimator $\widehat{\text{TE}}_{c,\ell}^{\text{ldv}}$ for limited dependent variables defined in Section 6.5.1 satisfies (6.29). Let $\widehat{\theta} = (\widehat{\alpha}, (\widehat{\alpha}_c)_{c \in \mathcal{C}}, (\widehat{\gamma}_t)_{t \in \{2, \dots, T\}}, (\widehat{\beta}_{c,\ell})_{c \in \mathcal{C}, \ell \in \{-c+2, \dots, -1, 1, T-(c-1)\}})$ denote the maximum likelihood estimator (MLE) of a binary choice model assuming

$$P(Y_{g,t} = 1) = L \left(\alpha + \sum_{c \in \mathcal{C}} \alpha_c 1\{g \in c\} + \sum_{t'=2}^T \gamma_{t'} 1\{t = t'\} + \sum_{c \in \mathcal{C}} \sum_{\ell=-(c-2), \ell \neq 0}^{T-(c-1)} \beta_{c,\ell} 1\{F_g = c, t = c-1+\ell\} \right).$$

Let us also recall that

$$\widehat{\text{TE}}_{c,\ell}^{\text{ldv}} = \bar{Y}_{c,c-1+\ell} - L \left(L^{-1}(\bar{Y}_{c,c-1}) + L^{-1}(\bar{Y}_{n,c-1+\ell}) - L^{-1}(\bar{Y}_{n,c-1}) \right). \quad (6.31)$$

Theorem 12 Assume that $\hat{\theta}$ exists. Then

$$\widehat{TE}_{c,\ell}^{ldv} = L(\hat{\alpha} + \hat{\alpha}_c + \hat{\gamma}_{c-1+\ell} + \hat{\beta}_{c,\ell}) - L(\hat{\alpha} + \hat{\alpha}_c + \hat{\gamma}_{c-1+\ell}).$$

Proof: let C denote the cardinality of \mathcal{C} . First, observe that the model is saturated in cohorts and time: the regressors are not collinear and there are $1 + C + T - 1 + C(T - 1) = (C + 1)T$ coefficients, compared to $(C + 1)$ cohorts (C cohorts in \mathcal{C} plus the never-treated groups) and T periods. Moreover, the linear combination of coefficients associated to cohort $c \in \mathcal{C}$ and t is $\hat{\alpha} + \hat{\alpha}_c + \hat{\gamma}_t + \hat{\beta}_{c,t-c+1}$ (with the conventions that $\hat{\gamma}_1 = 0$ and $\hat{\beta}_{c,0} = 0$), while the linear combination of coefficients associated to cohort n and t is $\hat{\alpha} + \hat{\gamma}_t$. Then, by Lemma 1 below, we have, for all $(c, t) \in \mathcal{C} \times \{1, \dots, T\}$,

$$\begin{aligned}\bar{Y}_{c,t} &= L(\hat{\alpha} + \hat{\alpha}_c + \hat{\gamma}_t + \hat{\beta}_{c,t-c+1}), \\ \bar{Y}_{n,t} &= L(\hat{\alpha} + \hat{\gamma}_t).\end{aligned}\tag{6.32}$$

As a result,

$$\begin{aligned}L^{-1}(\bar{Y}_{c,c-1}) + L^{-1}(\bar{Y}_{n,c-1+\ell}) - L^{-1}(\bar{Y}_{n,c-1}) &= (\hat{\alpha} + \hat{\alpha}_c + \hat{\gamma}_{c-1}) + (\hat{\alpha} + \hat{\gamma}_{c-1+\ell}) - (\hat{\alpha} + \hat{\gamma}_{c-1}) \\ &= \hat{\alpha} + \hat{\alpha}_c + \hat{\gamma}_{c-1+\ell}.\end{aligned}$$

The result follows from this equation, (6.31) and (6.32) applied to $t = c - 1 + \ell$.

A result on saturated binary choice models. Consider the model $Y_i = \mathbb{1}\{X'_i\theta + \varepsilon_i \geq 0\}$ where ε_i is independent of X_i and the cdf of $-\varepsilon_i$ is F , assumed to be differentiable with $F'(x) > 0$ for all x . Let $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_K)'$ denote the MLE of θ , assuming that the data are i.i.d. Then, we have the following result, which is used in the proof of Theorem 12:

Lemma 1 If $\hat{\theta}$ exists and the model is saturated, namely $X_i = (\mathbb{1}\{\widetilde{X}_i = x_1\}, \dots, \mathbb{1}\{\widetilde{X}_i = x_K\})'$ where $\{x_1, \dots, x_K\}$ is the support of \widetilde{X}_i , then, for all $k = 1, \dots, K$, $F(\hat{\theta}_k) = \bar{Y}_k$, where

$$\bar{Y}_k := \frac{1}{\sum_{i=1}^n \mathbb{1}\{\widetilde{X}_i = x_k\}} \sum_{i:\widetilde{X}_i=x_k} Y_i.$$

Proof of Lemma 1: The log-likelihood of the binary choice model is

$$\sum_{i=1}^n Y_i \ln(F(X'_i\theta)) + (1 - Y_i) \ln(1 - F(X'_i\theta)).$$

Since $\hat{\theta}$ exists, it satisfies the first-order conditions, which can be written, after some manipulations,

$$\sum_{i=1}^n X_i \frac{F'(X'_i \hat{\theta})}{F(X'_i \hat{\theta})(1 - F(X'_i \hat{\theta}))} (Y_i - F(X'_i \hat{\theta})) = 0.$$

For component k of this vector, we obtain, since the model is saturated,

$$\frac{F'(\hat{\theta}_k)}{F(\hat{\theta}_k)(1 - F(\hat{\theta}_k))} \sum_{i:\tilde{X}_i=x_k} (Y_i - F(\hat{\theta}_k)) = 0.$$

The result follows.

Chapter 7

Designs with variation in treatment dose

Heterogeneous adoption designs. In most of this chapter, we assume that our data contains only two time periods, $T = 2$. We assume that treatment follows an “heterogeneous-adoption design” (HAD): groups are untreated at period one, some or all groups receive a strictly positive treatment dose at period two, but the treatment dose can vary across treated groups, with some groups receiving larger doses than others. The variability in the dose received by treated groups is the key difference between HADs and the classical designs reviewed in Chapter 3. The period-two treatment dose could be a discrete variable taking a small number of values, like 1, 2, and 3. But the period-two treatment could also be a continuously distributed variable, taking as many different values as there are treated groups.

I.i.d. groups. In this chapter, we replace Assumption IND, which requires that groups are independent, by the slightly stronger assumption that the groups are an independent and identically distributed (i.i.d.) sample, drawn from an infinite super-population of groups. Introducing an infinite super-population of groups is necessary, to allow for a potentially continuous distribution of the period-two treatment. As groups are assumed i.i.d., we drop the g subscript, except when we introduce estimators. As different samples of groups lead to different treatments and potential outcomes, expectations are taken with respect to both the distribution of groups’ potential outcomes and treatments, while in all the unstarred sections that preceded, groups’

treatments (the study design \mathbf{D}) were implicitly conditioned upon (see starred Section 2.4 for further discussion). To highlight expectations taken with respect to both the distribution of groups' potential outcomes and treatments, we let $E_u[\cdot]$ denote such expectations. Finally, an estimand refers to a function of the probability distribution of the observed random variables, namely (Y_1, Y_2, D_1, D_2) in the simple setting we consider.

Formal definition of HAD. We can now formally define HADs.

Design HAD (*Heterogeneous adoption design*) $D_1 = 0, D_2 \geq 0, E_u(D_2) > 0$ and $V_u(D_2|D_2 > 0) > 0$.¹

HADs with stayers or quasi-stayers. Some of the results in this chapter apply to a subset of HADs, namely HADs with stayers or quasi-stayers.

Design HAD' (*Heterogeneous adoption design with stayers or quasi-stayers*) *The conditions in Design HAD hold and the support of D_2 includes 0.*²

The support condition in Design HAD' holds in two important cases. The first case is when there are groups whose period-two treatment is equal to zero: $P_u(D_2 = 0) > 0$. Hereafter, those groups are referred to as stayers: their treatment does not change from period one to two, they *stay* untreated. The second case is when there are no groups whose period-two treatment is equal to zero ($P_u(D_2 = 0) = 0$), but there are groups whose period-two treatment is “very close” to zero: for any $\delta > 0$, $P_u(0 < D_2 < \delta) > 0$. For instance, this condition holds if D_2 is continuously distributed on \mathbb{R}_+ with a continuous density that is strictly positive at 0. Hereafter, groups with a period-two treatment close to zero are referred to as quasi-stayers.

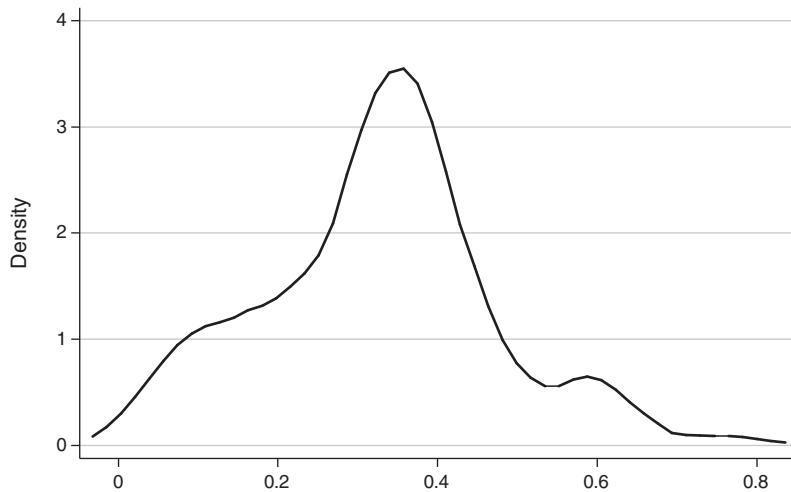
Chapter's running example. Before China joined the World Trade Organization (WTO), United States (US) imports from China were already subject to the low Normal Trade Relations (NTR) tariff rates reserved to WTO members, since 1980. However, those rates required uncertain and politically contentious annual renewals. Without renewal, US import tariffs on Chinese goods would have spiked to higher non-NTR tariff rates. In 2001, when China joined the WTO,

¹Note that $E_u(D_2) > 0$ implies $P_u(D_2 > 0) > 0$, so that $V_u(D_2|D_2 > 0)$ is well defined.

²Recall that the support of a random variable A is the smallest closed set C such that $P(A \in C) = 1$.

the US granted Permanent NTR (PNTR) to China. The reform eliminated a potential tariff spike, equal to the difference between the non-NTR and NTR tariff rates, referred to as the NTR gap. This NTR gap varies substantially across industries: without NTR renewal, in some industries there would have been a large increase in tariffs on Chinese imports, while in other industries the increase would have been smaller. Pierce and Schott (2016) study the effect of the NTR-gap treatment on US manufacturing employment. They define their treatment $D_{g,t}$ as the interaction of industry g 's NTR gap and t being after 2001. Letting for instance $t = 1$ denote year 2000 and $t = 2$ denote year 2001, $D_{g,1} = 0$, $D_{g,2} \geq 0$, and there is variability in the NTR gap of treated industries: the conditions in Design HAD are met. The NTR gap is strictly positive in all industries. Therefore, there are no stayers in this application. However, Figure 2 in Pierce and Schott (2016), reproduced below, suggests that the NTR gap's density is strictly positive in the neighborhood of 0: while the NTR gap is strictly positive in all industries, there are industries where it is close to zero. Therefore, there seems to be quasi-stayers in this application. Below, we show that a statistical test that there are quasi-stayers is not rejected.

Figure 7.1: Density of period-two treatment dose in Pierce and Schott (2016)



Dataset used in this chapter. To answer the green questions in this chapter, you need to use the `pierce_schott_didtextbook` dataset, which is constructed from the data used by Pierce and Schott (2016) to produce their Table 3. The data used by the authors to produce their other regression tables is proprietary. The `pierce_schott_didtextbook` dataset contains

the following variables, for 103 US industries:³

- **indusid:** an industry identifier;
- **ntrgap:** the industry's NTR gap;
- **ntrgapsq:** the square of the industry's NTR gap;
- **delta2001, delta2002, delta2004, and delta2005,** which respectively correspond to the change in the industry's log employment from 2000 to 2001, from 2000 to 2002, from 2000 to 2004, and from 2000 to 2005;⁴
- **delta1999, delta1998, and delta1997,** which respectively correspond to the change in the industry's log employment from 2000 to 1999, from 2000 to 1998, and from 2000 to 1997;
- **deltalintrend2001, deltalintrend2002, ..., deltalintrend2005,** which respectively mimick **delta2001, delta2002, ..., delta2005**, netting out industry-specific linear trends from the industry's employment evolution: the construction of those variables is detailed below;
- **deltalintrend1998 and deltalintrend1997,** which respectively mimick **delta1998** and **delta1997**, netting out industry-specific linear trends from the industry's employment evolution: the construction of those variables is detailed below;
- **lemp1997, lemp1998, lemp1999, and lemp2000:** the industry's log employment in 1997, 1998, 1999, and 2000;
- **cons:** a constant variable equal to one.

³While the version of the UNIDO dataset downloaded by the authors had 104 US industries, the version we downloaded in 2023 has 103 industries, presumably due to some industry regrouping.

⁴As noted by Pierce and Schott (2016), 2003 data is missing for all US industries in the UNIDO dataset used to produce their Table 3.

Fuzzy designs. HADs are often “fuzzy” DID designs (de Chaisemartin, 2011; de Chaisemartin and D’Haultfoeuille, 2018), where the treatment varies at the individual level, and $D_{g,t}$ is the treatment rate of individuals in cell (g,t) . For instance, in 1996 a new TV channel, called NTV, was introduced in Russia. At that time, it was the only TV channel in Russia not controlled by the government. Enikolopov, Petrova and Zhuravskaya (2011) study the effect of having access to this independent news source on voting behavior, using voting outcomes for the 1938 Russian subregions in the 1995 and 1999 elections. The authors define their treatment as $D_{g,t}$, the proportion of the population having access to NTV in region g and year t , hereafter referred to as the NTV exposure rate. By definition $D_{g,1} = 0$, $D_{g,2} \geq 0$, and there is variability in the NTV exposure rate across treated regions: the conditions in Design HAD are met. In the region with the lowest exposure rate to NTV, this rate is still equal to 27%, so there is no unexposed (stayer) or almost unexposed (quasi-stayer) region.⁵

7.1 Identifying assumptions

Throughout this chapter, we impose Assumption NA, the no-anticipation assumption. We also impose Assumption ND, thus ruling out dynamic effects of the treatment. With only two time periods and no treated unit at period one, Assumption ND is not of essence but imposing it simplifies notation. Finally, we replace Assumption PT by a different parallel-trends assumption, better suited to case with a non-binary treatment and i.i.d. groups we consider. For any variable X , let $\Delta X = X_2 - X_1$ denote the change in X from period one to two.

Assumption PTNB (*Parallel-trends with a non-binary treatment*) *There is a real number γ_2 such that $E_u[\Delta Y(0)|D_2] = \gamma_2$.*

[Interpret Assumption PTNB.](#)

⁵In a fuzzy design, our (g,t) -level potential outcome notation $Y_{g,t}(d)$ assumes that the outcome of g at t can only depend on the proportion of treated units in g at t , not on the identities of the treated units. Actually, the results below still hold if potential outcomes depend on the identities of the treated units. Letting $Y_{i,g,t}(0)$, $Y_{i,g,t}(1)$, and $Y_{i,g,t}$ denote the untreated, treated, and observed outcomes of unit i in group g at t , letting $N_{1,g,t}$ denote the number of treated units in group g at t , one just needs to redefine $\text{TE}_{g,t}$ as $\text{TE}_{g,t} = \frac{1}{N_{1,g,t}} \sum_{i:D_{i,g,t}=1} E_u[Y_{i,g,t}(1) - Y_{i,g,t}(0)]$.

Assumption PTNB is a parallel-trends assumption. It requires that groups' untreated outcome evolution is mean independent of their period-two treatment: without treatment, strongly and weakly treated groups would not have experienced systematically different outcome evolutions. This condition is similar to a strong-exogeneity assumption in panel data models. This shows that parallel trends and strong exogeneity are, essentially, two different ways of expressing the same idea.

Pre-trends test of Assumption PTNB. If the data contains another pre-period $t = 0$ where groups are all untreated, as in period $t = 1$, Assumption PTNB can be “placebo-tested”, for instance by regressing $Y_{g,1} - Y_{g,0}$ on $D_{g,2}$, because $Y_{g,1} - Y_{g,0} = Y_{g,1}(0) - Y_{g,0}(0)$ is an outcome evolution without treatment.

7.2 Target parameters

7.2.1 Building blocks: actual-versus-no-treatment slopes

The group-specific treatment effects we consider are

$$\text{TE}_2 = \frac{Y_2(D_2) - Y_2(0)}{D_2},$$

the slopes of groups' potential outcome functions between 0 and their actual treatments. We refer to TE_2 as “actual-versus-no-treatment” slopes.

Counterfactual-versus-no-treatment slopes? For a group receiving two doses of treatment at period two ($D_2 = 2$), $\text{TE}_2 = (Y_2(2) - Y_2(0))/2$. Of course, one may be interested in slopes of that group's potential outcome function at different treatment doses, such as $(Y_2(3) - Y_2(0))/3$. For $d_2 \neq D_2$, we refer to $(Y_2(d_2) - Y_2(0))/d_2$ as a counterfactual-versus-no-treatment slope. Ideally, one would like to learn groups' entire dose-slope function

$$d_2 \mapsto \frac{Y_2(d_2) - Y_2(0)}{d_2},$$

for instance to determine the dose that generates the highest return per dose. However, for a group with $D_2 = 2$, estimating, say, $(Y_2(3) - Y_2(0))/3$ is harder than estimating $(Y_2(2) - Y_2(0))/2$.

As $D_2 = 2$, $Y_2(2)$ is the group's actual, observed outcome, so estimating $(Y_2(2) - Y_2(0))/2$ only requires estimating one unobserved outcome, $Y_2(0)$. On the other hand, estimating $(Y_2(3) - Y_2(0))/3$ requires estimating two unobserved outcomes, $Y_2(3)$ and $Y_2(0)$. While estimating $Y_2(0)$ can be achieved under Assumption PTNB, a parallel-trends assumption whose plausibility can be assessed using a pre-trends test, estimating $Y_2(3)$ requires making assumptions whose plausibility cannot be assessed using a pre-trends test. This is why we focus on actual-versus-no-treatment slopes, rather than on counterfactual-versus-no-treatment slopes.

Actual-versus-counterfactual slopes? Another type of slopes one might be interested in is $(Y_2(D_2) - Y_2(d))/(D_2 - d)$ for $d > 0$, which we refer to as an actual-versus-counterfactual slope. When d tends to D_2 , and assuming that $Y_2(d)$ is almost surely differentiable everywhere, this slope converges towards $Y'_2(D_2)$, the derivative of the potential outcome function evaluated at D_2 . $E_u(Y'_2(D_2))$ is called the average marginal effect, a parameter that has often been studied in the literature. Like TE_2 , estimating an actual-versus-counterfactual slope only requires estimating one unobserved outcome. However, the unobserved outcome in the actual-versus-counterfactual slope, $Y_2(d)$, is not observed at $t = 1$ and in prior periods. Therefore, estimating actual-versus-counterfactual slopes again requires making assumptions whose plausibility cannot be assessed using a pre-trends test.

Target parameters: averages of actual-versus-no-treatment slopes. TE_2 is a group-specific effect, which cannot be consistently estimated. Instead, we will now focus on averages of those group-specific slopes, which can be consistently estimated.

Bounded-slope assumption.* Throughout this chapter, we assume that there exists a real number K such that for all $d_2 > 0$ in the support D_2 , $|Y_2(d_2) - Y_2(0)|/d_2 \leq K$ almost surely. This ensures that the expectations of the slopes $(Y_2(d_2) - Y_2(0))/d_2$ introduced below are well defined. This condition holds if $d_2 \mapsto Y_2(d_2)$ is differentiable with a bounded derivative.

7.2.2 Average of actual-versus-no-treatment slopes conditional on the dose received

The conditional-average-slope function. The first target we consider is

$$\text{CAS}(d_2) := E_u (\text{TE}_2 | D_2 = d_2), \text{ for all } d_2 \text{ in the support of } D_2.$$

$\text{CAS}(d_2)$ is the average of the slopes TE_2 , across all groups with treatment dose d_2 . Hereafter, $d_2 \mapsto \text{CAS}(d_2)$ is referred to as the conditional average slope function (CAS).

Assume that $\text{CAS}(3) < \text{CAS}(1)$. Can we conclude that three doses of treatment generate a lower return per dose than one dose, thus implying that treatment has diminishing returns? Or could we have that $\text{CAS}(3) < \text{CAS}(1)$ with constant or even increasing returns?

The CAS conflates a dose-slope relationship and a selection bias.

$$\text{CAS}(3) = E_u ((Y_2(3) - Y_2(0))/3 | D_2 = 3),$$

and

$$\text{CAS}(1) = E_u ((Y_2(1) - Y_2(0))/1 | D_2 = 1).$$

Therefore, $\text{CAS}(3)$ and $\text{CAS}(1)$ are averages of different slopes across different populations:

$\text{CAS}(3)$ averages **0-to-3 slopes** across groups that received *3 doses*,

while

$\text{CAS}(1)$ averages **0-to-1 slopes** across groups that received *1 dose*.

Therefore, $\text{CAS}(3) < \text{CAS}(1)$ could be due to diminishing returns, or it could be due to the fact that groups that received three doses have lower returns per dose than groups that received one dose. For instance, one could have that

$$\begin{aligned} E_u \left(\frac{Y_2(1) - Y_2(0)}{1} \middle| D_2 = 3 \right) &= E_u \left(\frac{Y_2(3) - Y_2(0)}{3} \middle| D_2 = 3 \right) \\ &< E_u \left(\frac{Y_2(1) - Y_2(0)}{1} \middle| D_2 = 1 \right) = E_u \left(\frac{Y_2(3) - Y_2(0)}{3} \middle| D_2 = 1 \right), \end{aligned}$$

in which case the difference between CAS(3) and CAS(1) is entirely driven by the selection bias (groups receiving one and three doses have different returns per dose), while returns per dose are constant within those two groups.

Find a supplementary assumption under which one can conclude that treatment has diminishing returns when $CAS(3) < CAS(1)$. When is that assumption plausible, and when is it implausible?

An homogeneous slope assumption. Let

$$ASF(d_2) = E_u \left(\frac{Y_2(d_2) - Y_2(0)}{d_2} \right)$$

denote the average 0-to- d_2 slope of the period-two potential outcome function, across all groups and not only across groups receiving a dose equal to d_2 . Hereafter, $d_2 \mapsto ASF(d_2)$ is referred to as the Average Slope Function. Assume that for all $d_2 > 0$ in the support of D_2 ,

$$E_u \left(\frac{Y_2(d_2) - Y_2(0)}{d_2} \middle| D_2 = d_2 \right) = E_u \left(\frac{Y_2(d_2) - Y_2(0)}{d_2} \right), \quad (7.1)$$

meaning that the average 0-to- d_2 slope across groups receiving a dose equal to d_2 is the same as across all groups. Under (7.1), the CAS and ASF are equal. Then, if one has, say, $CAS(3) < CAS(1)$, this also implies that $ASF(3) < ASF(1)$, which can be interpreted as evidence of diminishing returns per dose. However, (7.1) is a strong restriction. It is very unlikely to hold in a Roy selection model where groups select their dose based on the gains they expect from treatment. It holds if treatment doses are as good as randomly assigned, but this type of as-good-as-random assumptions should be substantiated by thorough balancing checks to demonstrate that high- and low-dose groups are similar on a number of observables. Moreover, if doses are as good as randomly assigned, there is no need to resort to a DID or TWFE estimator, one can just regress Y_2 on D_2 . Callaway, Goodman-Bacon and Sant'Anna (2021) study HADs, and propose a “strong-parallel-trends” assumption under which that function is equal to the dose-response relationship. Under Assumption PTNB, their “strong-parallel-trends” assumption is equivalent to the homogeneous slope assumption in (7.1).

7.2.3 Two unconditional averages of actual-versus-no-treatment slopes

ATT. The first unconditional average of slopes we consider is just

$$\text{ATT} := E_u(\text{TE}_2 | D_2 > 0),$$

the average of TE_2 across all treated groups. By the law of iterated expectations,

$$\text{ATT} = E_u(\text{TE}_2 | D_2 > 0) = E_u(E_u(\text{TE}_2 | D_2) | D_2 > 0) = E_u(\text{CAS}(D_2) | D_2 > 0) : \quad (7.2)$$

ATT is a weighted average of the conditional average slopes $E_u(\text{TE}_2 | D_2 = d_2)$. For instance, if a half of treated groups receive one dose of treatment at period two while the other half receive two doses,

$$\text{ATT} = \frac{1}{2}E(\text{TE}_2 | D_2 = 1) + \frac{1}{2}E(\text{TE}_2 | D_2 = 2).$$

If D_2 is a continuously distributed variable, $E_u(\text{TE}_2 | D_2 = d_2)$ receives a weight equal to $f_{D_2|D_2>0}(d_2)$, the density of D_2 conditional on $D_2 > 0$ evaluated at d_2 .

Weighted ATT. The second unconditional average of slopes we consider is

$$\text{WATT} := E_u \left[\frac{D_2}{E_u[D_2 | D_2 > 0]} \text{TE}_2 \middle| D_2 > 0 \right].$$

The WATT is a weighted average of treated groups' slopes, where groups with a larger period-two treatment receive more weight. For instance, if a half of treated groups receive one dose of treatment at period two while the other half received two doses, $E_u[D_2 | D_2 > 0] = 1/2 \times 1 + 1/2 \times 2 = 3/2$, and

$$\text{WATT} = \frac{1}{2} \frac{1}{3/2} E(\text{TE}_2 | D_2 = 1) + \frac{1}{2} \frac{2}{3/2} E(\text{TE}_2 | D_2 = 2) = \frac{1}{3} E(\text{TE}_2 | D_2 = 1) + \frac{2}{3} E(\text{TE}_2 | D_2 = 2).$$

Find a condition under which $\text{ATT} = \text{WATT}$.

ATT = WATT if and only if treated-groups' slopes are uncorrelated with their treatment dose:

$$\text{cov}(TE_2, D_2 | D_2 > 0) = 0. \quad (7.3)$$

This condition is similar to, but weaker than, (7.1). As that previous condition, it is unlikely to hold in a Roy selection model.

A statistical and an economic argument in favor of considering the WATT. The WATT may seem to be a less natural target parameter than the ATT. Yet, de Chaisemartin, D'Haultfœuille, Pasquier, Sow and Vazquez-Bare (2022) put forward some arguments to consider this parameter. First, estimating the ATT may sometimes be more difficult than estimating the WATT. When there are treated groups with a value of D_2 close to zero, the denominator of TE_2 is close to zero for those groups. Then, as we will discuss in more details later in this chapter, estimators of the ATT may suffer from a small-denominator problem, which could substantially increase their variance, and even affect their convergence rate (see Graham and Powell, 2012; Sasaki and Ura, 2021, for similar issues in related contexts). On the other hand,

$$\text{WATT} = E_u \left[\frac{D_2}{E_u[D_2 | D_2 > 0]} \frac{Y_2(D_2) - Y_2(0)}{D_2} \middle| D_2 > 0 \right] = \frac{E_u[Y_2(D_2) - Y_2(0) | D_2 > 0]}{E_u[D_2 | D_2 > 0]}. \quad (7.4)$$

As the WATT is a ratio of expectations rather than the expectation of a ratio, it is not affected by a small-denominator problem, even if there are treated groups with a value of D_2 close to zero. Second, de Chaisemartin et al. (2022) show that the WATT is the relevant quantity to consider in a cost-benefit analysis assessing if groups' period-two treatment D_2 is beneficial. Assume that the outcome is a measure of output, such as agricultural yields or wages, expressed in monetary units. Assume also that the treatment is costly, with a cost linear in dose, uniform across groups, and known to the analyst: the cost of giving d doses of treatment is $c \times d$ for some known $c > 0$. Then, D_2 is beneficial relative to a no-treatment counterfactual if and only if $E_u(Y_2(D_2) - cD_2) > E_u(Y_2(0))$, namely if and only if

$$\text{WATT} > c.$$

Then, comparing the WATT to the per-unit treatment cost is sufficient to evaluate if changing the treatment from 0 to D_2 was beneficial.

7.3 TWFE estimator in heterogeneous-adoption designs

7.3.1 Under parallel-trends, $\hat{\beta}^{\text{fe}}$ may not identify a convex combination of slopes

In an HAD, the TWFE estimator compares the outcome evolutions of more- and less-treated groups. Let $\hat{\beta}^{\text{fe}}$ denote the coefficient on $D_{g,t}$ in a regression of $Y_{g,t}$ on group FEs, an indicator for period 2, and $D_{g,t}$, as defined in (3.1), with $T = 2$:

$$Y_{g,t} = \sum_{g'=1}^G \hat{\alpha}_{g'} 1\{g = g'\} + \hat{\gamma}_2 1\{t = 2\} + \hat{\beta}^{\text{fe}} D_{g,t} + \hat{\epsilon}_{g,t}.$$

As $T = 2$, $\hat{\beta}^{\text{fe}}$ is numerically equivalent to the coefficient on $D_{g,2}$ in a first-difference regression of ΔY_g on a constant and $D_{g,2} - D_{g,1} = D_{g,2}$. Then, find a simple closed-form formula for $\hat{\beta}^{\text{fe}}$.

The regression of ΔY_g on a constant and $D_{g,2}$ is a univariate regression. Therefore, $\hat{\beta}^{\text{fe}}$ is equal to the sample covariance between ΔY_g and $D_{g,2}$, divided by the sample variance of $D_{g,2}$:

$$\hat{\beta}^{\text{fe}} = \frac{\sum_{g=1}^G (D_{g,2} - D_{.,2}) \Delta Y_g}{\sum_{g=1}^G (D_{g,2} - D_{.,2})^2}. \quad (7.5)$$

If $D_{g,2}$ were binary, $\hat{\beta}^{\text{fe}}$ would just compare the average outcome evolutions of treated and control groups. With a non-binary treatment, $\hat{\beta}^{\text{fe}}$ still implements a form of treated-versus-control comparison. In the numerator of $\hat{\beta}^{\text{fe}}$, the outcome evolutions of groups such that $D_{g,2} - D_{.,2} > 0$ enter with a positive sign, so those groups are used as “treatment groups”. On the other hand, groups such that $D_{g,2} - D_{.,2} < 0$ are used as “control groups”: their outcome evolutions are weighted negatively. Thus, $\hat{\beta}^{\text{fe}}$ compares the outcome evolutions of more- and less-treated groups.

Application to the NTR-gap example. Using the `pierce_schott_didtextbook` dataset, regress `delta2001`, industries’ evolutions of log employment from 2000 to 2001, on the NTR-gap treatment. Similarly, regress `delta2002`, `delta2004`, and `delta2005` on the NTR-gap treatment. According to these regressions, do PNTR with China have an effect on US employment?

```
reg delta2001 ntrgap, vce(hc2, dfadjust)
```

As $G = 103$ is not very large, we follow the recommendations from Section 3.3 and use HC2 standard errors with the DOF adjustment recommended by Bell and McCaffrey (2002) to obtain more reliable confidence intervals. The coefficient on `ntrgap` is small and insignificant at the 5% level ($\hat{\beta}^{\text{fe}} = -0.061$, 95% CI=[−0.143, 0.020]). The coefficient on `ntrgap` is larger and significant when we regress `delta2002` on `ntrgap` ($\hat{\beta}^{\text{fe}} = -0.260$, 95% CI=[−0.413, −0.107]), and becomes even larger when we regress `delta2004` ($\hat{\beta}^{\text{fe}} = -0.540$, 95% CI=[−0.849, −0.230]) and `delta2005` ($\hat{\beta}^{\text{fe}} = -0.532$, 95% CI=[−0.870, −0.194]) on `ntrgap`. This suggests that while the elimination of a potential tariffs' spike did not immediately impact US manufacturing employment, this treatment had a large negative effect after a few years. According to our TWFE regressions, eliminating a potential tariffs' spike of 100 percentage points reduces employment by more than 0.5 percentage points in 2004 and 2005.

Under parallel trends, $\hat{\beta}^{\text{fe}}$ may not estimate the ATT, or a convex combination of slopes. Let β^{fe} denote the probability limit of $\hat{\beta}^{\text{fe}}$ when $G \rightarrow +\infty$, assuming that $(D_{g,2}, \Delta Y_g)$ has a finite second moment to ensure this probability limit exists.

Theorem 13 *In Design HAD, if Assumptions NA, ND, and PTNB hold,*

$$\begin{aligned}\beta^{\text{fe}} &= \frac{\text{cov}_u(E_u(Y_2(D_2) - Y_2(0)|D_2), D_2)}{V(D_2)} \\ &= E_u\left(\frac{(D_2 - E_u(D_2))D_2}{E_u((D_2 - E_u(D_2))D_2|D_2 > 0)} E_u(\text{TE}_2|D_2) \middle| D_2 > 0\right).\end{aligned}\quad (7.6)$$

A proof of Theorem 13 is in this chapter's appendix, as all the other proofs given in this chapter. The first equality shows that β^{fe} is equal to the coefficient from a regression of the conditional average dose-response $E_u(Y_2(D_2) - Y_2(0)|D_2)$ on D_2 , so β^{fe} may be used to assess how those two variables correlate. On the other hand, the second equality shows that β^{fe} may not provide an easily interpretable measure of the treatment's effect: it is equal to a weighted sum of the conditional average slopes $E_u(\text{TE}_2|D_2 = d_2)$, where $E_u(\text{TE}_2|D_2 = d_2)$ receives a weight proportional to

$$(d_2 - E_u(D_2))d_2 P_u(D_2 = d_2 | D_2 > 0)$$

if D_2 is discrete, and proportional to

$$(d_2 - E_u(D_2))d_2 f_{D_2|D_2>0}(d_2)$$

if D_2 is continuously distributed. As $d_2 \mapsto (d_2 - E_u(D_2))d_2$ is not constant, (7.2) and (7.6) imply that $\hat{\beta}^{\text{fe}}$ may not be consistent for ATT. $\hat{\beta}^{\text{fe}}$ may not even be consistent for a convex combination of conditional average slopes: some weights in (7.6) are negative if $P_u(0 < D_2 < E_u(D_2)) > 0$. For instance, this condition always holds when there are no stayers.

Comparison with binary-and-staggered designs. Theorem 13 shows that $\hat{\beta}^{\text{fe}}$ may fail to identify a convex combination of effects, even without variation in treatment timing. In binary-and-staggered designs, we have seen that time-varying effects could lead $\hat{\beta}^{\text{fe}}$ to estimate a non-convex combination of effects. Here, treatment effects that vary across groups receiving different treatment doses could lead $\hat{\beta}^{\text{fe}}$ to estimate a non-convex combination of effects.

Bibliographic notes. Theorem 13 is an asymptotic version of Proposition S1 in de Chaisemartin and D'Haultfoeuille (2020).

Application to the NTR-gap example. To compute the weights attached to $\hat{\beta}^{\text{fe}}$ in the regression of `delta2001` on `ntrgap`, use the `twowayfeweights` Stata command, and run:

```
twowayfeweights delta2001 indusid cons ntrgap ntrgap, type(fdTR)
```

The `fdTR` option reflects the fact that $\hat{\beta}^{\text{fe}}$ arises from a first-difference regression of `delta2001` on `ntrgap`. Here the regression is estimated with only two periods and one first difference, so the regression does not have time FEs, and the time variable inputted to the command is just a constant. Finally, to compute weights attached to first-difference regressions with `twowayfeweights`, on top of the first-differenced outcome, group id, time id, and first-differenced treatment variables, one also needs to input the treatment of cell (g, t) , hence the command's fifth argument `ntrgap`. Here, this fifth argument is redundant, because in an HAD the treatment and the first-differenced treatment are equal. Interpret the results: does $\hat{\beta}^{\text{fe}}$ estimate a convex or almost convex combination of slopes of log-employment with respect to the NTR gap?

$\hat{\beta}^{\text{fe}}$ estimates a weighted sum of slopes of log-employment with respect to the NTR gap in 103

industries, where 62 estimated weights are strictly positive, while 41 are strictly negative. The negative weights sum to -0.32, so $\hat{\beta}^{\text{fe}}$ is far from estimating a convex combination of slopes.

7.3.2 The origin of the negative weights in heterogeneous-adoption designs

Intuitively, why is it that $\hat{\beta}^{\text{fe}}$ may estimate a non-convex combination of effects?

Intuition for the negative weights in Theorem 13. $\hat{\beta}^{\text{fe}}$ compares the outcome evolution of more- and less-treated groups. However, less-treated groups may still be treated, in which case their treatment effect gets differenced out and weighted negatively by $\hat{\beta}^{\text{fe}}$.

Forbidden comparisons in HADs: a tale of two patients who had a headache. To gain further intuition, let us consider a simple example, with only two groups m and ℓ , corresponding to two patients who start having a headache at $t = 1$. At $t = 2$, they go see their doctor, who prescribes two Ibuprofen pills to m , and one Ibuprofen pill to ℓ . Thus m is the more-treated patient, while ℓ is the less-treated one. An econometrician comes by, and immediately sees some research potential in this natural experiment. Accustomed to running TWFE regressions, they regress $Y_{g,t}$, the pain level of patient g at t , on patient FEs, period FEs, and the treatment received by patient g at t . With $G = 2$, one can show that

$$\hat{\beta}^{\text{fe}} = \frac{\Delta Y_m - \Delta Y_\ell}{D_{m,2} - D_{\ell,2}}. \quad (7.7)$$

As in our example, $D_{m,2} - D_{\ell,2} = 2 - 1 = 1$, (7.7) simplifies to

$$\hat{\beta}^{\text{fe}} = \Delta Y_m - \Delta Y_\ell :$$

$\hat{\beta}^{\text{fe}}$ is just a simple DID comparing the evolution of the pain of m and ℓ , before and after they take their Ibuprofen prescription. To his surprise, the econometrician finds that $\hat{\beta}^{\text{fe}} > 0$: the pain of patient m , who received more Ibuprofen, decreases less than that of patient ℓ , who received

less Ibuprofen. Is it correct to conclude that Ibuprofen increases pain, or could something else explain why $\hat{\beta}^{\text{fe}} > 0$?

It could be the case that the drug reduces the pain of both patients, but the effect of the drug per pill is more than twice lower for patient m than for patient ℓ , so their pain level decreases less than that of patient ℓ , despite the fact they received two pills while patient ℓ only received one pill. This scenario is not completely unlikely: perhaps the reason why the doctor prescribed two pills to m is because they believed that m would have a lower sensitivity to the drug. This shows that in HADs, negative weights arise from a new type of forbidden comparisons: comparing the outcome evolutions of more- and less-treated groups.

Forbidden comparisons in HADs: formal analysis. Mathematically, and assuming that m and ℓ have the same expected pain evolutions if they do not take Ibuprofen ($E_u(\Delta Y_m(0)) = E_u(\Delta Y_\ell(0))$),

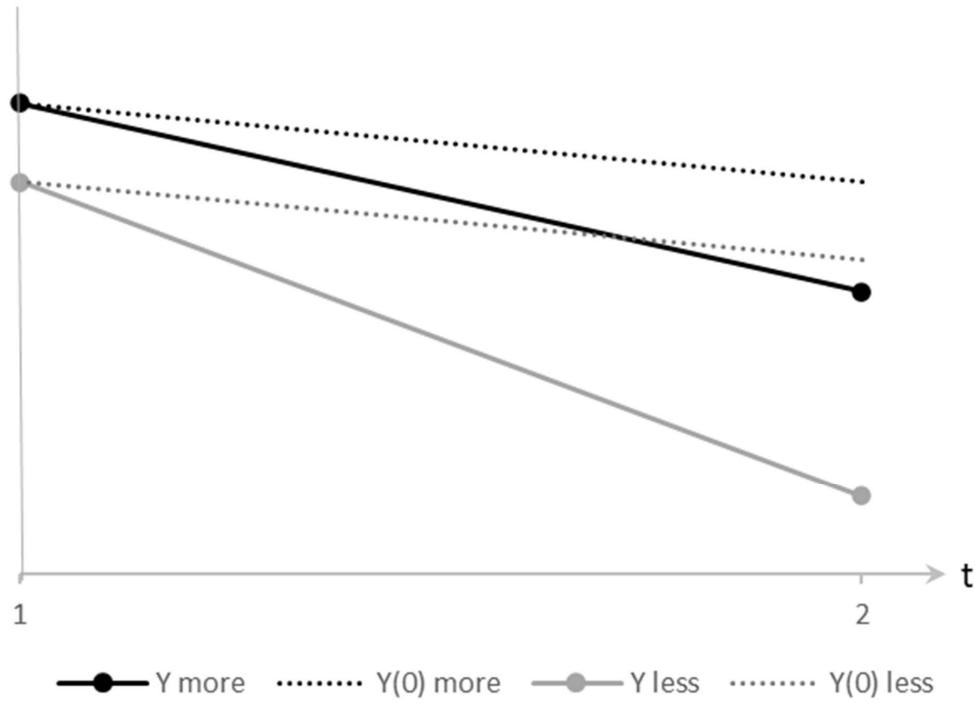
$$\begin{aligned} E_u[\hat{\beta}^{\text{fe}}] &= E_u(Y_{m,2}(2) - Y_{m,1}(0) - (Y_{\ell,2}(1) - Y_{\ell,1}(0))) \\ &= E_u(Y_{m,2}(0) - Y_{m,1}(0) - (Y_{\ell,2}(0) - Y_{\ell,1}(0))) + E_u(Y_{m,2}(2) - Y_{m,2}(0) - (Y_{\ell,2}(1) - Y_{\ell,2}(0))) \\ &= E_u(2\text{TE}_{m,2} - \text{TE}_{\ell,2}), \end{aligned}$$

where the last equality follows from the assumption that $E_u(\Delta Y_m(0)) = E_u(\Delta Y_\ell(0))$, and from the definitions of $\text{TE}_{m,2}$ and $\text{TE}_{\ell,2}$. The right-hand-side of the previous display is a weighted sum of m and ℓ 's treatment effects per dose of Ibuprofen, with weights summing to one, and where ℓ 's effect is weighted negatively. Intuitively, ℓ is also treated at period two, and $\hat{\beta}^{\text{fe}}$, which uses ℓ as a control group, subtracts its treatment effect out. If $\text{TE}_{m,2}$ and $\text{TE}_{\ell,2}$ are both negative, but m 's effect is more than twice smaller in absolute value than ℓ 's effect, $E_u[\hat{\beta}^{\text{fe}}] > 0$.

Forbidden comparisons in HADs: graphical representation. Figure 7.2 below shows the actual and counterfactual outcome evolution, in a numerical example where both patients benefit from Ibuprofen, but the treatment effect per dose is more than twice larger for the patient

that receives one dose than for the patient that receives two, thus leading to a positive TWFE coefficient.

Figure 7.2: A tale of two patients who had a headache.



Bibliographic notes. The right hand side of (7.7) is a Wald-DID estimator, that may not estimate a convex combination of treatment effects under a parallel-trends assumption (Blundell and Costa-Dias, 2009; de Chaisemartin, 2011; de Chaisemartin and D'Haultfœuille, 2018).

7.3.3 Assuming randomized treatment dose rather than parallel trends*

If groups' treatment dose is randomly assigned, $\hat{\beta}^{fe}$ estimates a convex combination of effects. Theorem 13 holds under no-anticipation and parallel-trends assumptions. If instead of parallel of trends, one is ready to assume that the treatment dose is as good as randomly assigned, then one can show, leveraging a decomposition from Yitzhaki (1996), that $\hat{\beta}^{fe}$ is unbiased for a convex combination of treatment effects. Let us give some intuition for that result in the simple example we considered earlier, with two groups, a more-treated group that receives

two doses of treatment, and a less-treated groups that receives one dose. Assume that the more treated group is chosen at random: with probability 1/2, the more-treated group is group 1, and with probability 1/2 the more-treated group is group 2. Thus, m and ℓ are now random variables, with $P_u(m = 1) = P_u(m = 2) = 1/2$, and $\ell = 3 - m$. Assume also that potential outcomes are non-stochastic. For any random variable X , let $E_m(X)$ denote the expectation of X with respect to the identity of the more-treated group. Then,

$$\begin{aligned} & E_m [\hat{\beta}^{\text{fe}}] \\ &= E_m [\Delta Y_m - \Delta Y_\ell] \\ &= E_m [Y_{m,2}(2) - Y_{m,1}(0) - (Y_{\ell,2}(1) - Y_{\ell,1}(0))] \\ &= P_u(m = 1) [Y_{1,2}(2) - Y_{1,1}(0) - (Y_{2,2}(1) - Y_{2,1}(0))] + P_u(m = 2) [Y_{2,2}(2) - Y_{2,1}(0) - (Y_{1,2}(1) - Y_{1,1}(0))] \\ &= \frac{1}{2} [Y_{1,2}(2) - Y_{1,1}(0) - (Y_{2,2}(1) - Y_{2,1}(0))] + \frac{1}{2} [Y_{2,2}(2) - Y_{2,1}(0) - (Y_{1,2}(1) - Y_{1,1}(0))] \\ &= \frac{1}{2} [Y_{1,2}(2) - Y_{1,2}(1)] + \frac{1}{2} [Y_{2,2}(2) - Y_{2,2}(1)]. \end{aligned}$$

Therefore, $\hat{\beta}^{\text{fe}}$ is unbiased for the average effect of increasing the treatment from one to two doses at period two, a convex combination of treatment effects.

If treatment doses are randomly assigned, cross-sectional regressions controlling for the baseline outcome may be more efficient than TWFE regressions. In Chapter 3, we saw that if all groups are untreated at period 1, and some are randomly assigned to a binary treatment at period 2, $\hat{\beta}^{\text{fe}}$ can have a higher variance than the treatment coefficient in a regression of $Y_{g,2}$ on an intercept and $D_{g,2}$ controlling for $Y_{g,1}$ (Frison and Pocock, 1992; McKenzie, 2012). There is no reason to suspect that this conclusion would change in an RCT where the treatment dose can vary at period 2.

As-good-as-random treatment dose is a strong assumption, which should be thoroughly tested. In natural experiments where researchers do not effectively randomize the treatment dose, a randomization claim should be substantiated with thorough and preferably pre-specified balancing checks showing that groups' treatment dose does not predict covariates correlated with the outcome. If the treatment dose $D_{g,2}$ is randomly assigned, one should have

$$D_{g,2} \perp\!\!\!\perp (Y_{g,1}(0), Y_{g,2}(0)).$$

Groups' observed outcome at period one is equal to their untreated outcome. Then, the previous display implies that

$$D_{g,2} \perp\!\!\!\perp Y_{g,1}, \quad (7.8)$$

an equation that only involves observed variables, and can therefore be tested. To test (7.8), one can regress $Y_{g,1}$ on $D_{g,2}$. When several pre-treatment periods are available, one can regress $D_{g,2}$ on all pre-treatment outcomes.

Application to the NTR-gap example. The data starts in 1997 while treatment starts in 2001, so we observe industries' employment levels for four years before the PNTR treatment. Therefore, we regress the NTR-gap treatment on $(Y_{g,1997}, \dots, Y_{g,2000})$, and run an F-test that all coefficients are equal to zero. Using the `pierce_schott_didtextbook` dataset, regress `ntrgap` on `lemp1997`, `lemp1998`, `lemp1999`, and `lemp2000`. Do you reject the null that the coefficients on `lemp1997`, `lemp1998`, `lemp1999`, and `lemp2000` are all equal to zero? Accordingly, does the NTR-gap treatment seem to be as good as randomly assigned?

```
reg ntrgap lemp1997 lemp1998 lemp1999 lemp2000, vce(hc2, dfadjust)
```

The p-value of the test that the coefficients on `lemp1997`, `lemp1998`, `lemp1999`, and `lemp2000` are all equal to zero is 0.056. Thus, there is some indication that industries' NTR gap is predicted by their pre-treatment employment levels, which suggests that the NTR-gap treatment may not be as good as randomly assigned.

7.3.4 A testable condition under which $\beta^{\text{fe}} = \text{ATT}$ under parallel trends

7.3.4.1 A constant-and-linear-effect assumption

Assume that for all $d_2 > 0$ in the support of D_2 ,

$$E_u(\text{TE}_2 | D_2 = d_2) = \text{ATT}, \quad (7.9)$$

meaning that the conditional average slopes of groups receiving different doses do not differ. For instance, (7.9) requires that the average 0-to-1 slope of groups receiving one dose be equal to the average 0-to-2 slope of groups receiving two doses. Is (7.9) weaker or stronger than (7.1)?

Strictly speaking, (7.9) is neither weaker nor stronger than (7.1). However, one can show that (7.9) holds if (7.1) holds and if

$$d_2 \mapsto E_u(Y_2(d_2)) \text{ is linear.} \quad (7.10)$$

Thus, (7.9) holds if the treatment's effect is homogeneous across groups receiving different doses, as in (7.1), and linear, as in (7.10). To ease exposition, we sometimes refer to (7.9) as a constant-and-linear-effect assumption, though strictly speaking (7.1) and (7.10) are sufficient but not necessary for (7.9) to hold.

If (7.9) holds, do we have that $\hat{\beta}^{\text{fe}}$ is consistent for the ATT under Assumptions NA, ND, and PTNB?

If (7.9) holds,

$$\begin{aligned} \beta^{\text{fe}} &= E_u \left(\frac{(D_2 - E_u(D_2))D_2}{E_u((D_2 - E_u(D_2))D_2 | D_2 > 0)} E_u(\text{TE}_2 | D_2) \middle| D_2 > 0 \right) \\ &= \text{ATT} \times E_u \left(\frac{(D_2 - E_u(D_2))D_2}{E_u((D_2 - E_u(D_2))D_2 | D_2 > 0)} \middle| D_2 > 0 \right) \\ &= \text{ATT}. \end{aligned}$$

Condition (7.9) is sufficient but not necessary to have $\text{ATT} = \beta^{\text{fe}}$. We have $\text{ATT} = \beta^{\text{fe}}$ if and only if

$$\text{cov}_u \left(\frac{(D_2 - E_u(D_2))D_2}{E_u((D_2 - E_u(D_2))D_2 | D_2 > 0)}, E_u(\text{TE}_2 | D_2) \middle| D_2 > 0 \right) = 0, \quad (7.11)$$

a condition weaker than (7.9) (see Corollary 1 in de Chaisemartin and D'Haultfœuille, 2020).

7.3.4.2 The constant-and-linear-effect assumption is testable

de Chaisemartin, Ciccia, D'Haultfoeuille and Knau (2024) show that the constant-and-linear-effect assumption in (7.9) has a testable implication, and is fully testable in HADs with stayers or quasi-stayers. Let $\beta_0 = E_u(\Delta Y) - \beta^{\text{fe}}E_u(D_2)$ denote the probability limit of the intercept of the TWFE regression.

Theorem 14 Suppose that Assumption PTNB holds.

1. In Design HAD, if (7.9) holds then $E(\Delta Y|D_2) = \beta_0 + \beta^{\text{fe}}D_2$.
2. In Design HAD', $E(\Delta Y|D_2) = \beta_0 + \beta^{\text{fe}}D_2$ implies that (7.9) holds.

Theorem 14 implies that if Assumption PTNB and (7.9) hold, $E_u(\Delta Y|D_2)$ has to satisfy a certain property, which it may or may not satisfy, thus opening the possibility of testing Assumption PTNB and (7.9). What is this property, and how could you test it?

Point 1 of Theorem 14 shows that if Assumption PTNB and (7.9) hold, then $E_u(\Delta Y|D_2)$ has to be linear. In practice, $E_u(\Delta Y|D_2)$ could be non-linear. For instance, one may have that $E_u(\Delta Y|D_2) = \alpha_0 + \alpha_1 D_2 + \alpha_2 D_2^2$. To test whether $E_u(\Delta Y|D_2)$ is linear, one could regress ΔY on, say, an intercept, D_2 , D_2^2 , and D_2^3 , and test that the coefficients on D_2^2 and D_2^3 are both equal to zero. However, this test can only detect some, but not all non-linearities in $E_u(\Delta Y|D_2)$. For instance, if $E_u(\Delta Y|D_2)$ is a non-linear but non-polynomial function, this test might fail to reject the null of linearity, even asymptotically.

Interpreting a rejection of the linearity of $E_u(\Delta Y|D_2)$. Point 1 of Theorem 14 shows that under Assumption PTNB, if (7.9) holds, then $E_u(\Delta Y|D_2)$ is linear. By contraposition, if $E_u(\Delta Y|D_2)$ is not linear, then (7.9) cannot hold. This may suggest that $\hat{\beta}^{\text{fe}}$ is not consistent for the ATT. However, as mentioned above, the constant-and-linear-effect assumption in (7.9) is sufficient but not necessary to have that $\hat{\beta}^{\text{fe}}$ is consistent for the ATT.

Interpreting a failure to reject the linearity of $E_u(\Delta Y|D_2)$, in designs with stayers or quasi-stayers. In designs with stayers or quasi-stayers, Point 2 of Theorem 14 shows that under Assumption PTNB, there is an “if and only if” relationship between the constant-and-linear-effect assumption in (7.9) and the linearity of $E_u(\Delta Y|D_2)$. Therefore, if $E_u(\Delta Y|D_2)$ is linear then the constant-and-linear-effect assumption holds, thus implying that $\hat{\beta}^{\text{fe}}$ is consistent for the ATT. This suggests the following estimation rule: in designs with stayers or quasi-stayers, when a linearity test of $E_u(\Delta Y|D_2)$ and a pre-trends test of Assumption PTNB are not rejected, one may use $\hat{\beta}^{\text{fe}}$.

Interpreting a failure to reject the linearity of $E_u(\Delta Y|D_2)$, in designs without quasi-stayers. If there are no stayers or quasi-stayers, we no longer have an “if and only if” between (7.9) and $E_u(\Delta Y|D_2) = \beta_0 + \beta^{\text{fe}}D_2$: $E_u(\Delta Y|D_2) = \beta_0 + \beta^{\text{fe}}D_2$ could hold but $E_u(TE_2|D_2) = \beta^{\text{fe}} + (\beta_0 - \gamma_2)/D_2$, thus implying that (7.9) fails if $\beta_0 - \gamma_2 \neq 0$. Therefore, without stayers or quasi-stayers, $\hat{\beta}^{\text{fe}}$ may not be consistent for the ATT even if $E_u(\Delta Y|D_2)$ is linear.

In the general designs that we will review in the next chapter, linearity tests may be less useful to assess the validity of TWFE estimators.* Assume one uses a two-periods panel data set to estimate a treatment’s effect, but $D_1 \neq 0$ and $V_u(D_1) > 0$, so the conditions in Design HAD are not met: the treatment dose varies at period one. Then, the TWFE estimator is the coefficient on ΔD_g in a regression of ΔY_g on ΔD_g . Letting $TE_t = (Y_t(D_t) - Y_t(0))/D_t$, $Y_t = Y_t(0) + D_t TE_t$. If one is ready to assume that the treatment effect is constant over time ($TE_2 = TE_1$), then

$$\Delta Y = \Delta Y(0) + \Delta D \times TE_2.$$

Then, one can show that under Assumption PTNB, if there are stayers or quasi-stayers there is an “if and only if” relationship between

$$E_u(\Delta Y|\Delta D) = \beta_0 + \beta^{\text{fe}}\Delta D$$

and $E_u(TE_2|\Delta D) = E_u(TE_2|\Delta D \neq 0)$, a condition under which the TWFE estimator is consistent for $E_u(TE_2|\Delta D \neq 0)$. However, this “if and only if” relationship only holds if the treatment effect is constant over time. If the treatment effect varies over time, as is often likely to be the

case, then one might have that $E_u(\Delta Y|\Delta D)$ is linear but the TWFE estimator is not consistent for the ATT or for a convex combination of effects.

7.3.4.3 Testing the constant-and-linear-effect assumption

Propose a simple method to test that $E_u(\Delta Y|D_2)$ is linear when D_2 takes a finite number of values.

A non-parametric and tuning-parameter-free test of the linearity of $E_u(\Delta Y|D_2)$, when D_2 takes a finite number of values. Assume that D_2 takes K values. If $K = 2$, $E_u(\Delta Y|D_2)$ is necessarily linear, with

$$\alpha_1 = \frac{E_u(\Delta Y|D_2 = \bar{d}) - E_u(\Delta Y|D_2 = \underline{d})}{\bar{d} - \underline{d}}$$

and $\alpha_0 = E_u(\Delta Y|D_2 = \underline{d}) - \underline{d}\alpha_1$, where $\underline{d} < \bar{d}$ denote the two values D_2 can take. Then, there is no room for testability. This is not an issue if 0 is one of the two values D_2 can take, as then we are back to a classical DID design, where one can show that $\hat{\beta}^{\text{fe}}$ is unbiased for the ATT under Assumption PTNB. On the other hand, if $K = 2$ and D_2 takes two values different from 0, lack of testability is an issue. If $K > 2$, to test that $E_u(\Delta Y|D_2)$ is linear one can just regress ΔY_g on a constant, $D_{g,2}$, $D_{g,2}^2$, ... $D_{g,2}^{K-1}$, and test that the coefficients on $D_{g,2}^2$, ... $D_{g,2}^{K-1}$ are all equal to zero.

A non-parametric and tuning-parameter-free test of the linearity of $E_u(\Delta Y|D_2)$, when D_2 is continuous. When D_2 is continuous, one can rely on Stute (1997) and Stute, Manteiga and Quindimil (1998) to test the linearity of $E_u(\Delta Y|D_2)$. The test therein has desirable properties: it has asymptotically correct size, is consistent under any fixed alternative, and has non-trivial power against local alternatives converging towards the null at the $1/G^{1/2}$ rate. Moreover, it follows from Corollary 1 in de Chaisemartin and D'Haultfœuille (2024) that under the null, inference on β^{fe} conditional on not rejecting the Stute pre-test is conservative. Thus, under the

null of parallel trends and constant and linear effects, the pre-testing rule we propose cannot make post-test inference liberal.

Some details on the Stute test.* Under the null hypothesis that $E_u[\Delta Y|D_2]$ is linear, then $(\hat{\varepsilon}_{\text{lin},g})_{g=1,\dots,G}$, the residuals of the linear regression of ΔY_g on $D_{g,2}$, should not be correlated with any function of D_2 . Then, consider the so-called cusum process of the residuals:

$$c_G(d) := G^{-1/2} \sum_{g=1}^G \mathbb{1}\{D_{g,2} \leq d\} \hat{\varepsilon}_{\text{lin},g}.$$

Stute (1997) shows that under the null hypothesis, c_G , as a process indexed by d , converges to a Gaussian process. On the other hand, under the alternative, $c_G(d)$ tends to infinity for some d . Then, one can for instance consider the following Cramer-von Mises test statistics based on $c_G(d)$:

$$\text{CVM} = \frac{1}{G} \sum_{g=1}^G c_G^2(D_g).$$

The limiting distribution of CVM under the null is complicated, but Stute et al. (1998) show that one can approximate it using the wild bootstrap. Specifically, consider i.i.d. random variables $(V_g)_{g=1,\dots,G}$ with $E_u[V_g] = 0$, $E_u[V_g^2] = E_u[V_g^3] = 1$.⁶ Then, let $\hat{\varepsilon}_{\text{lin},g}^* = V_g \hat{\varepsilon}_{\text{lin},g}$ and

$$\Delta Y_g^* = \hat{\beta}_0 + \Delta D_g \hat{\beta}^{\text{fe}} + \hat{\varepsilon}_{\text{lin},g}^*.$$

Then, letting CVM^* denote the bootstrap counterpart of CVM based on the sample $(D_g, \Delta Y_g^*)_{g=1,\dots,G}$, Stute et al. (1998) show that as $G \rightarrow \infty$, the conditional distribution of CVM^* tends to the limiting distribution of CVM under the null.

A non-parametric and tuning-parameter-free pre-trends test of Assumption PTNB.

Assumption PTNB is a mean-independence condition. If the data contains another pre-period $t = 0$ where groups are all untreated, as in period $t = 1$, regressing $Y_{g,1} - Y_{g,0}$ on $D_{g,2}$ is a pre-trends test of a condition weaker than Assumption PTNB, namely that $\Delta Y(0)$ is uncorrelated with D_2 . A pre-trends test of Assumption PTNB should test the null that $Y_1 - Y_0$ is mean independent of D_2 , something that can be achieved with another version of the Stute test.

⁶For instance, one can use the standard two-point distribution: $V_g = (1 + \sqrt{5})/2$ with probability $(\sqrt{5} - 1)/(2\sqrt{5})$, $V_g = (1 - \sqrt{5})/2$ otherwise.

Computation: Stata and R commands to implement the Stute test. The `stute_test` Stata (see de Chaisemartin, Ciccia, D'Haultfœuille, Knau and Sow, 2024) and R (see de Chaisemartin, Ciccia, D'Haultfœuille, Knau and Sow, 2024d) commands implement the Stute test. The syntax of the Stata command is:

```
stute_test ΔY D2.
```

By default, the command tests that the conditional mean of ΔY given D_2 is linear. To test that ΔY and D_2 are mean independent, as one would do in a pre-trends test, one needs to specify the `order(0)` option.

7.3.4.4 Testing the null that there are quasi-stayers

Without stayers, the interpretation of the linearity tests crucially depends on whether there are quasi-stayers, namely groups with a period-two treatment very close to zero. Therefore, de Chaisemartin, Ciccia, D'Haultfœuille and Knau (2024) propose tests of the null hypothesis that there are quasi-stayers. One of their test statistics is $QS = D_{(1),2}/(D_{(2),2} - D_{(1),2})$, where $D_{(1),2} \leq \dots \leq D_{(G),2}$ denotes the order statistic of $(D_{g,2})_{g=1,\dots,G}$. The critical region is $W_\alpha := \{QS > 1/\alpha - 1\}$. Intuitively, we reject the null if the distance between $D_{(1),2}$ and 0 is more than $1/\alpha - 1$ times larger than the distance between $D_{(2),2}$ and $D_{(1),2}$: then, $D_{(1),2}$ is too far from zero for it to be plausible that $D_{(1),2}$ would converge to 0 if the sample size were to grow to infinity. That test is asymptotically valid if the density of D_2 is strictly positive at 0, and it has nontrivial local power.⁷

7.3.4.5 Application to the NTR-gap example

A test that $E_u(\Delta Y_t|D_2)$ is linear for all t is not rejected. Using the `pierce_schott_didtextbook` dataset, run a Stute test of linearity of the conditional expectation of `delta2001` given `ntrgap`. Do you reject linearity?

⁷If one worries about the positive density assumption, one can use instead the statistic $D_{(1),2}^2/(D_{(2),2}^2 - D_{(1),2}^2)$. Then, the test remains valid if the density of D_2 vanishes at 0, provided its derivative is strictly positive at 0.

```
stute_test delta2001 ntrgap
```

The test's p-value is equal to 0.04, so we reject the linearity assumption in 2001. We proceed similarly to test that the conditional expectations of `delta2002`, `delta2004`, and `delta2005` given `ntrgap` are linear. P-values are respectively equal to 0.13, 0.48, and 0.72. By running:

```
preserve
```

```
reshape long delta deltaintrend, i(indusid) j(year)
stute_test delta ntrgap indusid year if year>=2001, seed(1)
restore,
```

one can perform a joint test pooling the four years together. The test is not rejected (p-value =0.54). The interpretation of those linearity tests depends on whether there are quasi-stayers. $D_{2,(1)} / (D_{2,(2)} - D_{2,(1)}) = 6.15$, so the null that there are quasi-stayers is not rejected (p-value=0.14). It seems that there are quasi-stayers in this application. Then, there is an “if and only if” between the null of the Stute test and the constant-and-linear effect assumption.

Pre-trend tests of Assumption PTNB are rejected. Importantly, remember that it is only if Assumption PTNB holds that we have an “if and only if” relationship between the constant-and-linear-effect assumption in (7.9) and linearity of $E_u(\Delta Y|D_2)$. We now run pre-trend tests of Assumption PTNB. Using the `pierce_schott_didtextbook` dataset, regress `delta1999`, `delta1998`, and `delta1997` on `ntrgap`. Do you reject the null that industries' employment evolutions prior to the treatment are uncorrelated with the NTR-gap treatment?

For `delta1999`, we run:

```
reg delta1999 ntrgap, vce(hc2, dfadjust)
```

The coefficient on `ntrgap` is significant at the 5% level ($\hat{\beta}^{\text{fe}} = 0.058$, 95% CI=[0.001, 0.115]). We proceed similarly with `delta1998` and `delta1997`, and we also find significant coefficients (in 1998: $\hat{\beta}^{\text{fe}} = 0.141$, 95% CI=[0.030, 0.252]; in 1997: $\hat{\beta}^{\text{fe}} = 0.163$, 95% CI=[0.018, 0.307]). Thus there is clear evidence of differential pre-trends: Assumption PTNB does not seem plausible in

this application.⁸ Note however that the pre-trends TWFE estimators are substantially smaller than the actual TWFE estimators, so it does not seem that pre-trends can fully account for the estimated treatment effects.

Pre-trend tests of Assumption PTNB are no longer rejected when industry-specific linear trends are controlled for. The pre-trend estimators increase as we look at employment evolutions over a longer horizon. Then, the violation of Assumption PTNB in this data might be due to industry-specific linear trends that are correlated with industries' NTR gaps, and Assumption PTNB might be plausible when such linear trends are controlled for. Accordingly, we replace Assumption PTNB by

$$E_u(Y_{g,t}(0) - Y_{g,2000}(0) - (t - 2000) \times (Y_{g,2000}(0) - Y_{g,1999}(0))|D_{g,2001}) = \gamma_t. \quad (7.12)$$

$Y_{g,2000}(0) - Y_{g,1999}(0)$ captures industry g 's linear trend without treatment. Then, $Y_{g,t}(0) - Y_{g,2000}(0) - (t - 2000) \times (Y_{g,2000}(0) - Y_{g,1999}(0))$ is g 's deviation from its linear trend from 2000 to t . Therefore, (7.12) requires that industries' deviations from their linear trend are mean-independent from the NTR-gap treatment, which is similar to Assumption CDLT in Chapter 4. Under this assumption, treatment effect estimators can be obtained by regressing, for $t \in \{2001, 2002, 2004, 2005\}$, $Y_{g,t} - Y_{g,2000} - (t - 2000) \times (Y_{g,2000} - Y_{g,1999})$ on the NTR-gap treatment. Similarly, to test that prior to 2001, industries' deviations from their linear trends are unrelated to the NTR-gap treatment, one can regress, for $t \in \{1998, 1997\}$, $Y_{g,t} - Y_{g,1999} - (t - 1999) \times (Y_{g,2000} - Y_{g,1999})$ on $D_{g,2001}$. We now implement this test. For 1998, run:

```
reg delta1998 ntrgap, vce(hc2, dfadjust)
```

Interpret the results: do we still have differential pre-trends once industry-specific linear trends are accounted for?

⁸Those findings are at odds with those from Figure 4 in Pierce and Schott (2016). Therein, the authors run the exact same pre-trend tests as we do, but on the proprietary dataset they use for most of their analysis, and they do not find statistically significant pre-trends. In the dataset we use, industries are at the four-digit ISIC level, while in the proprietary dataset industries are defined at a more disaggregated level. It seems that while disaggregated NTR gaps are uncorrelated with industries' employment pre-trends, the aggregated variables are correlated, a version of the so-called "ecological inference problem".

The coefficient on `ntrgap` is small and insignificant ($\hat{\beta}^{\text{fe}} = -0.025$, 95% CI=[−0.093, 0.044]). We proceed similarly for 1997, and we also find a small and insignificant coefficient ($\hat{\beta}^{\text{fe}} = -0.046$, 95% CI=[−0.136, 0.044]). Thus, it seems we no longer have differential pre-trends once industry-specific linear trends are accounted for. However, those pre-trend tests are testing whether $Y_{g,t} - Y_{g,1999} - (t - 1999) \times (Y_{g,2000} - Y_{g,1999})$ and $D_{g,2001}$ are uncorrelated for $t \in \{1998, 1997\}$, while (7.12) is a mean-independence assumption. To assess whether $Y_{g,t} - Y_{g,1999} - (t - 1999) \times (Y_{g,2000} - Y_{g,1999})$ is mean independent of $D_{g,2001}$ for $t \in \{1998, 1997\}$, run:

```
stute_test deltalintrend1998 ntrgap, order(0) seed(1)
stute_test deltalintrend1997 ntrgap, order(0) seed(1)
```

Interpret the results: do we reject the null that `deltalintrend1998` and `deltalintrend1997` are mean independent of `ntrgap`?

Those tests are not rejected (p-value=0.30 and p-value=0.51, respectively). A joint test pooling the two years together is also not rejected (p-value =0.47). This lends plausibility to (7.12).

The Stute test is still not rejected with industry-specific linear trends. As it seems that Assumption PTNB fails but (7.12) holds, we run the Stute test of linearity again, on $Y_{g,t} - Y_{g,2000} - (t - 2000) \times (Y_{g,2000} - Y_{g,1999})$, for $t \in \{2001, 2002, 2004, 2005\}$. None of the four tests is rejected at the 5% level. A joint test is also not rejected (p-value =0.40).

TWFE estimators are smaller with than without industry-specific linear trends, but they are still negative and marginally significant. Overall, our tests suggest that TWFE estimators with industry-specific linear trends might be reliable in this application, or at least there is no strong, detectable indication that they are not. Those estimators are shown in Table 7.1 below, together with all the other estimators and tests computed in this replication. While TWFE estimators with industry-specific linear trends are smaller and less significant than TWFE estimators without linear trends, the estimated effect in 2004 is significant at the 5% level, and that in 2002 is significant at the 10% level.

Table 7.1: Effects, on US employment, of eliminating potential tariffs spikes on imports from China

	Panel A: Effects			
	2001	2002	2004	2005
$\hat{\beta}^{\text{fe}}$	-0.06	-0.26	-0.54	-0.53
95% CI	[-0.14, 0.02]	[-0.41, -0.11]	[-0.85, -0.23]	[-0.87, -0.19]
P-value Stute test of linearity	0.04	0.13	0.48	0.72
P-value joint test of linearity			0.54	
	Panel B: Pre-trend estimators			
	1999	1998	1997	
$\hat{\beta}^{\text{fe}}$	0.06	0.14	0.16	
95% CI	[0.00, 0.11]	[0.03, 0.25]	[0.02, 0.31]	
	Panel C: Pre-trend estimators with industry-specific linear trends			
	1998	1997		
$\hat{\beta}^{\text{fe}}$	-0.02	-0.05		
95% CI	[-0.093, 0.044]	[-0.136, 0.044]		
P-value Stute test of mean indep.	0.30	0.51		
P-value joint test of mean indep.		0.47		
	Panel D: Effects with industry-specific linear trends			
	2001	2002	2004	2005
$\hat{\beta}^{\text{fe}}$	-0.00	-0.14	-0.31	-0.24
95% CI	[-0.08, 0.08]	[-0.30, 0.01]	[-0.62, 0.00]	[-0.58, 0.10]
P-value Stute test of linearity	0.38	0.06	0.38	0.53
P-value joint test of linearity			0.40	
Observations	103	103	103	103

Notes: This table shows estimated effects, on US employment, of eliminating potential tariffs spikes on imports from China (Panels A and D), and pre-trend estimators (Panels B and C). Estimation uses log employment data for a panel of 103 US industries from 1997 to 2002 and from 2004 to 2005. In Panels A and B, TWFE regressions are shown. In Panels C and D, TWFE regressions with industry-specific linear trends are shown. Some panels also show p-values of Stute tests of mean independence and linearity.

7.4 Heterogeneity-robust estimators

Motivation. There are at least three instances where one may prefer avoiding the TWFE estimator, even if pre-trend tests of Assumption PTNB are not rejected. First, the test of the constant-and-linear-effect assumption in (7.9) may be rejected. Second, even when that test is not rejected, one may still worry that the power of the test is low. Third, even when that test is not rejected and one is not concerned with its power, one may be in a design without stayers or quasi-stayers, in which case (7.9) could fail even when its testable implication holds. In this section, we review heterogeneity-robust estimators one could then use. First, we restrict attention to HADs with stayers or quasi-stayers, before considering HADs without stayers or quasi-stayers.

A fundamental decomposition of the conditional expectation of group's outcome evolutions under Assumption PTNB. In Design HAD, under Assumption PTNB,

$$\begin{aligned}
 E(\Delta Y|D_2) &= E(Y_2(D_2) - Y_1(0)|D_2) \\
 &= E(Y_2(0) - Y_1(0)|D_2) + E(Y_2(D_2) - Y_2(0)|D_2) \\
 &= E(\Delta Y(0)|D_2) + D_2 E(\text{TE}_2|D_2) \\
 &= \gamma_2 + D_2 \text{CAS}(D_2) :
 \end{aligned} \tag{7.13}$$

$E(\Delta Y|D_2)$ can be decomposed into groups' counterfactual outcome evolutions without treatment γ_2 , and $D_2 \text{CAS}(D_2)$. It directly follows from this decomposition that if γ_2 is identified, then the CAS is identified, and therefore the ATT and WATT are also identified. With that in mind, we now propose heterogeneity-robust estimators of the CAS, ATT, and WATT, depending on whether we have stayers, quasi-stayers, or neither stayers nor quasi-stayers.

7.4.1 Designs with stayers or quasi-stayers

7.4.1.1 Identification.

Theorem 15 Suppose that we are in Design HAD' and Assumptions NA, ND and PTNB hold. Then,

1. For all $d_2 > 0$ in the support of D_2 ,

$$CAS(d_2) = \frac{E_u[\Delta Y|D_2 = d_2] - E_u[\Delta Y|D_2 = 0]}{d_2}. \quad (7.14)$$

2. If there exists a strictly positive real number η such that $P_u(0 < D_2 < \eta) = 0$,

$$ATT = E_u \left[\frac{\Delta Y - E_u[\Delta Y|D_2 = 0]}{D_2} \middle| D_2 > 0 \right]. \quad (7.15)$$

3.

$$WATT = \frac{E_u[\Delta Y|D_2 > 0] - E_u[\Delta Y|D_2 = 0]}{E_u[D_2|D_2 > 0]}. \quad (7.16)$$

Theorem 15 readily follows from (7.13), and from the fact that with stayers or quasi-stayers, $E_u[\Delta Y|D_2 = 0]$, the outcome evolution of untreated groups, identifies γ_2 , the counterfactual outcome evolution that treated groups would have experienced without treatment. Then, $d_2 \mapsto CAS(d_2)$, the ATT, and the WATT are identified by DID estimands comparing the outcome evolutions of treated and untreated groups, and scaling that comparison by the treatment of treated groups. [Intuitively, why is it that the estimands in Theorem 15 identify average effects under Assumption PTNB alone, unlike \$\beta^{fe}\$?](#)

β^{fe} may identify a non-convex combination of effects under Assumption PTNB alone because it leverages forbidden comparisons of the outcome evolutions of more- and less-treated groups. Instead, the estimands in Theorem 15 only compare treated and untreated groups: weakly treated groups are not used as control groups by those estimands.

Identification of ATT with quasi-stayers.* Importantly, for ATT the identification result in Theorem 15 assumes that there are no quasi-stayers: the dose of treatment received by treated groups should be bounded below and cannot be arbitrarily close to zero. Otherwise, under weak conditions one can show that

$$E_u \left[\left| \frac{\Delta Y - E_u[\Delta Y|D_2 = 0]}{D_2} \right| \middle| D_2 > 0 \right] = +\infty,$$

meaning that the estimand in (7.15) is not well defined. Intuitively, this is due to the fact that with quasi-stayers, D_2 can be arbitrarily close to zero while $\Delta Y - E_u[\Delta Y|D_2 = 0]$ may not be close to zero. One can show that with quasi-stayers, the ATT is identified by

$$\lim_{\eta \rightarrow 0} E_u \left[\frac{\Delta Y - E_u[\Delta Y|D_2 = 0]}{D_2} \middle| D_2 > \eta \right].$$

The estimand in the previous display is a limiting estimand. It trims quasi-stayers from the estimand in (7.15), and lets the trimming go to zero, as in Graham and Powell (2012), who consider a related estimand with quasi-stayers. Accordingly, with quasi-stayers the ATT is irregularly identified by a limiting estimand. Then, de Chaisemartin et al. (2022) conjecture that the ATT cannot be estimated at the \sqrt{n} -rate, as in Graham and Powell (2012), and as is often the case with target parameters identified by limiting estimands. Due to this limitation, we will not consider estimation of the ATT with quasi stayers.

7.4.1.2 Estimation with stayers ($P_u(D_2 = 0) > 0$).

Estimation of $d_2 \mapsto \text{CAS}(d_2)$, when the treatment takes a small number of values. If D_2 takes a small number of values relative to the number of groups, to estimate $d_2 \mapsto \text{CAS}(d_2)$ one can follow Theorem 15, replacing expectations by sample averages:

$$\widehat{\text{CAS}}(d_2) := \frac{\frac{1}{G_{1,d_2}} \sum_{g:D_{g,2}=d_2} \Delta Y_g - \frac{1}{G_0} \sum_{g:D_{g,2}=0} \Delta Y_g}{d_2},$$

where G_{1,d_2} denote the number of treated groups such that $D_{g,2} = d_2$. The numerator of $\widehat{\text{CAS}}(d_2)$ is numerically equivalent to the coefficient on $1\{D_{g,2} = d_2\}$ in a linear regression of ΔY_g on indicators for all the strictly positive values that $D_{g,2}$ can take.

Estimation of $d_2 \mapsto \text{CAS}(d_2)$, when the treatment takes a large number of values. D_2 might take a large number of values, or could even be continuously distributed conditional on $D_2 > 0$. Then, estimating $d_2 \mapsto \text{CAS}(d_2)$ requires estimating $E_u[\Delta Y|D_2 = d_2]$, a univariate conditional expectation. To conduct that estimation, Callaway et al. (2021) adapt a non-parametric estimator proposed by Chen, Christensen and Kankanala (2024). Importantly, that estimator is fully data-driven, and does not require that the user choose some tuning parameters.

Estimation of the ATT, if there are no quasi-stayers. To estimate the ATT, if there are no quasi-stayers, one can follow Theorem 15, and use

$$\widehat{\text{ATT}}^s := \frac{1}{G_1} \sum_{g:D_{g,2}>0} \frac{\Delta Y_g - \frac{1}{G_0} \sum_{g:D_{g,2}=0} \Delta Y_g}{D_{g,2}}.$$

de Chaisemartin et al. (2022) show that this estimator is \sqrt{G} -consistent and asymptotically normal when $G \rightarrow +\infty$, and derive its asymptotic variance.

Estimation of the WATT. To estimate the WATT, one can also follow Theorem 15, and use

$$\widehat{\text{WATT}}^s := \frac{\frac{1}{G_1} \sum_{g:D_{g,2}>0} \Delta Y_g - \frac{1}{G_0} \sum_{g:D_{g,2}=0} \Delta Y_g}{\frac{1}{G_1} \sum_{g:D_{g,2}>0} D_{g,2}}.$$

de Chaisemartin et al. (2022) show that this estimator is \sqrt{G} -consistent and asymptotically normal when $G \rightarrow +\infty$, and derive its asymptotic variance. [Find a binary instrumental variable \$Z_g\$ such that \$\widehat{\text{WATT}}^s\$ is equal to the coefficient on \$D_{g,2}\$ in a 2SLS regression of \$\Delta Y_g\$ on a constant and \$D_{g,2}\$, using \$Z_g\$ as the instrument for \$D_{g,2}\$.](#)

The coefficient on the treatment in a 2SLS regression with one binary instrument Z_g and no controls is the Wald ratio

$$\frac{\frac{1}{\#\{g:Z_g=1\}} \sum_{g:Z_g=1} \Delta Y_g - \frac{1}{\#\{g:Z_g=0\}} \sum_{g:Z_g=0} \Delta Y_g}{\frac{1}{\#\{g:Z_g=1\}} \sum_{g:Z_g=1} D_{g,2} - \frac{1}{\#\{g:Z_g=0\}} \sum_{g:Z_g=0} D_{g,2}},$$

where for any set A , $\#A$ denotes the number of elements of A , i.e. its cardinality. With $Z_g = 1\{D_{g,2} > 0\}$, the previous display is equal to $\widehat{\text{WATT}}^s$. Thus, to compute $\widehat{\text{WATT}}^s$, we can run a 2SLS regression of ΔY_g on a constant and $D_{g,2}$, using $1\{D_{g,2} > 0\}$ as the instrument for $D_{g,2}$. Note that when control variables are included in this regression, it is no longer guaranteed to unbiasedly estimate the WATT, and it may not even estimate a convex combination of slopes.

$\widehat{\text{WATT}}^s$ can be much more precise than $\widehat{\text{ATT}}^s$. Proposition 1 in de Chaisemartin et al. (2022) shows that under some assumptions, the asymptotic variance of $\widehat{\text{ATT}}^s$ is strictly larger

than that of $\widehat{\text{WATT}}^s$. The difference can be very large: in their empirical application, the variance of $\widehat{\text{ATT}}^s$ is more than seven times larger than that of $\widehat{\text{WATT}}^s$. Then, there may be cases where even if one's target is the ATT, $\widehat{\text{WATT}}^s$ may have a lower mean-squared error for that target than $\widehat{\text{ATT}}^s$.

Intuition for why $\widehat{\text{WATT}}^s$ is more precise than $\widehat{\text{ATT}}^s$.* One has

$$\widehat{\text{WATT}}^s = \frac{1}{G_1} \sum_{g:D_{g,2}>0} \frac{D_{g,2}}{\frac{1}{G_1} \sum_{g':D_{g',2}>0} D_{g',2}} \frac{\Delta Y_g - \frac{1}{G_0} \sum_{g:D_{g,2}=0} \Delta Y_g}{D_{g,2}}.$$

Compare the previous display to the definition of $\widehat{\text{ATT}}^s$. Intuitively, why is it that $\widehat{\text{WATT}}^s$ is more precise than $\widehat{\text{ATT}}^s$?

Like $\widehat{\text{ATT}}^s$, $\widehat{\text{WATT}}^s$ is an average of the group-specific-slope estimators

$$\widehat{TE}_{g,2} := \frac{\Delta Y_g - \frac{1}{G_0} \sum_{g:D_{g,2}=0} \Delta Y_g}{D_{g,2}},$$

but unlike $\widehat{\text{ATT}}^s$, $\widehat{\text{WATT}}^s$ downweights $\widehat{TE}_{g,2}$ for “weakly-treated” groups ($D_{g,2} < \frac{1}{G_1} \sum_{g':D_{g',2}>0} D_{g',2}$), and it upweights $\widehat{TE}_{g,2}$ for strongly-treated groups. One has that

$$\begin{aligned} \widehat{TE}_{g,2} &= \frac{Y_{g,2}(D_{g,2}) - Y_{g,1}(0) - \frac{1}{G_0} \sum_{g:D_{g,2}=0} \Delta Y_g}{D_{g,2}} \\ &= TE_{g,2} + \frac{\Delta Y_g(0) - \frac{1}{G_0} \sum_{g:D_{g,2}=0} \Delta Y_g(0)}{D_{g,2}}, \end{aligned}$$

where the second equality follows from adding and subtracting $Y_{g,2}(0)$. Then, if $TE_{g,2} = \delta$ for some real number δ and $V_u(\Delta Y_g(0)|D_{g,2}) = \sigma^2$ for some real number σ^2 ,

$$\begin{aligned} V(\widehat{TE}_{g,2}|D_{g,2}) &= V\left(TE_{g,2} + \frac{\Delta Y_g(0) - \frac{1}{G_0} \sum_{g:D_{g,2}=0} \Delta Y_g(0)}{D_{g,2}} \middle| D_{g,2}\right) \\ &= \frac{\sigma^2(1 + 1/G_0)}{D_{g,2}^2}. \end{aligned}$$

Thus, the conditional variance of $\widehat{TE}_{g,2}$ is larger for weakly-treated than for strongly-treated groups, hence $\widehat{\text{WATT}}^s$'s greater precision.

Computation: Stata and R commands. The `did_multiplegt_stat` Stata (see de Chaisemartin, Ciccia, D'Haultfoeuille, Knau and Sow, 2024c) command can be used to estimate the ATT and the WATT, including in cases where the WATT cannot be estimated via a simple 2SLS regression because one wants to control for covariates. The syntax of the command is:

```
did_multiplegt_stat outcome groupid timeid treatment, estimator(as was) exact_match
```

We are not aware of a Stata or R command computing the estimator of $d_2 \mapsto \text{CAS}(d_2)$ proposed by Callaway et al. (2021) when the treatment takes a large number of values.

7.4.1.3 Estimation of the WATT without stayers but with quasi-stayers.

In this section, our target parameter is the WATT: to our knowledge, with no stayers but some quasi-stayers, only heterogeneity-robust estimators of the WATT have been proposed so far. [Without stayers, which part of the estimands identifying the WATT in Theorem 15 becomes difficult to estimate?](#)

Estimation problem. In Theorem 15, the estimand identifying the WATT compares treated groups' average outcome evolution to $E_u[\Delta Y|D_2 = 0]$, their counterfactual outcome evolution without treatment under Assumption PTNB. Without stayers, estimating $E_u[\Delta Y|D_2 = 0]$ is not straightforward: $P_u(D_2 = 0) = 0$, so no group in the sample is such that $D_{g,2} = 0$, and we cannot merely compute the sample average of the outcome evolutions of untreated groups.

A possible solution: use the outcome evolution of quasi-stayers to estimate groups' counterfactual trend without treatment. de Chaisemartin, Ciccia, D'Haultfoeuille and Knau (2024) propose to use “quasi-stayers”, namely observations with D_2 “close” to 0. Specifically, they propose to run a regression of ΔY on D_2 among the subsample with D_2 lower than some bandwidth h , and use the regression's intercept to estimate $E_u[\Delta Y|D_2 = 0]$. Intuitively, as the bandwidth h increases, the bias of the estimator increases, as it uses groups that received

a higher treatment dose to infer groups' counterfactual trend without treatment. At the same time, as h increases, the variance of the estimator decreases, as it estimates groups' counterfactual trend without treatment out of a larger sample. This suggests that there might exist an optimal bandwidth, that trades off the estimator's bias and variance optimally. de Chaisemartin, Ciccia, D'Haultfoeuille and Knau (2024) leverage results from the regression discontinuity designs (RDDs) and non-parametric estimation literature (see Imbens and Kalyanaraman, 2012; Calonico, Cattaneo and Titiunik, 2014; Calonico, Cattaneo and Farrell, 2018), to propose an optimal bandwidth that minimizes an asymptotic approximation of the estimator's mean squared-error, and a robust confidence interval accounting for the estimator's first-order bias. The resulting estimator of the WATT, $\widehat{\text{WATT}}_{\widehat{h}_G^*}^{qs}$, converges at the $G^{2/5}$ rate, the standard univariate non-parametric rate. With stayers, the estimator of the WATT converges at the faster $G^{1/2}$ rate, so moving from a design with stayers to a design with quasi-stayers comes with a precision cost. $\widehat{\beta}^{\text{fe}}$ also converges at the faster $G^{1/2}$ rate, so moving from $\widehat{\beta}^{\text{fe}}$ to the heterogeneity-robust estimator $\widehat{\text{WATT}}_{\widehat{h}_G^*}^{qs}$ also comes with a precision cost. This is a further reason why, in designs without stayers but with quasi-stayers, researchers might want to run the test of the constant-and-linear-effect assumption discussed in the previous section, to ensure that using an heterogeneity-robust estimator is warranted.

Estimator's definition.* We define the following estimators, indexed by the bandwidth h :

$$\widehat{\text{WATT}}_h^{qs} = \frac{\frac{1}{G} \sum_{g=1}^G \Delta Y_g - \widehat{\gamma}_h}{\frac{1}{G} \sum_{g=1}^G D_{g,2}},$$

with $\widehat{\gamma}_h$ the intercept in the local linear regression of ΔY_g on $D_{g,2}$, weighting observations by $k(D_{g,2}/h)/h$, for a kernel function k and a bandwidth $h > 0$.

Estimator's asymptotic distribution.* Let $m(d_2) = E_u(\Delta Y | D_2 = d_2)$. One can derive the asymptotic behavior of $\widehat{\text{WATT}}_h^{qs}$ under the conditions below:

Assumption RC (*Regularity conditions*)

1. *The cumulative distribution function of D_2 is differentiable at 0, with derivative denoted by $f_{D_2}(0)$. Moreover, $f_{D_2}(0) > 0$.*
2. *m , defined on $\text{Supp}(D_2)$, is twice differentiable at $d = 0$.*

3. $\sigma^2(d) := V_u(\Delta Y|D_2 = d)$, defined on $\text{Supp}(D_2)$, is continuous at 0 and $\sigma^2(0) > 0$.

4. k is bounded and has bounded support.

5. As $G \rightarrow \infty$, the bandwidth h_G satisfies $h_G \rightarrow 0$ and $Gh_G \rightarrow \infty$.

Assumption RC imposes standard regularity conditions on the distribution of D_2 , $E_u(\Delta Y|D_2)$, the kernel function and the bandwidth. We also introduce the following notation. Let $\kappa_j = \int_0^\infty t^j k(t) dt$ for $j \in \mathbb{N}$ and

$$\begin{aligned} k^*(t) &= \frac{\kappa_2 - \kappa_1 t}{\kappa_0 \kappa_2 - \kappa_1^2} k(t), \\ C &= \frac{\kappa_2^2 - \kappa_1 \kappa_3}{\kappa_0 \kappa_2 - \kappa_1^2}. \end{aligned}$$

Since $\sum_{g=1}^G \Delta Y_g/G$ and $\sum_{g=1}^G D_{g,2}/G$ are root- G consistent, their randomness is negligible compared to that of $\hat{\gamma}_h$. Thus,

$$G^{2/5} \left(\widehat{\text{WATT}}_{h_G}^{qs} - \text{WATT} \right) = G^{2/5} \frac{\widehat{\gamma}_{h_G} - m(0)}{E_u[D_{g,2}]} + o_P(1).$$

Then, following the exact same reasoning as in, say, Imbens and Kalyanaraman (2012), we obtain that

$$\sqrt{Gh_G} \left(\widehat{\text{WATT}}_{h_G}^{qs} - \text{WATT} - h_G^2 \frac{Cm''(0)}{2E_u[D_{g,2}(0)]} \right) \xrightarrow{d} \mathcal{N} \left(0, \frac{\sigma^2(0) \int_0^\infty k^*(u)^2 du}{E_u[D_{g,2}(0)]^2 f_{D_2}(0)} \right). \quad (7.17)$$

The fastest rate of convergence is obtained with $G^{1/5}h_G \rightarrow c > 0$, in which case

$$G^{2/5} \left(\widehat{\text{WATT}}_{h_G}^{qs} - \text{WATT} \right) \xrightarrow{d} \mathcal{N} \left(\frac{c^2 Cm''(0)}{2E_u[D_{g,2}(0)]}, \frac{\sigma^2(0) \int_0^\infty k^*(u)^2 du}{c E_u[D_{g,2}(0)]^2 f_{D_2}(0)} \right). \quad (7.18)$$

Optimal bandwidth and robust confidence interval.* Based on (7.18), one can derive a so-called optimal bandwidth, which, as in RDDs (see Imbens and Kalyanaraman, 2012), minimizes the asymptotic mean squared error of $\widehat{\text{WATT}}_{h_G}^{qs}$. Then, inference on the WATT is not straightforward, because the asymptotic distribution of

$$\sqrt{Gh_G^*} (\widehat{\text{WATT}}_{\hat{h}_G^*}^{qs} - \text{WATT})$$

has a first-order bias that needs to be accounted for. However, the general approach for local-polynomial regressions in Calonico et al. (2018) can be applied here. de Chaisemartin, Ciccia, D'Haultfœuille and Knau (2024) rely on their results and on their software implementation (see Calonico, Cattaneo and Farrell, 2019) to:

1. estimate an optimal bandwidth \hat{h}_G^* ;
2. compute an estimator $\widehat{\gamma}_{\hat{h}_G^*}$ of $E_u[\Delta Y|D_2 = 0]$;
3. compute $\widehat{M}_{\hat{h}_G^*}$, an estimator of $\widehat{\gamma}_{\hat{h}_G^*}$'s first-order bias;
4. compute $\widehat{V}_{\hat{h}_G^*}$, an estimator of the variance of $\widehat{\gamma}_{\hat{h}_G^*} - \widehat{M}_{\hat{h}_G^*}$.

With those inputs, de Chaisemartin, Ciccia, D'Haultfœuille and Knau (2024) simply define their estimator of the WATT with quasi-stayers as

$$\widehat{\text{WATT}}_{\hat{h}_G^*}^{qs} = \frac{\frac{1}{G} \sum_{g=1}^G \Delta Y_g - \widehat{\gamma}_{\hat{h}_G^*}}{\frac{1}{G} \sum_{g=1}^G D_{g,2}},$$

and its bias-corrected confidence interval as

$$\left[\widehat{\text{WATT}}_{\hat{h}_G^*}^{qs} + \frac{\widehat{M}_{\hat{h}_G^*}}{\frac{1}{G} \sum_{g=1}^G D_{g,2}} \pm \frac{q_{1-\alpha/2} \sqrt{\widehat{V}_{\hat{h}_G^*}/(G\hat{h}_G^*)}}{\frac{1}{G} \sum_{g=1}^G D_{g,2}} \right], \quad (7.19)$$

where q_x denotes the quantile of order x of a standard normal distribution.

Computation: Stata and R commands to compute $\widehat{\text{WATT}}_{\hat{h}_G^*}^{qs}$ and confidence intervals for the WATT. The `did_had` Stata (see de Chaisemartin, Ciccia, D'Haultfœuille, Knau and Sow, 2024b) and R (see de Chaisemartin, Ciccia, D'Haultfœuille, Knau and Sow, 2024a) commands can be used to compute the estimators $\widehat{\text{WATT}}_{\hat{h}_G^*}^{qs}$ and confidence intervals for the WATT, in HADs with no stayers but some quasi-stayers. The syntax of the Stata command is:

```
did_had outcome groupid timeid treatment
```

Importantly, `did_had` heavily relies on the `nprobust` package of Calonico et al. (2019), which should be cited, together with Calonico et al. (2018), whenever `did_had` is used.

7.4.2 Designs without stayers or quasi-stayers

Identification problem. Without stayers or quasi-stayers, there is no group that receives a treatment dose of zero at period two, even in the super-population. Accordingly, $\gamma_2 = E_u[\Delta Y(0)]$ is not point identified under Assumption PTNB. Then, it follows from the fundamental decomposition in (7.13) that the CAS, ATT, and WATT are not point identified under Assumption PTNB alone: identifying those objects requires making additional assumptions.

Constant effect of the lowest treatment dose. Results above for designs with stayers or quasi-stayers can be extended to designs without stayers or quasi-stayers, at the expense of imposing the following constant-effect assumption. Hereafter, let $\underline{d} = \inf \text{Supp}(D_2)$ be the infimum of the support of the period-two treatment, so that $\underline{d} = 0$ in designs with stayers or quasi-stayers.

Assumption CELD (*Constant effect of lowest treatment dose*)

$$E [Y_2(\underline{d}) - Y_2(0)|D_2] = E [Y_2(\underline{d}) - Y_2(0)].$$

Assumption CELD requires that the effect of receiving the lowest treatment dose \underline{d} be mean independent of units' actual period-two dose D_2 . While strong, Assumption CELD is arguably less strong than $E_u[\text{TE}_2|D_2 = \underline{d}] = \text{ATT}$, the constant-and-linear-effect assumption in (7.9). [With data from a second pre-treatment period, can we use a pre-trends test to assess the plausibility of Assumption CELD?](#)

Contrary to Assumption PTNB, one cannot assess the plausibility of Assumption CELD via a pre-trends test: if the data contains another pre-period $t = 0$ where groups are all untreated, $Y_{g,1} - Y_{g,0}$ is an outcome evolution without treatment, which is not the period-one equivalent of $Y_{g,2}(\underline{d}) - Y_{g,2}(0)$, the effect of the lowest treatment dose.

Actual-versus-lowest-treatment slopes. Under Assumption CELD, instead of the actual-versus-no-treatment slopes TE_2 , we consider actual-versus-lowest-treatment slopes:

$$\text{TE}_{2,\underline{d}} := \frac{Y_2(D_2) - Y_2(\underline{d})}{D_2 - \underline{d}}.$$

Accordingly, we also define the following counterparts of the ATT and WATT:

$$\begin{aligned} \text{ATT}_{\underline{d}} &:= E [\text{TE}_{2,\underline{d}}] \\ \text{WATT}_{\underline{d}} &:= E \left[\frac{D_2 - \underline{d}}{E[D_2 - \underline{d}]} \text{TE}_{2,\underline{d}} \right]. \end{aligned}$$

Results on TWFE regressions without quasi-stayers, but assuming homogeneous effects of receiving the lowest treatment dose. Under Assumptions PTNB and CELD, we can obtain a decomposition, similar to that in (7.6), of β^{fe} as a weighted sum of $E(\text{TE}_{2,\underline{d}}|D_2)$ instead of $E(\text{TE}_2|D_2)$. Then, if

$$E(\text{TE}_{2,\underline{d}}|D_2) = \text{ATT}_{\underline{d}}, \quad (7.20)$$

$\beta^{\text{fe}} = \text{ATT}_{\underline{d}}$. The following theorem shows that under Assumptions PTNB and CELD, one has an equivalence between (7.20) and linearity of $E[\Delta Y|D_2]$: (7.20) is fully testable.

Theorem 16 *Suppose that Assumptions PTNB and CELD hold. In Design HAD, (7.20) holds if and only if $E(\Delta Y|D_2) = \beta_0 + \beta^{\text{fe}} D_2$.*

The proof is similar to that of Theorem 14 and thus omitted.

Identification and estimation of $\text{WATT}_{\underline{d}}$ under Assumption CELD. The following result mimicks Point 3 of Theorem 15 for designs without quasi-stayers:

Theorem 17 *Suppose that we are in Design HAD and Assumptions PTNB and CELD hold. Then,*

$$\text{WATT}_{\underline{d}} = \frac{E[\Delta Y] - E[\Delta Y|D_2 = \underline{d}]}{E[D_2 - \underline{d}]} \quad (7.21)$$

The proof is similar to that of Theorem 15 and thus omitted. If $P(D_2 = \underline{d}) > 0$, to estimate $\text{WATT}_{\underline{d}}$ we can use a similar estimator as the 2SLS WATT estimator introduced in Section 7.4.1.2: we can run a 2SLS regression of ΔY on D_2 using $1\{D_2 > \underline{d}\}$ as the instrument. Similarly, if $P(D_2 = \underline{d}) = 0$, to estimate $\text{WATT}_{\underline{d}}$ we can use a similar estimator as the WATT estimator with quasi-stayers introduced in Section 7.4.1.3, replacing D_2 by $D_2 - \underline{d}$ in the estimator's definition. We let $\widehat{\text{WATT}}_{\underline{d}, \widehat{h}_G^*}$ denote that estimator.

A condition under which the estimand in Theorem 17 identifies the sign of the WATT.* Under Assumption PTNB,

$$E[\Delta Y] - E[\Delta Y|D_2 = \underline{d}] = E(Y_2(D_2) - Y_2(0)) - E(Y_2(\underline{d}) - Y_2(0)|D_2 = \underline{d}).$$

Then, $\frac{E[\Delta Y] - E[\Delta Y|D_2 = \underline{d}]}{E[D_2 - \underline{d}]}$ has the same sign as $E(Y_2(D_2) - Y_2(0))$, and therefore as the WATT, if and only if the following condition holds:

$$\begin{aligned} E(Y_2(D_2) - Y_2(0)) \text{ and } E(Y_2(\underline{d}) - Y_2(0)|D_2 = \underline{d}) \text{ are of opposite signs,} \\ \text{or } |E(Y_2(D_2) - Y_2(0))| \geq |E(Y_2(\underline{d}) - Y_2(0)|D_2 = \underline{d})|. \end{aligned} \quad (7.22)$$

Then, the only scenario where the sign of $\frac{E[\Delta Y] - E[\Delta Y|D_2 = \underline{d}]}{E[D_2 - \underline{d}]}$ can differ from the sign of the WATT is if $E(Y_2(D_2) - Y_2(0))$ and $E(Y_2(\underline{d}) - Y_2(0)|D_2 = \underline{d})$ are of the same sign and $|E(Y_2(D_2) - Y_2(0))| < |E(Y_2(\underline{d}) - Y_2(0)|D_2 = \underline{d})|$. This second condition can only hold if the least treated groups have a larger response per dose of treatment than the average group, thus offsetting the fact that they receive a smaller treatment dose than the average group. Therefore, (7.22) is fairly plausible when \underline{d} is much lower than $E(D_2)$.

Alternative assumptions without stayers or quasi-stayers. If one is not willing to assume a constant effect of the lowest treatment dose, de Chaisemartin, Ciccia, D'Haultfœuille and Knau (2024) propose two alternative identifying assumptions. First, they show that the CAS, ATT, and WATT can be identified if one is ready to make a parametric functional-form assumption for $d_2 \mapsto \text{CAS}(d_2)$. Second, they show that if one is ready to assume that the CAS function is bounded, then the ATT and WATT can be bounded as well.

7.4.2.1 Application to the effect of having access to independent news on voting behavior

Design and data. Remember that in 1996, a new TV channel called NTV was introduced in Russia. At that time, NTV was the only TV channel in Russia not controlled by the government. Enikolopov et al. (2011) study the effect of having access to this independent news source on voting behavior, using voting outcomes for the 1938 Russian subregions in the 1995 and 1999 elections. After 1996, NTV's coverage rate is heterogeneous across regions: while a large fraction of the population receives it in urbanized regions, a smaller fraction receives it in more rural regions. Yet, in the region with the lowest exposure rate to NTV, this rate is still equal to 27%, so there is no unexposed (stayer) or almost unexposed (quasi-stayer) region. The authors define their treatment as $D_{g,t}$, the proportion of the population having access to NTV in region g and year t , hereafter referred to as the NTV exposure rate.

TWFE estimators. In their Table 3, Enikolopov et al. (2011) use $\hat{\beta}^{\text{fe}}$ to estimate NTV's effect on five outcomes: the share of the electorate voting for the SPS and Yabloko parties, two opposition parties supported by NTV; the share of the electorate voting for the KPRF and LDPR parties, two parties not supported by NTV; and electoral turnout. Specifically, they regress those outcomes on region FEs, an indicator for the 1999 election, and the NTV exposure rate in region g and period t . As this TWFE regression only has two time periods, $\hat{\beta}^{\text{fe}}$, the coefficient on the NTV exposure rate, is numerically equivalent to the coefficient on the NTV exposure rate in a regression of outcomes' first difference from 1995 to 1999 on a constant and $D_{g,2}$. $\hat{\beta}^{\text{fe}} = 6.65$ (s.e.= 1.40) for the SPS voting rate, and $\hat{\beta}^{\text{fe}} = 1.84$ (s.e.= 0.76) for the Yabloko voting rate. According to these regressions, increasing the NTV exposure rate from 0 to 100% increases the share of votes for the SPS and Yabloko opposition parties by 6.65 and 1.84 percentage points, respectively. $\hat{\beta}^{\text{fe}}$ is small and insignificant for the remaining three outcomes.

If one only makes a parallel-trends assumption, $\hat{\beta}^{\text{fe}}$ is very far from estimating a convex combination of effects. We use the `twowayfweights` Stata package to compute the weights in (7.6), a decomposition of the probability limit of $\hat{\beta}^{\text{fe}}$ that makes no further assumptions than the parallel-trends condition in Assumption PTNB. We find that $\hat{\beta}^{\text{fe}}$ estimates a weighted sum of the effects of NTV in 1999 in the 1,938 regions, where 918 estimated weights are strictly positive, while 1,020 are strictly negative. The negative weights sum to -2.26.

The test of the constant-and-linear-effect assumption in (7.9) is rejected for four outcomes out of five. In spite of its negative weights, $\hat{\beta}^{\text{fe}}$ is still consistent for the ATT if on top of Assumption PTNB, the constant-and-linear-effect assumption in (7.9) also holds. Then, $E_u(\Delta Y|D_2)$ should be linear, a null hypothesis that we test using the Stute test discussed above. As shown in Table 7.2 below, the test is rejected at the 1% level for three outcomes out of five, and at the 5% level for four outcomes out of five. Unfortunately, the data does not contain electoral outcomes for another election before the introduction of NTV than the 1995 one, so we cannot run a pre-trends test of Assumption PTNB.

The Stute test suggests that $\hat{\beta}^{\text{fe}}$ may not be reliable. Which heterogeneity-robust estimator can we use in this application?

Estimators assuming a constant effect of the lowest treatment dose are more noisy than $\hat{\beta}^{\text{fe}}$, and sometimes take implausible values. Under the assumption that the effect of raising the NTV exposure rate from 0 to $\underline{d} = 0.27$ is constant across subregions, we can estimate $\text{WATT}_{\underline{d}}$. As the NTV exposure rate is continuously distributed, $P(D_2 = \underline{d}) = 0$, so we use $\widehat{\text{WATT}}_{\underline{d}, \hat{h}_G^*}$, an estimator similar to the estimator with quasi-stayers in Section 7.4.1.3, replacing D_2 by $D_2 - \underline{d}$. The bottom panel of Table 7.2 shows $\widehat{\text{WATT}}_{\underline{d}, \hat{h}_G^*}$, its standard error, and its bias-corrected 95% confidence interval. For the SPS vote outcome, $\widehat{\text{WATT}}_{\underline{d}, \hat{h}_G^*}$ is close to $\hat{\beta}^{\text{fe}}$ but not significantly different from zero, because its standard error is almost three times larger. For the Yabloko vote outcome, $\widehat{\text{WATT}}_{\underline{d}, \hat{h}_G^*}$ is nine times larger than $\hat{\beta}^{\text{fe}}$. Its standard error is more than ten times larger than that of $\hat{\beta}^{\text{fe}}$ but it is still significantly different from zero. In view of the low nationwide voting rate for the Yabloko party (3.2 percentage points), the value of $\widehat{\text{WATT}}_{\underline{d}, \hat{h}_G^*}$ is implausibly large, which suggests that either Assumption PTNB or Assumption CELD fails. For the KPRF outcome, $\widehat{\text{WATT}}_{\underline{d}, \hat{h}_G^*}$ is much more negative than $\hat{\beta}^{\text{fe}}$, and it is significant, even though its standard error is almost four times larger. Finally, for the remaining two outcomes $\widehat{\text{WATT}}_{\underline{d}, \hat{h}_G^*}$ is insignificant, with a standard error much larger than $\hat{\beta}^{\text{fe}}$.

Table 7.2: The Effects of NTV on Voting Behavior

	SPS vote	Yabloko vote	KPRF vote	LDPR vote	Turnout
$\hat{\beta}^{\text{fe}}$	6.65	1.84	-2.20	1.18	-2.06
(s.e.)	(1.40)	(0.76)	(2.12)	(1.38)	(2.01)
P-value Stute test	0.014	0.006	0.112	<0.001	0.002
$\widehat{\text{WATT}}_{\underline{d}, \hat{h}_G^*}$	4.55	16.76	-10.45	-1.51	-13.63
(s.e.)	(5.00)	(8.52)	(7.96)	(9.69)	(7.91)
95% CI	[−1.51, 18.09]	[2.18, 35.59]	[−36.15, −4.96]	[−8.02, 29.98]	[−27.42, 3.58]
Observations	1,938	1,938	1,938	1,938	1,938

Notes: This table shows estimated effects of the exposure rate to independent information on voting outcomes in Russia, using voting data for 1,938 Russian subregions in the 1995 and 1999 elections. In the first line, effects are estimated using TWFE regressions. In the fourth line, effects are estimated using a parametric heterogeneity-robust estimator, relying on the assumption that the effect of the lowest treatment dose is constant across regions.

Conclusion. It is only when one uses the TWFE estimator that one can conclude that access to NTV increases votes for opposition parties. That estimator crucially relies on Assumption PTNB and on (7.11), two assumptions which cannot be tested. However, the stronger condition (7.9), which is testable together with Assumption PTNB, is strongly rejected by our Stute test, thus suggesting that treatment effects are indeed heterogeneous in this application, or that the parallel-trends condition in Assumption PTNB fails.

7.5 Appendix*

7.5.1 Proof of Theorem 13

In Design HAD,

$$\begin{aligned}
 \Delta Y_g &= Y_{g,2} - Y_{g,1} \\
 &= Y_{g,2}(D_{g,2}) - Y_{g,1}(0) \\
 &= Y_{g,2}(D_{g,2}) - Y_{g,2}(0) + Y_{g,2}(0) - Y_{g,1}(0) \\
 &= D_{g,2} \frac{Y_{g,2}(D_{g,2}) - Y_{g,2}(0)}{D_{g,2}} + \Delta Y_g(0) \\
 &= D_{g,2} \text{TE}_{g,2} + \Delta Y_g(0),
 \end{aligned} \tag{7.23}$$

with the convention that $D_{g,2} \text{TE}_{g,2} = 0$ if $D_{g,2} = 0$. Justify the second equality of this derivation.

The second equality follows from the fact that $D_{g,1} = 0$ in an HAD. Plugging (7.23) into (7.5) and multiplying numerators and denominators by $\frac{1}{G}$,

$$\widehat{\beta}^{\text{fe}} = \frac{\frac{1}{G} \sum_{g=1}^G (D_{g,2} - D_{.,2}) \Delta Y_g(0)}{\frac{1}{G} \sum_{g=1}^G (D_{g,2} - D_{.,2})^2} + \frac{\frac{1}{G} \sum_{g=1}^G (D_{g,2} - D_{.,2}) D_{g,2} \text{TE}_{g,2}}{\frac{1}{G} \sum_{g=1}^G (D_{g,2} - D_{.,2})^2}. \tag{7.24}$$

It follows from (7.24), the weak law of large numbers and the continuous mapping theorem that

$$\beta^{\text{fe}} = \frac{\text{cov}(D_2, \Delta Y(0))}{V_u(D_2)} + \frac{E_u((D_2 - E_u(D_2)) D_2 \text{TE}_2)}{V_u(D_2)}. \tag{7.25}$$

Under Assumption PTNB,

$$\begin{aligned}
\text{cov}(D_2, \Delta Y(0)) &= E_u(D_2 \Delta Y(0)) - E_u(D_2)E_u(\Delta Y(0)) \\
&= E_u(D_2 E_u(\Delta Y(0)|D_2)) - E_u(D_2)E_u(\Delta Y(0)) \\
&= E_u(D_2)E_u(\Delta Y(0)) - E_u(D_2)E_u(\Delta Y(0)) \\
&= 0,
\end{aligned} \tag{7.26}$$

where the second equality follows from the law of iterated expectations. Next, it follows from the definition of $\text{TE}_{g,2}$ and the law of iterated expectations that

$$\begin{aligned}
E_u((D_2 - E_u(D_2))D_2 \text{TE}_2) &= E_u((D_2 - E_u(D_2))(Y_2(D_2) - Y_2(0))) \\
&= E_u((D_2 - E_u(D_2))E[Y_2(D_2) - Y_2(0)|D_2]) \\
&= \text{Cov}_u(D_2, E[Y_2(D_2) - Y_2(0)|D_2]).
\end{aligned}$$

The first result follows from plugging the preceding display and (7.26) into (7.25).

Using again the law of iterated expectations, we have

$$\begin{aligned}
E_u((D_2 - E_u(D_2))D_2 \text{TE}_2) &= E_u((D_2 - E_u(D_2))D_2 E_u(\text{TE}_2|D_2)) \\
&= E_u((D_2 - E_u(D_2))D_2 E_u(\text{TE}_2|D_2)|D_2 > 0) P_u(D_2 > 0).
\end{aligned}$$

Finally,

$$V_u(D_2) = E_u(D_2^2) - E_u(D_2)^2 = E_u[(D_2 - E_u(D_2))D_2] = E_u((D_2 - E_u(D_2))D_2|D_2 > 0) P_u(D_2 > 0).$$

The second result follows from plugging the two preceding displays and (7.26) into (7.25). **QED.**

7.5.2 Proof of Theorem 14

Point 1 of Theorem 14 directly follows from plugging (7.9) into (7.13) and from the fact that if $E(U|V) = a_0 + a_1 V$ then it is equal to the linear regression of U on $(1, V)$.

Then, assume that

$$E(\Delta Y|D_2) = \beta_0 + \beta^{\text{fe}} D_2. \tag{7.27}$$

(7.13) and (7.27) imply that

$$E(TE_2|D_2) = \beta^{\text{fe}} + (\beta_0 - \gamma_2)/D_2.$$

Then, if $\beta_0 \neq \gamma_2$, $\lim_{d_2 \rightarrow 0} E(\text{TE}_2|D_2 = d_2) = \infty$, thus contradicting the fact that $|E(TE_2|D_2)| \leq K$ under the bounded-slope assumption we maintain in this chapter. Therefore, $\beta_0 = \gamma_2$. Then, equating (7.13) and (7.27) implies that

$$D_2 E(\text{TE}_2|D_2) = \beta^{\text{fe}} D_2,$$

and dividing by $D_2 > 0$ yields $E(\text{TE}_2|D_2) = \beta^{\text{fe}}$. Taking expectations on both sides finally yields $\beta^{\text{fe}} = \text{ATT}$. This proves Point 2. **QED.**

7.5.3 Proof of Theorem 15

$E_u[\Delta Y|D_2 = 0]$ is well-defined in Design HAD'. Moreover, for all d_2 in the support of D_2 ,

$$\begin{aligned} & \frac{E_u[\Delta Y|D_2 = d_2] - E_u[\Delta Y|D_2 = 0]}{d_2} \\ &= \frac{E_u[Y_2(d_2) - Y_1(0)|D_2 = d_2] - E_u[\Delta Y(0)|D_2 = 0]}{d_2} \\ &= \frac{E_u[Y_2(d_2) - Y_2(0)|D_2 = d_2] + E_u[\Delta Y(0)|D_2 = d_2] - E_u[\Delta Y(0)|D_2 = 0]}{d_2} \\ &= E_u \left[\frac{Y_2(d_2) - Y_2(0)}{d_2} \middle| D_2 = d_2 \right]. \end{aligned}$$

Justify each step of this derivation.

The first equality follows from the fact we are in Design HAD'. The third equality follows from Assumption PTNB. This proves (7.14). (7.15) follows from (7.2), (7.14), and the law of iterated expectations. The previous display also shows that

$$E_u[\Delta Y|D_2 = d_2] - E_u[\Delta Y|D_2 = 0] = E_u [Y_2(d_2) - Y_2(0)|D_2 = d_2].$$

(7.16) follows from (7.4), the law of iterated expectations, the previous equation, and the law of iterated expectations **QED.**

Chapter 8

General designs

Motivation. Of the 26 highly-cited AER papers estimating a TWFE regression in the survey of de Chaisemartin and D'Haultfœuille (2025), two have an absorbing and binary treatment with no variation in treatment timing, four more have an absorbing and binary treatment with variation in treatment timing, and two more have an heterogeneous-adoption design. However, 18 have a design that differs from those we have studied so far. Below, we open this chapter by reviewing some of the designs in those 18 papers: non-absorbing binary treatments, absorbing treatments with variation in treatment timing and dose, and treatments that vary at baseline. Yet, we will not analyze in turn each non-binary and or non-staggered design ever encountered by a social scientist in their research. Doing so would lead to an encyclopedia, rather than an already long textbook. Instead, we believe that equipped with the fundamental insights from the two preceding chapters, we are now able to provide generic results, that apply to any design. The only restriction we impose throughout is that $D_{g,t} \geq 0$: the treatment should be a positive variable, as is most often the case, at least up to a normalization. Alongside generic results, we will also discuss what we see as particularly interesting design-specific results, but we will not exhaust all that can be said on say, non-absorbing binary treatments. We believe that more research on common non-binary and/or non-staggered designs would be very useful.

Non-absorbing binary treatments. First, social scientists are often interested in the effect of a non-absorbing binary treatment. For instance, Burgess, Jedwab, Miguel, Morjaria and Padró i Miquel (2015) study the effect, in Kenya, of sharing the ethnicity of a country's president,

on a district's volume of public expenditures. Districts can enter and leave the treatment (sharing the president's ethnicity) twice over the study period. An interesting special case is when groups can join and leave treatment once:

$$D_{g,t} = 1\{E_g \geq t \geq F_g\}. \quad (8.1)$$

When (8.1) holds, groups may get treated and leave treatment once, at heterogeneous dates F_g and E_g .

Absorbing treatments with variation in treatment timing and dose. Second, social scientists are often interested in the effect of an absorbing treatment with variation in treatment timing and dose:

$$D_{g,t} = I_g 1\{t \geq F_g\} \quad (8.2)$$

If (8.2) holds, treatment is absorbing but there is variation across groups in the period at which they start receiving the treatment, and in the dose they receive. For instance, Favara and Imbs (2015) study the effect, in the US, of financial deregulations conducted during the 1990s, on the volume of credit and housing prices. Their design almost satisfies (8.2): US states deregulate at heterogeneous times and with heterogeneous intensities. The only difference is that a small number of states deregulate more than once over the study period, so strictly speaking the treatment is not absorbing.

Treatments that vary at baseline. Third, social scientists are often interested in the effect of treatments whose intensity varies across groups at all time periods, including at period one:

$$\exists(g, g') : D_{g,1} \neq D_{g',1}. \quad (8.3)$$

For instance, Gentzkow et al. (2011) study the effect, in the US, of the number of newspapers in circulation in a county on turnout in presidential elections in that county. In 1868, the first presidential election used in their analysis, counties' number of newspapers ranges from 0 to 33. Another example is Fuest, Peichl and Siegloch (2018), who study the effect, in Germany, of the local business tax rate on wages. In 1993, the first period in their data, municipalities have business tax rates ranging from 10 to 37 percentage points.

Two common misconceptions about general designs. Before starting our study of general designs, it is important to clear up two common misconceptions. First, non-absorbing designs do not only arise when those that receive the treatment self-select in and out of it. In most of the aforementioned examples, the multiple treatment changes come from laws that are changed several times or repealed after having been enacted: twists and turns in policy making were not invented by the current US president. Therefore, while it is important to document the reasons that led the legislator to further or cancel an initial policy change, parallel-trends assumptions are not by construction less plausible in non-absorbing designs. Second, treatments continuously distributed across groups within periods are not necessarily continuously distributed over time within groups. Specifically, there exists designs where different groups all receive a different dose ($D_{g,t} \neq D_{g',t}$ for all t and $g \neq g'$), but where some groups have the same dose at different periods ($D_{g,t} = D_{g,t'}$ for some g and $t \neq t'$). For instance, in Fuest, Peichl and Siegloch (2018), in any given year the local business tax rate is close to being continuously distributed across municipalities, but many municipalities have the same tax rate in several years.

Chapter's running example: the effect of newspapers on voters' participation in elections. As in Chapter 5, our running example in this chapter is Gentzkow et al. (2011), who use a US panel data set at the county \times presidential-election level, with 1,195 counties and from the 1872 to the 1928 election, to measure the effect of newspapers on voters' participation in elections. To answer the green questions in this chapter, you need to use the `gentzkowetal_didtextbook` dataset. You can refer to the introduction of Chapter 5 for a description of this dataset.

Chapter's roadmap. First, we will see that in general designs, TWFE estimators may be even less robust to heterogeneous effects than in the previous chapters. Strikingly, even assuming that treatments are randomly assigned may not be enough to guarantee that they estimate a convex combination of effects. Second, we will see that heterogeneity-robust estimators can be extended to general designs, by combining insights from Chapters 6 and 7, and further ensuring that one compares switchers and stayers with the same baseline treatment.

8.1 Static Two-Way Fixed Effects estimator

No dynamic effects. In this section, we assume that the treatment has no dynamic effects, namely, we maintain Assumption ND. This is consistent with the TWFE regression in (3.1), where the current treatment $D_{g,t}$ is one of the independent variables, but the lagged treatments $D_{g,t-1}$, $D_{g,t-2}$ etc. are not part of the independent variables, thus implicitly ruling out dynamic treatment effects. We will relax Assumption ND in some of the chapter's next sections.

Two periods. Unless otherwise noted, in this section we also assume that $T = 2$. Doing so simplifies the exposition, without much substantive loss. When $T > 2$, for all $k \in \{1, \dots, T-1\}$ and $t \in \{k+1, \dots, T\}$ let $\hat{\beta}_{k,t}^{\text{fe}}$ denote the TWFE coefficient estimated restricting the sample to periods t and $t-k$. It follows from Theorem 1 in Ishimaru (2021) that $\hat{\beta}^{\text{fe}}$ is a weighted average of the coefficients $\hat{\beta}_{k,t}^{\text{fe}}$ across k and t . Therefore, even when $T > 2$, $\hat{\beta}^{\text{fe}}$ is a weighted average of two-periods TWFE coefficients. Without dynamic effects, each two-periods regression can be analyzed in isolation, because groups' outcomes at those two periods do not depend on their treatments at other periods. Then, once a “causal” decomposition of $\hat{\beta}_{k,t}^{\text{fe}}$ as a weighted sum of treatment effects has been obtained, one can plug it in the decomposition of $\hat{\beta}^{\text{fe}}$ as a weighted average of the $\hat{\beta}_{k,t}^{\text{fe}}$ s to finally obtain a “causal” decomposition of $\hat{\beta}^{\text{fe}}$.

(g,t) -specific treatment effects. To accommodate potentially non-binary treatments, we use the same definition of the (g,t) -specific treatment effects as in Chapter 5: for all (g,t) such that $D_{g,t} \neq 0$, we let

$$\text{TE}_{g,t} = \frac{E[Y_{g,t}(D_{g,t}) - Y_{g,t}(0)]}{D_{g,t}}.$$

Roadmap. In this section, we will show that $\hat{\beta}^{\text{fe}}$ is even less robust to heterogeneous treatment effects in general designs than in binary-and-staggered designs or in heterogeneous-adoption designs. Unsurprisingly, $\hat{\beta}^{\text{fe}}$ may not estimate a convex combination of effects under Assumption PT, or under a different parallel-trends assumption that may be better suited to general designs. More surprisingly, $\hat{\beta}^{\text{fe}}$ may still not estimate a convex combination of effects even if treatments are as-good-as randomly assigned.

8.1.1 Decomposition of TWFE regressions under Assumption PT

Recall that in the decomposition of $\hat{\beta}^{\text{fe}}$ in Theorem 8, the weights $W_{g,t}$ depend on the residuals $\hat{u}_{g,t}$ from a regression of $D_{g,t}$ on group and period FEs. Recall also that $\hat{u}_{g,t} = D_{g,t} - D_{g,.} - D_{.,t} + D_{.,.}$, where $D_{g,.}$ is the average treatment of group g across time periods, $D_{.,t}$ is the average treatment at period t across groups, and $D_{.,.}$ is the average treatment across groups and periods. When $T = 2$, the formula for $\hat{u}_{g,t}$ simplifies as follows:

$$\begin{aligned}\hat{u}_{g,1} &= D_{g,1} - (D_{g,1} + D_{g,2})/2 - D_{.,1} + (D_{.,1} + D_{.,2})/2 = 1/2(D_{g,1} - D_{g,2} - (D_{.,1} - D_{.,2})), \\ \hat{u}_{g,2} &= D_{g,2} - (D_{g,1} + D_{g,2})/2 - D_{.,2} + (D_{.,1} + D_{.,2})/2 = 1/2(D_{g,2} - D_{g,1} - (D_{.,2} - D_{.,1})).\end{aligned}\quad (8.4)$$

Therefore, $\hat{u}_{g,1} = -\hat{u}_{g,2}$. Recall that Δ denotes the first-difference operator, and let $\Delta D_+ = D_{.,2} - D_{.,1}$. Then, it directly follows from Theorem 8 that

$$E \left[\hat{\beta}^{\text{fe}} \right] = \sum_{g=1}^G \sum_{t=1}^2 \frac{(\Delta D_g - \Delta D_+) D_{g,t} (1\{t=2\} - 1\{t=1\})}{\sum_{g'=1}^G \sum_{t'=1}^2 (\Delta D_{g'} - \Delta D_+) D_{g',t'} (1\{t'=2\} - 1\{t'=1\})} \text{TE}_{g,t}. \quad (8.5)$$

Assume that $D_{g,t} > 0$ and $\Delta D_g \neq \Delta D_+$ for all (g, t) . Then, which proportion of treatment effects $\text{TE}_{g,t}$ are weighted negatively?

If $D_{g,t} > 0$ for all (g, t) and $\Delta D_g \neq \Delta D_+$, all effects receive a non-zero weight. As the weight on $\text{TE}_{g,1}$ is equal to the weight on $\text{TE}_{g,2}$ multiplied by -1 , exactly a half of the weights are negative: for every g such that $\Delta D_g \neq \Delta D_+$, either $\text{TE}_{g,2}$ or $\text{TE}_{g,1}$ is weighted negatively, a fact first noted by de Chaisemartin and Lei (2021). This feature is specific to two-periods TWFE regressions: if $T > 2$, the proportion of effects weighted negatively by $\hat{\beta}^{\text{fe}}$ might differ from a half.

Intuitively, why is it that $\hat{\beta}^{\text{fe}}$ may weight negatively exactly a half of treatment effects?

The TWFE estimator compares the outcome evolution of groups whose treatment increases more to the outcome evolution of groups whose treatment increases less.

When $T = 2$, $\hat{\beta}^{\text{fe}}$ is numerically equivalent to $\hat{\beta}^{\text{fd}}$, the coefficient from the first-difference regression of ΔY_g on an intercept and ΔD_g . Thus, it follows from standard formulas for coefficients in regressions with one non-constant explanatory variable that

$$\hat{\beta}^{\text{fe}} = \frac{\sum_{g=1}^G (\Delta D_g - \Delta D.) \Delta Y_g}{\sum_{g=1}^G (\Delta D_g - \Delta D.)^2}. \quad (8.6)$$

When we regress a dependent variable on an intercept and a binary treatment variable, the coefficient on the treatment compares the average of the dependent variable in the treatment and control groups. Here, ΔD_g is not binary, but (8.6) still shows that the coefficient on ΔD_g has a similar “treated-versus-control” interpretation. $\hat{\beta}^{\text{fe}}$ gives a positive weight to the outcome evolution ΔY_g of groups whose treatment change from period one to two is larger than the average ($\Delta D_g > \Delta D.$): those groups are used as “treatment groups” by $\hat{\beta}^{\text{fe}}$. But then, as

$$\Delta Y_g = Y_{g,2}(0) - Y_{g,1}(0) + Y_{g,2} - Y_{g,2}(0) - (Y_{g,1} - Y_{g,1}(0)) = Y_{g,2}(0) - Y_{g,1}(0) + D_{g,2}\text{TE}_{g,2} - D_{g,1}\text{TE}_{g,1},$$

if those groups are treated at period one, their $\text{TE}_{g,1}$ is weighted negatively by $\hat{\beta}^{\text{fe}}$. Similarly, $\hat{\beta}^{\text{fe}}$ gives a negative weight to the outcome evolution ΔY_g of groups whose treatment change from period one to two is lower than the average ($\Delta D_g < \Delta D.$): those groups are used as “control groups”. But then, if those groups are treated at period two, their $\text{TE}_{g,2}$ is weighted negatively by $\hat{\beta}^{\text{fe}}$.

With two groups, $\hat{\beta}^{\text{fe}}$ reduces to a Wald-DID estimator. With two groups m and ℓ , (8.6) reduces to

$$\hat{\beta}^{\text{fe}} = \frac{Y_{m,2} - Y_{m,1} - (Y_{\ell,2} - Y_{\ell,1})}{D_{m,2} - D_{m,1} - (D_{\ell,2} - D_{\ell,1})}. \quad (8.7)$$

The right hand side of (8.7) is a so-called Wald-DID estimator which compares the outcome evolution of a group whose treatment increases more to the outcome evolution of a group whose treatment increases less ($D_{m,2} - D_{m,1} > D_{\ell,2} - D_{\ell,1}$). A Wald-DID can estimate a non-convex combinations of treatment effects, if treatment effects vary between groups (Blundell and Costa-Dias, 2009) or if they change over time (de Chaisemartin, 2011; de Chaisemartin and D’Haultfœuille, 2018). To see that, consider a simple example where $D_{m,1} = 2$, $D_{m,2} = 4$, $D_{\ell,1} = 1$, and

$D_{\ell,2} = 2$: group m receives two units of treatment at period one and four units at period two, and group ℓ receives one unit of treatment at period one, and two at period two. Then, $D_{m,2} - D_{m,1} - (D_{\ell,2} - D_{\ell,1}) = 1$, and

$$\begin{aligned} E[\hat{\beta}^{\text{fe}}] &= E[Y_{m,2}(4) - Y_{m,1}(2) - (Y_{\ell,2}(2) - Y_{\ell,1}(1))] \\ &= E[Y_{m,2}(0) - Y_{m,1}(0) - (Y_{\ell,2}(0) - Y_{\ell,1}(0))] + 4\text{TE}_{m,2} - 2\text{TE}_{m,1} - 2\text{TE}_{\ell,2} + \text{TE}_{\ell,1} \\ &= 4\text{TE}_{m,2} - 2\text{TE}_{m,1} - 2\text{TE}_{\ell,2} + \text{TE}_{\ell,1}, \end{aligned} \quad (8.8)$$

where the last equality follows from Assumption PT. The right hand side of the previous display is a weighted sum of m and ℓ 's treatment effects at periods 1 and 2, with weights summing to one, and where two effects enter with negative weights. Assuming that treatment effects are constant across groups but not over time ($\text{TE}_{m,t} = \text{TE}_{\ell,t} := \text{TE}_t$), the previous display reduces to

$$E[\hat{\beta}^{\text{fe}}] = 2\text{TE}_2 - \text{TE}_1,$$

so $\hat{\beta}^{\text{fe}}$ still estimates a non-convex combination of effects. Assuming that treatment effects are constant over time but not across groups ($\text{TE}_{g,2} = \text{TE}_{g,1} := \text{TE}_g$), the previous display reduces to

$$E[\hat{\beta}^{\text{fe}}] = 2\text{TE}_m - \text{TE}_\ell,$$

so $\hat{\beta}^{\text{fe}}$ still estimates a non-convex combination of effects.

8.1.2 Decomposition of TWFE regressions under an alternative parallel-trends assumption

Parallel trends if groups' treatment does not change. Instead of Assumption PT, assume that

$$E[Y_{g,2}(D_{g,1}) - Y_{g,1}(D_{g,1})] \text{ does not vary across } g. \quad (8.9)$$

Interpret (8.9).

(8.9) is a parallel-trends assumption in the counterfactual where groups' treatment does not change between periods one and two, while Assumption PT is a parallel-trends assumption in the counterfactual where they remain untreated. In general designs, it might be the case that very few groups are untreated at period one. Then, it might be more natural to impose a parallel-trends assumption on $Y_{g,t}(D_{g,1})$ than on $Y_{g,t}(0)$. For instance, if a third period of data, period zero, is available, and all groups keep the same treatment from period zero to one, then $Y_{g,1}(D_{g,1}) - Y_{g,0}(D_{g,1})$ is observed for all groups. Then, one can conduct a pre-trends test of (8.9), while one cannot conduct a pre-trends test of Assumption PT as $Y_{g,1}(0) - Y_{g,0}(0)$ is not observed for all groups.

$\hat{\beta}^{\text{fe}}$ may not estimate a convex combination of effects under (8.9). If $\Delta D_g \neq 0$, let

$$\text{TE}_{g,2}^\Delta = \frac{E[Y_{g,2}(D_{g,2}) - Y_{g,2}(D_{g,1})]}{D_{g,2} - D_{g,1}}$$

denote the expectation of the slope of group g 's potential outcome function at period two, between its period-one and its period-two treatment. Let

$$W_{g,t}^\Delta = \frac{(\Delta D_g - \Delta D.)\Delta D_g}{\sum_{g'=1}^G (\Delta D_{g'} - \Delta D.)^2}.$$

Theorem 18 *If $T = 2$ and Assumptions NA, ND, and (8.9) hold,*

$$E[\hat{\beta}^{\text{fe}}] = \sum_{g:\Delta D_g \neq 0}^G W_{g,t}^\Delta \text{TE}_{g,2}^\Delta. \quad (8.10)$$

To fix ideas, assume that $\Delta D. > 0$: the average treatment increases from period one to two. Then, if there are groups whose treatment increases from period one to two ($\Delta D_g > 0$), but increases less than the average increase in the population ($\Delta D_g - \Delta D. < 0$), their slope $\text{TE}_{g,2}^\Delta$ is weighted negatively by $\hat{\beta}^{\text{fe}}$: $\hat{\beta}^{\text{fe}}$ may still estimate a non-convex combination of effects. Like all the other theorems in this chapter, Theorem 18 is proven in the appendix of this chapter.

If there is at least one untreated group at period one, Assumption PT and (8.9) imply that some treatment effects are constant over time. Assume that Assumption PT and (8.9) both hold. Then, there exists real numbers γ_2 and γ_2^Δ such that $E[Y_{g,2}(0) - Y_{g,1}(0)] = \gamma_2$

and $E[Y_{g,2}(D_{g,1}) - Y_{g,1}(D_{g,1})] = \gamma_2^\Delta$ for all g . Now, assume that there exists at least one group g_0 that is untreated at period one: $D_{g_0,1} = 0$. Then,

$$\gamma_2^\Delta = E[Y_{g_0,2}(D_{g_0,1}) - Y_{g_0,1}(D_{g_0,1})] = E[Y_{g_0,2}(0) - Y_{g_0,1}(0)] = \gamma_2.$$

Therefore, for every g ,

$$\begin{aligned} E[Y_{g,2}(D_{g,1}) - Y_{g,1}(D_{g,1})] &= E[Y_{g,2}(0) - Y_{g,1}(0)] \\ \Leftrightarrow E[Y_{g,2}(D_{g,1}) - Y_{g,2}(0)] &= E[Y_{g,1}(D_{g,1}) - Y_{g,1}(0)] : \end{aligned} \quad (8.11)$$

the effect of switching the treatment from 0 to $D_{g,1}$ has to be the same at periods one and two. (8.9) only imposes restrictions on one potential outcome per group, $Y_{g,t}(D_{g,1})$, so that condition alone does not impose restrictions on treatment effects. However, when combined with Assumption PT, it implies that some treatment effects have to be constant over time. Then, to have that (8.9) does not restrict effects' heterogeneity over time, Assumption PT has to fail: groups have to be on parallel trends in the counterfactual where they keep the same treatment in periods one and two, but they have to experience differential trends in the counterfactual where they remain untreated at both dates. Such a scenario might be hard to rationalize, so we view (8.9) as "essentially" assuming constant effects over time.

With a binary treatment, $\hat{\beta}^{\text{fe}}$ estimates a convex combination of effects under parallel-trends assumptions on the untreated and treated outcomes. Note that if $D_{g,t}$ is binary, the weights $W_{g,t}^\Delta$ are necessarily positive: for all g such that $\Delta D_g \neq 0$, either $\Delta D_g = -1$, in which case $\Delta D_g - \Delta D. < 0$ and $(\Delta D_g - \Delta D.)\Delta D_g > 0$, or $\Delta D_g = 1$, in which case $\Delta D_g - \Delta D. > 0$ and $(\Delta D_g - \Delta D.)\Delta D_g > 0$. Then, if $D_{g,t}$ is binary $\hat{\beta}^{\text{fe}}$ estimates a convex combination of effects under (8.9), but we have seen that (8.9) essentially assumes constant effects over time. Actually, Fabre (2023) shows that if $D_{g,t}$ is binary, $\hat{\beta}^{\text{fe}}$ still estimates a convex combination of effects under the following condition:

$$\begin{aligned} E[Y_{g,2}(0) - Y_{g,1}(0)] &\text{ does not vary across } g \\ E[Y_{g,2}(1) - Y_{g,1}(1)] &\text{ does not vary across } g. \end{aligned} \quad (8.12)$$

(8.12) requires that all groups have the same expected evolutions of their untreated and treated outcomes. This condition implies that all groups should experience the same evolution of their

treatment effect from period one to two, but unlike (8.9) it does not imply that the treatment effect is constant over time. On the other hand, when $D_{g,t}$ is not binary, assuming parallel trends for all potential outcomes rather than just for the untreated outcome is not enough to ensure that $\hat{\beta}^{\text{fe}}$ estimates a convex combination of effects.

Bibliographic notes. When $T = 2$, the weights in the decomposition on $\hat{\beta}^{\text{fe}}$ in Theorem 18 coincide with the weights in the decomposition of $\hat{\beta}^{\text{fe}}$ in Theorem S2 of de Chaisemartin and D'Haultfoeuille (2020). Instead of assuming (8.9), the decomposition therein assumes Assumption PT and constant treatment effects over time, but we have seen that the two sets of conditions are closely connected.

8.1.3 Decomposition of TWFE regressions with randomly-assigned treatments*

In binary-and-staggered designs and in heterogeneous-adoption designs, we saw that while the TWFE regression can estimate a non-convex combination of effects under parallel trends, it does estimate a convex combination of effects under the assumption that the treatment timing or the treatment dose is as-good-as randomly assigned. Strikingly, this is no longer true in general designs. Then, de Chaisemartin and Lei (2024) show that the TWFE regression can estimate a non-convex combination of effects even if treatments are as-good-as random.

I.i.d. groups. In this section, to simplify the exposition we momentarily take the sampling-based perspective and assume that the G groups we observe are a random sample drawn from an infinite super population. Thus, we replace Assumption IND by Assumption IID and we drop the g subscript. Instead of $E[\hat{\beta}^{\text{fe}}]$, the estimand we consider is

$$\beta^{\text{fe}} := \frac{\text{cov}_u(D_2 - D_1, Y_2 - Y_1)}{V_u(D_2 - D_1)},$$

the probability limit of $\hat{\beta}^{\text{fe}}$.

Linear-treatment-effect assumption. We also assume that treatment has a linear effect on the outcome:

$$Y_t(d) = Y_t(0) + S_t d. \quad (8.13)$$

The slope S_t is the treatment's effect. It is a random variable, that may vary across groups. It is also indexed by t , as the treatment's effect might be time-varying. Under (8.13), one has

$$\begin{aligned}\Delta Y &= Y_2(0) + S_2 D_2 - Y_1(0) - S_1 D_1 \\ &= Y_2(0) - Y_1(0) + S_2 \times \Delta D + \Delta S \times D_1.\end{aligned}\tag{8.14}$$

As good as randomly assigned first-differenced treatment. Assume that

$$\Delta D \perp\!\!\!\perp (Y_2(0) - Y_1(0), S_1, S_2).\tag{8.15}$$

(8.15) requires that $\Delta D \perp\!\!\!\perp Y_2(0) - Y_1(0)$, a condition similar in spirit to the parallel-trends condition in Assumption PT. But (8.15) also requires that $\Delta D \perp\!\!\!\perp (S_1, S_2)$, groups' treatment change should be independent of the level of their treatment effect at periods one and two. Requiring that ΔD is independent of the level of those unobservables essentially amounts to assuming that ΔD is as-good-as randomly assigned.

Omitted variable bias (OVB) formula. Assume that (8.15) holds. Then,

$$\begin{aligned}\beta^{\text{fe}} &= \frac{\text{cov}_u(\Delta D, \Delta Y)}{V_u(\Delta D)} \\ &= \frac{\text{cov}_u(\Delta D, Y_2(0) - Y_1(0)) + \text{cov}_u(\Delta D, S_2 \times \Delta D) + \text{cov}_u(\Delta D, \Delta S \times D_1)}{V_u(\Delta D)} \\ &= \frac{\text{cov}_u(\Delta D, Y_2(0) - Y_1(0)) + E_u((\Delta D)^2 S_2) - E_u(\Delta D) E_u(\Delta D S_2) + \text{cov}_u(\Delta D, \Delta S \times D_1)}{V_u(\Delta D)} \\ &= E_u(S_2) + \frac{\text{cov}_u(\Delta D, \Delta S \times D_1)}{V_u(\Delta D)},\end{aligned}\tag{8.16}$$

where the second equality follows from (8.14) and the fourth follows from (8.15). (8.16) is akin to a standard omitted variable bias (OVB) formula. It shows that β^{fe} identifies the average treatment effect at period two $E_u(S_2)$, plus the OVB term

$$\frac{\text{cov}_u(\Delta D, \Delta S \times D_1)}{V_u(\Delta D)}.$$

Intuition for the OVB formula. Letting $\alpha = E_u(Y_2(0) - Y_1(0))$, and letting

$$\varepsilon = Y_2(0) - Y_1(0) - E(Y_2(0) - Y_1(0)) + (S_2 - E_u(S_2)) \times \Delta D + \Delta S \times D_1,$$

(8.14) is equivalent to

$$\Delta Y = \alpha + E_u(S_2) \times \Delta D + \varepsilon. \quad (8.17)$$

Then, the residual of a regression of ΔY on ΔD is ε , and ΔD has to be uncorrelated to ε for the regression not to suffer from an OVB. (8.15) ensures that ΔD is uncorrelated to $Y_2(0) - Y_1(0) - E(Y_2(0) - Y_1(0))$ and to $(S_2 - E_u(S_2)) \times \Delta D$. But if the treatment effect changes over time ($S_2 \neq S_1$), ε is also a function of D_1 , so having that ΔD and ε are uncorrelated essentially requires that

$$\Delta D \perp\!\!\!\perp D_1 : \quad (8.18)$$

ΔD and D_1 should be independent. If ΔD and D_1 are correlated, β^{fe} attributes to ΔD changes in Y_t that would have happened without any treatment change, that are caused by changes in the treatment's effect, the $\Delta S \times D_1$ term in ε .

Interpreting and testing (8.18). (8.18) is a strong condition. For instance, it cannot hold if D_1 and D_2 take the same bounded set of values, and the distribution of $\Delta D|D_1 = d$ is non-degenerate for all d . If $(D_t)_{t \geq 1}$ follows an AR(1) process (namely, $D_{t+1} = \lambda_0 + \lambda_1 D_t + \epsilon_{t+1}$ for all $t \geq 1$, with $(\epsilon_t)_{t \geq 1}$ independent across t), (8.18) can only hold in the knife-edge case where $\lambda_1 = 1$, meaning that the process is non-stationary. (8.18) is also a testable condition. [How can one test \(8.18\)?](#)

For instance, one can regress ΔD on D_1 , or one can use a Kolmogorov-Smirnov test similar to that discussed in Section 3.2.1.1. Then, researchers assuming that ΔD is as good as random to justify a TWFE regression should start by testing if ΔD and D_1 are independent. If they are not, and if the treatment effect varies over time ($S_1 \neq S_2$), the regression may be subject to an OVB, even if ΔD is independent of the unobservables $(Y_2(0) - Y_1(0), S_1, S_2)$.

Even if the treatment paths (D_1, D_2) are randomly assigned, β^{fe} may identify a non-convex combination of $E_u(S_1)$ and $E_u(S_2)$. As the OVB in the previous display is a

function of ΔS , one may wonder if under reasonable assumptions this OVB can simplify so that β^{fe} identifies a relatively interpretable causal effect, like a convex combination of the effects at period one and two. The following result shows this is not necessarily the case, even under the following, stronger assumption:

$$(D_1, D_2) \perp\!\!\!\perp (Y_1(0), Y_2(0), S_1, S_2). \quad (8.19)$$

(8.19) requires that groups' treatment paths (D_1, D_2) be as-good-as random, which is stronger than assuming that their treatment changes ΔD are as good as random.

Theorem 19 *If Assumptions NA, ND, (8.13) and (8.19) hold,*

$$\beta^{fe} = \sum_{t=1}^2 \frac{V_u(D_t) - \text{cov}_u(D_1, D_2)}{\sum_{t'=1}^2 [V_u(D_{t'}) - \text{cov}_u(D_1, D_2)]} E_u(S_t).$$

Find conditions on the distribution of (D_1, D_2) under which the weights in Theorem 19 are all positive.

Cases where the weights in Theorem 19 are guaranteed to be positive. Theorem 19 shows that under (8.13) and (8.19), β^{fe} estimates a weighted sum of the average slopes $E_u(S_1)$ and $E_u(S_2)$. Weights are guaranteed to be positive if the treatment is binary.¹ This is consistent with the results of Athey and Imbens (2022) and Arkhangelsky, Imbens, Lei and Luo (2024), who have shown that with a binary randomized treatment, OLS TWFE regressions always estimate a convex combination of effects. By Cauchy-Schwarz's inequality, weights are also guaranteed to be positive if $V_u(D_1) = V_u(D_2)$. Weights are also guaranteed to be positive if the treatments are not serially correlated ($\text{cov}_u(D_1, D_2) = 0$) or negatively correlated ($\text{cov}_u(D_1, D_2) < 0$). However, those positive results only hold under (8.19), which requires that the treatment paths, rather than their first differences, be randomly assigned. Accordingly, invoking Theorem 19 to justify a TWFE regression requires advocating for random assignment of the paths.

¹To see this, note that $V_u(D_1) - \text{cov}_u(D_1, D_2) = E_u((D_1 - E_u(D_1))(D_2 - E_u(D_1)))$, and $(D_1 - E_u(D_1))(D_2 - E_u(D_1)) \geq 0$ almost surely for binary variables.

Cases where one of the two weights in Theorem 19 could be negative. One of the two weights in Theorem 19 could be negative if D_1 and D_2 are non-binary, positively correlated, and have different variances. The weights in Theorem 19 can be estimated, to assess if in a given application, $\hat{\beta}^{\text{fe}}$ estimates a convex combination of effects under (8.13) and (8.19).

Adding group FEs can do more harm than good. Under (8.19), estimating a TWFE regression is not necessary anymore: one can merely regress Y_1 on an intercept and D_1 to estimate $E_u(S_1)$, and one can regress Y_2 on an intercept and D_2 to estimate $E_u(S_2)$. Thus, adding group FEs to the regression is not necessary, and adding those FEs can actually do more harm than good, by leading the researcher to estimate a non-convex combination of treatment effects.

Application to the effect of Chinese imports on US employment. de Chaisemartin and Lei (2024) apply the results above to a two-periods TWFE regression of US industries' employment evolutions on their Chinese imports penetration, using years 1999 and 2007 of the panel data set constructed by Acemoglu, Autor, Dorn, Hanson and Price (2016). They find that $\hat{\beta}^{\text{fe}} = -0.78$ (s.e.=0.22): according to this regression, an increase of Chinese imports decreases US employment. To understand the assumptions underlying this finding, de Chaisemartin and Lei (2024) start by testing (8.18), by regressing ΔD , the change in industries' import-penetration ratio from 1999 to 2007, on D_1 , industries' import-penetration ratio in 1999. The coefficient on D_1 is equal to 0.74, and it is highly significant (s.e.=0.16). Thus, the industries that experienced the largest growth of their imports-penetration ratio from 1999 to 2007 already had a larger imports-penetration ratio in 1999. Then, $\hat{\beta}^{\text{fe}}$ may be subject to an OVB even if ΔD is as-good-as randomly assigned. Then, they estimate the weights attached to $\hat{\beta}^{\text{fe}}$ in Theorem 19. They find that $\hat{\beta}^{\text{fe}}$ estimates a weighted sum of the average effects of import penetration in 1999 and 2007, where the 2007 average effect receives a weight equal to 1.3, while the 1999 average effect receives a weight equal to -0.3. Thus, even if industries' import paths are as-good-as random, $\hat{\beta}^{\text{fe}}$ does not estimate a convex combination of time-varying treatment effects. Then, a causal interpretation of $\hat{\beta}^{\text{fe}}$ requires assuming that the effect of Chinese imports is constant over time.

Application to the newspapers example. Gentzkow et al. (2011) estimate a so-called stacked “first-difference” regression with more than two time periods, but results similar to

those above apply to such regressions. Rather than invoking a parallel-trends assumption to justify this regression, one could instead assume that (8.15) holds, meaning that the change in counties' number of newspapers is as good as random. Then, we have seen that the first-difference regression may still be subject to an OVB if that change is correlated to counties' lagged number of newspapers. Using `gentzkowetal_didtextbook`, regress `changedailies` on `lag_numdailies`, clustering standard errors at the county level. Interpret the results.

The coefficient on `lag_numdailies` is negative and significant. Then, the first-difference regression may still be subject to an OVB, even if the change in counties' number of newspapers is as good as random. To assess the plausibility of (8.15), one may regress the change in counties' number of newspapers on some counties' characteristics that are unlikely to be affected by that change, in the spirit of a balancing check in a randomized controlled trial. Using `gentzkowetal_didtextbook`, regress `changedailies` on `lag_ishare_urb`, counties' lagged urbanization rate, clustering standard errors at the county level. Interpret the results.

The coefficient on `lag_ishare_urb` is negative and significant: more-urbanized counties are less likely to experience an increase in their number of newspapers. This suggests that the change in counties' number of newspapers may not be as good as random.

8.1.4 Extensions

8.1.4.1 TWFE regressions with several treatments

A decomposition of TWFE regressions with several treatments. In this section, we no longer assume that $T = 2$ and we assume that we are interested in the effect of K treat-

ments. For instance, one may be interested in estimating separately the effects of laws legalizing Marijuana consumption for medical and for recreational purposes. For every $(k, g, t) \in \{1, \dots, K\} \times \{1, \dots, G\} \times \{1, \dots, T\}$, let $D_{g,t}^k$ denote the value of treatment k for group g at period t . To simplify the exposition, we assume that the treatments are binary. Let $\hat{\beta}^{\text{fe}}$ denote the coefficient on $D_{g,t}^1$ in a regression of $Y_{g,t}$ on group fixed effects, period fixed effects, and the vector $(D_{g,t}^1, \dots, D_{g,t}^K)$. Let $\text{TE}_{g,t}^1$ denote the effect, in cell (g, t) , of moving the first treatment from zero to one while keeping the other treatments at their observed values. Let $\text{TE}_{g,t}^{-1}$ denote the effect, in cell (g, t) , of moving the other treatments from zero to their actual values, while keeping the first treatment at zero. Under a no-anticipation assumption and a parallel-trends assumption in the counterfactual where groups do not receive any of the K treatments, de Chaisemartin and D'Haultfœuille (2023a) show that

$$E[\hat{\beta}^{\text{fe}}] = \sum_{(g,t):D_{g,t}^1=1} W_{g,t}^1 \text{TE}_{g,t}^1 + \sum_{(g,t):(D_{g,t}^2, \dots, D_{g,t}^K) \neq \mathbf{0}_{K-1}} W_{g,t}^{-1} \text{TE}_{g,t}^{-1},$$

where $\sum_{(g,t):D_{g,t}^1=1} (N_{g,t}/N_1) W_{g,t}^1 = 1$. Thus, $\hat{\beta}^{\text{fe}}$ does not estimate the ATT of $D_{g,t}^1$, and it may not even estimate a convex combination of the effects of that treatment, as some of the weights $W_{g,t}^1$ may be negative. Moreover, $\hat{\beta}^{\text{fe}}$ is contaminated by the effects of the other treatments, a phenomenon very similar to that we discussed in the context of TWFE ES regressions in Chapter 6. A difference with TWFE ES regressions is that the contamination weights $W_{g,t}^{-1}$ do not always sum to zero. Accordingly, even if the effects of all treatments are constant, $\hat{\beta}^{\text{fe}}$ may still be biased for the first treatment's effect. There are still three special cases where the contamination weights sum to zero: if $K = 2$, or if the treatments $D_{g,t}^2, \dots, D_{g,t}^K$ are mutually exclusive, or if for all (g, t) , there exists $(\delta_{g,t}^k)_{k=2,\dots,K}$ such that

$$\text{TE}_{g,t}^{-1} = \sum_{k=2}^K D_{g,t}^k \delta_{g,t}^k, \quad (8.20)$$

meaning that there is no interaction effect between the treatments $D_{g,t}^2, \dots, D_{g,t}^K$.

The origin of the contamination weights. Consider a simple example with four groups and two periods. No group is treated at period 1. At period two, group two receives the first treatment, group three receives the second treatment, and group four receives both treatments. Then one can show that

$$\hat{\beta}^{\text{fe}} = \frac{1}{2} (Y_{2,2} - Y_{2,1} - (Y_{1,2} - Y_{1,1})) + \frac{1}{2} (Y_{4,2} - Y_{4,1} - (Y_{3,2} - Y_{3,1})).$$

The second DID in the previous display compares the period-one-to-two outcome evolution of group 4, that starts receiving the first and second treatments at period 2, to that of group 3, that only starts receiving the second treatment. If the effect of the second treatment is constant across groups, that DID unbiasedly estimates the effect of the first treatment. But if the effect of the second treatment is heterogeneous, that DID is contaminated by the effect of the second treatment.

Computation: Stata and R commands to compute the weights attached to TWFE regressions with several treatments. The `twowayfeweights` Stata and R commands can be used to compute the weights attached to TWFE regressions with several treatments, using the `other_treatments` option.

8.1.4.2 Two-stage-least-squares TWFE regressions, and Bartik regressions*

To simplify the discussion, in this section we assume that $T = 2$ and groups are an i.i.d. sample drawn from a larger population, thus allowing us to drop the g subscript except when we consider estimators.

Motivation. Sometimes, it may be more plausible to assume parallel-trends with respect to an instrument rather than the treatment. For instance, one may be interested in estimating the price-elasticity of a good. If prices respond to demand shocks, groups' counterfactual consumption trends may be correlated with their change in price, thus violating Assumption PT. Instead, one may impose a parallel-trends assumption with respect to taxes, and use taxes as an instrument for prices. Let $\hat{\beta}_{2SLS}^{fe}$ denote the coefficient on $D_{g,t}$ in a so-called 2SLS-TWFE regression of $Y_{g,t}$ on group and period fixed effects and $D_{g,t}$, using a variable $Z_{g,t}$ (e.g.: taxes) as the instrument for $D_{g,t}$ (e.g.: prices).

2SLS-TWFE regressions, outside of the classical IV-DID design. In Section 3.8, we showed that in a classical IV-DID design, with two periods and a binary instrument such that a subset of units receive the instrument at period two, a 2SLS-TWFE regression estimates a LATE under reduced-form and first-stage parallel-trends assumptions. On the other hand, outside of a classical IV-DID design, those two assumptions are not sufficient to ensure that 2SLS-TWFE

regressions estimate a LATE or even just a convex combination of (g, t) -specific effects. As $T = 2$ and groups are i.i.d., the probability limit of $\hat{\beta}_{2SLS}^{\text{fe}}$ is

$$\beta_{2SLS}^{\text{fe}} := \frac{\text{cov}(\Delta Y, \Delta Z)}{\text{cov}(\Delta D, \Delta Z)}.$$

Consider the following first-stage and reduced-form parallel-trends assumptions:

$$\text{cov}(D_2(Z_1) - D_1(Z_1), \Delta Z) = 0, \quad \text{cov}(Y_2(D_2(Z_1)) - Y_1(D_1(Z_1)), \Delta Z) = 0,$$

respectively requiring that the instrument change ΔZ be uncorrelated with the counterfactual trends of the treatment and outcome if the instrument had not changed. Under those assumptions, as

$$\begin{aligned}\Delta D &= D_2(Z_2) - D_1(Z_1) = D_2(Z_2) - D_2(Z_1) + D_2(Z_1) - D_1(Z_1) \\ \Delta Y &= Y_2(D_2(Z_2)) - Y_1(D_1(Z_1)) = Y_2(D_2(Z_2)) - Y_2(D_2(Z_1)) + Y_2(D_2(Z_1)) - Y_1(D_1(Z_1)),\end{aligned}$$

one has that

$$\begin{aligned}\beta_{2SLS}^{\text{fe}} &= \frac{\text{cov}(Y_2(D_2(Z_2)) - Y_2(D_2(Z_1)), \Delta Z)}{\text{cov}(D_2(Z_2) - D_2(Z_1), \Delta Z)} \\ &= E \left(W^{2SLS} \frac{Y_2(D_2(Z_2)) - Y_2(D_2(Z_1))}{D_2(Z_2) - D_2(Z_1)} \right),\end{aligned}$$

where

$$W^{2SLS} = \frac{(\Delta Z - E(\Delta Z))(D_2(Z_2) - D_2(Z_1))}{E[(\Delta Z - E(\Delta Z))(D_2(Z_2) - D_2(Z_1))]},$$

and we use the convention that $0/0 = 0$. The previous display shows that β_{2SLS}^{fe} may estimate a non-convex combination of the slopes $(Y_2(D_2(Z_2)) - Y_2(D_2(Z_1)))/(D_2(Z_2) - D_2(Z_1))$, even if one further makes a monotonicity assumption such as $Z_2 \geq Z_1 \Rightarrow D_2(Z_2) \geq D_2(Z_1)$.

Bartik regressions. A common special case of 2SLS-TWFE regressions are Bartik-TWFE regressions, where ΔZ_g has a so-called shift-share structure: $\Delta Z_g = \sum_{s=1}^S Q_{s,g} \Delta Z_s$. Typically, ΔZ_s is a sector-specific shock, and $Q_{s,g}$ is the share that sector s accounts for in, say, the employment of location g . For instance, in Autor et al. (2013), ΔZ_s is the change in imports from China to high-income countries in sector s , and $Q_{s,g}$ is the share that sector s accounts for in the employment of US commuting zone (CZ) g . Under a parallel-trends assumption, requiring that

the shares $Q_{s,g}$ are uncorrelated to groups' outcome evolutions without treatment, Goldsmith-Pinkham et al. (2020) show that if the treatment effect is constant then $\hat{\beta}_{2SLS}^{fe}$ is consistent for the constant-effect parameter. As Bartik-TWFE regressions are just a special case of 2SLS-TWFE regressions, the result in the previous paragraph implies that this conclusion no longer holds if treatment effects are heterogeneous: then, Bartik regressions may estimate a non-convex combination of effects. Accordingly, de Chaisemartin and Lei (2021) find that under parallel trends, $\hat{\beta}_{2SLS}^{fe}$ estimates, in Autor et al. (2013), a highly non-convex combination of CZ-and-year-specific effects of Chinese imports on US employment. Moreover, the weights are correlated with some CZ characteristics like the offshorability of their employment, which are themselves likely to be correlated with their employment elasticity to Chinese imports. Instead of assuming parallel trends, Borusyak et al. (2022) assume that the shocks ΔZ_s are as good as random. In line with the results we discussed in Section 8.1.3, de Chaisemartin and Lei (2021) show that even if the shocks are as good as random, $\hat{\beta}_{2SLS}^{fe}$ may still fail to estimate a convex combination of treatment effects if those effects change over time, and if ΔZ_s and $Z_{s,1}$ are correlated. When they revisit Autor et al. (2013), de Chaisemartin and Lei (2021) find that ΔZ_s and $Z_{s,1}$ are very strongly positively correlated, and that some-industry level characteristics predict ΔZ_s , thus suggesting that ΔZ_s may not be as good as random.

8.2 Distributed-Lag Two-Way Fixed Effects estimators

In this section, we no longer impose Assumption ND and we allow for dynamic effects. We also no longer assume that $T = 2$.

Distributed-lag regressions. In general designs, a commonly-used estimator of dynamic treatment effects, for instance discussed in Equation (5.2.6) of Angrist and Pischke (2009), is the distributed-lag TWFE estimator. For $\ell \in \{0, \dots, K\}$, let $\hat{\beta}_\ell^{dl}$ denote the coefficient on $D_{g,t-\ell}$ in a regression of $Y_{g,t}$ on group and period FEs and $(D_{g,t-\ell})_{\ell \in \{0, \dots, K\}}$, in the subsample such that $t \geq K + 1$:

$$Y_{g,t} = \sum_{g'=1}^G \hat{\alpha}_{g'} 1\{g = g'\} + \sum_{t'=1}^T \hat{\gamma}_{t'} 1\{t = t'\} + \sum_{\ell=0}^K \hat{\beta}_\ell^{dl} D_{g,t-\ell} + \hat{\epsilon}_{g,t}. \quad (8.21)$$

In practice, researchers may slightly augment or modify (8.21). They may include treatment leads in the regression, to test Assumptions NA and PT. They may define the lagged treatments as equal to 0 at time periods when they are not observed, and estimate the regression in the full sample. They may also estimate the regression in first difference and without group fixed effects. Finally, they may include control variables. Results similar to Theorem 20 below apply to all those variations on (8.21).

A decomposition of the coefficients in distributed-lag regressions.

Theorem 20 Suppose that Assumptions NA and PT hold, and that for all g and $t \geq K+1$ there exists real numbers $(\gamma_{g,t}^\ell)_{\ell \in \{0, \dots, K\}}$ such that for all \mathbf{d} in the support of \mathbf{D} , $Y_{g,t}(\mathbf{d}) = Y_{g,t}(\mathbf{0}_t) + \sum_{\ell=0}^K \gamma_{g,t}^\ell d_{t-\ell}$. Then, for all $\ell \in \{0, \dots, K\}$,

$$E[\hat{\beta}_\ell^{dl}] = \sum_{\substack{(g,t): D_{g,t-\ell} \neq 0, \\ t \geq K+1}} W_{g,t}^{dl,\ell} \gamma_{g,t}^\ell + \sum_{\substack{\ell'=0 \\ \ell' \neq \ell}}^K \sum_{\substack{(g,t): D_{g,t-\ell'} \neq 0, \\ t \geq K+1}} W_{g,t}^{dl,\ell} \gamma_{g,t}^{\ell'},$$

where

$$\sum_{\substack{(g,t): D_{g,t-\ell} \neq 0, \\ t \geq K+1}} W_{g,t}^{dl,\ell} = 1, \quad \sum_{\substack{(g,t): D_{g,t-\ell'} \neq 0, \\ t \geq K+1}} W_{g,t}^{dl,\ell} = 0 \quad \forall \ell' \neq \ell.$$

The decomposition in Theorem 20 assumes that

$$Y_{g,t}(\mathbf{d}) = Y_{g,t}(\mathbf{0}_t) + \sum_{\ell=0}^K \gamma_{g,t}^\ell d_{t-\ell}, \tag{8.22}$$

meaning that the functional form of the distributed-lag regression is correctly specified: only the first K treatment lags affect the outcome, their effect is linear and those lags do not interact. Even under those strong assumptions, $\hat{\beta}_\ell^{dl}$, the coefficient on the ℓ th treatment lag, estimates the sum of $K+1$ terms. The first term is a weighted sum of the effect of the ℓ th treatment lag, across all (g, t) cells for which that lag is not equal to 0, with weights that sum to one but may be negative. This term may be biased for the average effect of the ℓ th treatment lag, if that effect varies across (g, t) cells. The remaining K terms are weighted sums of the effects of other treatment lags, with weights summing to zero. If the effects of the other lags vary across (g, t) cells, those terms may differ from zero and may contaminate $\hat{\beta}_\ell^{dl}$.

Application to the newspapers example. Using `gentzkowetal_didtextbook`, regress turnout on the number of newspapers and its lag and on county and year FEs, clustering standard errors at the county level. Interpret the results.

```
areg prestout i.year numdailies lag_numdailies, absorb(cnty90) cluster(cnty90)
```

$\hat{\beta}_0^{dl} = -0.0008$ (s.e.= 0.0014) and $\hat{\beta}_1^{dl} = 0.0050$ (s.e.= 0.0015): according to this distributed-lag TWFE regression, increasing the current number of newspapers insignificantly reduces turnout by 0.08 percentage points, while increasing its first lag significantly increases turnout by 0.5 percentage points. Use the `twowayfeweights` Stata package and the `other_treatments` option to decompose $\hat{\beta}_0^{dl}$ and $\hat{\beta}_1^{dl}$, and interpret the results.

```
twowayfeweights prestout cnty90 year numdailies, other_treatments(lag_numdailies)
type(feTR)

twowayfeweights prestout cnty90 year lag_numdailies, other_treatments(numdailies)
type(feTR)
```

$\hat{\beta}_0^{dl}$ estimates the sum of two terms. The first term is a weighted sum of 10,056 (g, t) -specific effects of current newspapers, where 5,754 effects receive a positive weight, 4,302 receive a negative weight, and where positive (resp. negative) weights sum to 1.85 (resp. -0.85). Negative weights are almost twice as large as in the basic TWFE regression without the lagged treatment. The second term is a weighted sum of 9,339 (g, t) -specific effects of lagged newspapers, where 4,721 effects receive a positive weight, 4,618 receive a negative weight, and where positive (resp. negative) weights sum to 1.21 (resp. -1.21): the contamination of $\hat{\beta}_0^{dl}$ by effects of the lagged treatment is very substantial in this application. Results are similar for $\hat{\beta}_1^{dl}$.

Bibliographic notes. Theorem 20 is an application, to distributed-lag regressions, of one of the decompositions in de Chaisemartin and D'Haultfœuille (2023a) of TWFE regressions with several treatments.

8.3 Heterogeneity-robust estimators

This section introduces the heterogeneity-robust DID estimators that we proposed in de Chaisemartin and D'Haultfœuille (2025), for designs where the treatment may be non-binary and/or non-absorbing, and where the outcome may be affected by treatment lags. Those estimators are computed by the `did_multiplegt_dyn` Stata and `didmultiplegtdyn` R commands. The syntax of the Stata command is described in Chapter 6.

Date of first treatment change. For all g , let $F_g = \min\{t : t \geq 2, D_{g,t} \neq D_{g,t-1}\}$ denote the date at which a group's treatment changes for the first time. We adopt the convention that $F_g = T + 1$ if g 's treatment never changes. In a binary and staggered design, if no group is treated at $t = 1$ then F_g reduces to the first date at which g is treated, which is why we use the same notation as in Chapter 6.

DID estimators applicable in any design where groups do not all experience their first treatment change at the same date. The estimators below are applicable to any design where the following condition holds:

Design STAY (*Designs with some stayers*) $\exists(g, g')$ such that: (i) $D_{g,1} = D_{g',1}$, (ii) $F_g \neq F_{g'}$.

(i) requires that there exist groups with the same period-one treatment. If groups' period-one treatments are i.i.d. draws from a continuous distribution, $D_{g,1} \neq D_{g',1}$ for all (g, g') , so (i) fails. In Section 8.3.4.2, we will extend the estimators below to designs where (i) fails, so (i) is not really of essence to what follows. (ii) requires that there is heterogeneity in the date at which groups change treatment for the first time. There are many applications where (ii) holds. Still, it fails in designs without stayers, where $D_{g,1} \neq D_{g,2}$ and $F_g = 2$ for all g , as will for instance be the case if $D_{g,t}$ is the amount of rainfall or the average temperature in location g and year

t : all locations will experience different precipitations or temperatures in years one and two. (ii) also fails if groups all change treatment for the first time at the same date t_0 , for instance due to a universal policy affecting them all: $F_g = t_0$ for all g . In such cases, if all groups are untreated at period one and receive heterogeneous treatment doses at t_0 , the design is actually and heterogeneous adoption design and one can then use the estimators reviewed in Chapter 7.

8.3.1 Parallel-trends assumption

Parallel trends if groups' treatment never changes, conditional on groups' period-one treatment. Let $\mathcal{D}_1^r = \{d : \exists(g, g') \in \{1, \dots, G\}^2 : D_{g,1} = D_{g',1} = d, F_g \neq F_{g'}\}$ be the set of period-one-treatment values such that two groups with different values of F_g have that period-one treatment. For any d in \mathcal{D}_1^r and any t , let \mathbf{d}_t denote a $1 \times t$ vector of ds , and let $\mathbf{D}_{g,1,t}$ be a $1 \times t$ vector whose coordinates are all equal to $D_{g,1}$. $Y_{g,t}(\mathbf{D}_{g,1,t})$ is g 's period- t outcome in a counterfactual where it keeps its period-one treatment till period t . We refer to it as its “status quo” potential outcome.

Assumption PTNC (*Parallel trends if groups' treatment never changes, conditional on their period-one treatment*) $\forall(g, g'), \text{ if } D_{g,1} = D_{g',1} \in \mathcal{D}_1^r, \text{ then } \forall t \geq 2,$

$$E[Y_{g,t}(\mathbf{D}_{g,1,t}) - Y_{g,t-1}(\mathbf{D}_{g,1,t-1})] = E[Y_{g',t}(\mathbf{D}_{g',1,t}) - Y_{g',t-1}(\mathbf{D}_{g',1,t-1})].$$

Interpret Assumption PTNC.

Assumption PTNC requires that if two groups have the same period-one treatment, then they have the same expected outcome evolution if their treatment never changes. If all groups are untreated at period 1, $\mathcal{D}_1^r = \{0\}$, so Assumption PTNC is equivalent to Assumption PT, the standard parallel-trends assumption if groups are never treated. Note that Assumption PTNC restricts only one potential outcome per group, so Assumption PTNC alone does not restrict groups' treatment effects.

Comparing Assumption PTNC to unconditional parallel-trends if groups' treatment never changes. Consider the following condition: $\forall(g, g'), \forall t \geq 2$,

$$E[Y_{g,t}(\mathbf{D}_{g,1,t}) - Y_{g,t-1}(\mathbf{D}_{g,1,t-1})] = E[Y_{g',t}(\mathbf{D}_{g',1,t}) - Y_{g',t-1}(\mathbf{D}_{g',1,t-1})]. \quad (8.23)$$

(8.23) is stronger than Assumption PTNC: it requires that all groups, and not just those with the same period-one treatment, have the same expected evolution if their treatment never changes. To simplify the remainder of the discussion, let us assume that treatment is binary and there is at least one group g_0 that is untreated at period 1. Then, (8.23) implies that for all groups g treated at period one and for all $t \geq 2$,

$$E[Y_{g,t}(\mathbf{1}_t) - Y_{g,t-1}(\mathbf{1}_{t-1})] = E[Y_{g_0,t}(\mathbf{0}_t) - Y_{g_0,t-1}(\mathbf{0}_t)],$$

while Assumption PT, the standard parallel-trends assumption on groups' never-treated outcome, implies that

$$E[Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})] = E[Y_{g_0,t}(\mathbf{0}_t) - Y_{g_0,t-1}(\mathbf{0}_t)].$$

Therefore, when combined with Assumption PT, (8.23) implies that for all groups treated at period one and for all $t \geq 2$,

$$\begin{aligned} E[Y_{g,t}(\mathbf{1}_t) - Y_{g,t-1}(\mathbf{1}_{t-1})] &= E[Y_{g,t}(\mathbf{0}_t) - Y_{g,t-1}(\mathbf{0}_{t-1})], \\ \iff E[Y_{g,t}(\mathbf{1}_t) - Y_{g,t}(\mathbf{0}_t)] &= E[Y_{g,t-1}(\mathbf{1}_{t-1}) - Y_{g,t-1}(\mathbf{0}_{t-1})]. \end{aligned} \quad (8.24)$$

Interpret (8.24).

(8.24) means that for initially treated groups, the effect of being treated for t periods should be the same as the effect of being treated for $t - 1$ periods. By iteration, the effect of being treated for t periods should be the same as the effect of being treated for one period. This is an unpalatable restriction: it fails whenever lagged treatments affect the outcome, and it also rules out time-varying effects. By contrast, when combined with the standard parallel-trends

assumption on groups' never-treated outcome, Assumption PTNC implies that for all groups such that $D_{g,1} = 1$ and for all $t \geq 2$,

$$E(Y_{g,t}(\mathbf{1}_t) - Y_{g,t}(\mathbf{0}_t)) - (E(Y_{g,t-1}(\mathbf{1}_{t-1}) - Y_{g,t-1}(\mathbf{0}_{t-1}))) \text{ does not vary across } g.$$

This means that the incremental effect of one treatment period should be the same in every initially treated group. This is a strong restriction, but unlike (8.24), it does not rule out effects of lagged treatments on the outcome, it allows for time-varying effects, and it also allows for heterogeneous effects across groups as the initial effect of being treated for one period can differ across groups.

8.3.2 Target parameters and estimators

8.3.2.1 Non-normalized actual-versus-status-quo event-study effects

Definition of the non-normalized actual-versus-status-quo effects. For every g , let

$$T_g = \max_{g': D_{g',1} = D_{g,1}} F_{g'} - 1$$

denote the last period where there is still a group with the same period-one treatment as g and whose treatment has not changed since the start of the panel. For any g such that $F_g \leq T_g$, and for any $\ell \in \{1, \dots, T_g - (F_g - 1)\}$, let

$$\text{AVSQ}_{g,\ell} = E \left[Y_{g,F_g-1+\ell} - Y_{g,F_g-1+\ell}(D_{g,1}, \dots, D_{g,1}) \right]$$

be the expected difference between group g 's actual outcome at $F_g - 1 + \ell$ and its counterfactual "status quo" outcome if its treatment had remained equal to its period-one value from period one to $F_g - 1 + \ell$. We refer to $\text{AVSQ}_{g,\ell}$ as the actual-versus-status-quo (AVSQ) event-study (ES) effect of g at $F_g - 1 + \ell$. In a binary and staggered design, if groups are all untreated at period 1, then $\text{AVSQ}_{g,\ell}$ reduces to the effect $\text{TE}_{g,\ell}^r$ that we considered in Chapter 6.

Estimation of the non-normalized actual-versus-status-quo effects. For all (g, ℓ) , let

$$\mathcal{C}_{g,\ell} = \{g' : D_{g',1} = D_{g,1}, F_{g'} > F_g - 1 + \ell\}$$

be the set of groups g' with the same period-one treatment as g , and which have kept the same treatment from period 1 to $F_g - 1 + \ell$. Recall that for any set A , $\#A$ denotes its number of elements, i.e. its cardinality. For every g such that $F_g \leq T_g$, and every $\ell \in \{1, \dots, T_g - (F_g - 1)\}$, $\#\mathcal{C}_{g,\ell} > 0$. To estimate $\text{AVSQ}_{g,\ell}$, we use

$$\widehat{\text{AVSQ}}_{g,\ell} = Y_{g,F_g-1+\ell} - Y_{g,F_g-1} - \frac{1}{\#\mathcal{C}_{g,\ell}} \sum_{g' \in \mathcal{C}_{g,\ell}} (Y_{g',F_g-1+\ell} - Y_{g',F_g-1}),$$

a DID estimator comparing the $F_g - 1$ -to- $F_g - 1 + \ell$ outcome evolution of g to that of groups with the same baseline treatment, and that have kept that treatment from period 1 to $F_g - 1 + \ell$, the not-yet-switchers.

Theorem 21 *If Assumptions NA and PTNC hold, then for every (g, ℓ) such that $1 \leq \ell \leq T_g - (F_g - 1)$, $E[\widehat{\text{AVSQ}}_{g,\ell}] = \text{AVSQ}_{g,\ell}$.*

No-crossing condition. Assume that g_0 is such that $D_{g_0,1} = 1, D_{g_0,2} = 2, D_{g_0,3} = 0$. Then,

$$\begin{aligned} \text{AVSQ}_{g_0,2} &= E[Y_{g_0,3}(1, 2, 0) - Y_{g_0,3}(1, 1, 1)] \\ &= E[Y_{g_0,3}(1, 2, 0) - Y_{g_0,3}(1, 1, 0)] - E[Y_{g_0,3}(1, 1, 1) - Y_{g_0,3}(1, 1, 0)] \end{aligned}$$

is the difference between the effect of increasing g_0 's period-2 treatment from 1 to 2, and the effect of increasing g_0 's period-3 treatment from 0 to 1. One could have that both effects are positive but $\text{AVSQ}_{g_0,2}$ is negative. Beyond this example, for all (g, ℓ) such that ℓ periods after its first treatment change, g has experienced both a treatment strictly below and a treatment strictly above its period-one treatment, $\text{AVSQ}_{g,\ell}$ can be written as a linear combination, with negative weights, of the effects of increasing different treatment lags. Throughout this section, we assume away the existence of such (g, ℓ) s.

$$\forall g \in \{1, \dots, G\}, \text{ either } D_{g,t} \geq D_{g,1} \forall t, \text{ or } D_{g,t} \leq D_{g,1} \forall t. \quad (8.25)$$

(8.25) automatically holds if all groups are untreated at baseline. It also holds automatically when the treatment is binary, or when groups' treatment can only change once. When (8.25) fails, one can discard from the sample all cells (g, t) such that, from period two to t , g has experienced both a treatment strictly below and a treatment strictly above its period-one treatment. The `did_multiplegt_dyn` command automatically drops those cells. This yields an unbalanced panel of groups where (8.25) holds by construction, on which the estimators below can be applied. We impose (8.25) to avoid the notational burden of defining estimators on an unbalanced panel.

Definition of the aggregated non-normalized AVSQ ES effects. Let $L = \max_g(T_g - (F_g - 1))$ denote the largest ℓ such that $\text{AVSQ}_{g,\ell}$ can be estimated for at least one g . Under Design STAY, $L \geq 1$. For every $\ell \in \{1, \dots, L\}$, let

$$\mathcal{S}_\ell = \{g : F_g - 1 + \ell \leq T_g\}$$

be the set of groups for which $\text{AVSQ}_{g,\ell}$ can be estimated. For all g such that $F_g \leq T$, let

$$S_g = 1\{D_{g,F_g} > D_{g,1}\} - 1\{D_{g,F_g} < D_{g,1}\}$$

be equal to 1 (resp. -1) for groups whose treatment increases (resp. decreases) at F_g . Then, let

$$\text{AVSQ}_\ell = \frac{1}{\#\mathcal{S}_\ell} \sum_{g \in \mathcal{S}_\ell} S_g \text{AVSQ}_{g,\ell}, \quad (8.26)$$

be the average of $S_g \text{AVSQ}_{g,\ell}$, referred to as non-normalized AVSQ ES effect ℓ . Why is it that for groups such that $S_g = -1$, $\text{AVSQ}_{g,\ell}$ is multiplied by -1 in the definition of AVSQ_ℓ ?

Under (8.25), for groups with $S_g = 1$, $D_{g,t} \geq D_{g,1}$ for all t , so $\text{AVSQ}_{g,\ell}$ is the effect of having been exposed to a weakly higher treatment dose for ℓ periods. Conversely, for groups with $S_g = -1$, $D_{g,t} \leq D_{g,1}$ for all t , so $\text{AVSQ}_{g,\ell}$ is the effect of having been exposed to a weakly lower dose for ℓ periods. Taking the negative of $\text{AVSQ}_{g,\ell}$ for those groups ensures that AVSQ_ℓ is an average effect of having been exposed to a weakly larger dose for ℓ periods. In a binary and staggered design, AVSQ_ℓ reduces to ATT_ℓ , the average effect of having been treated rather than untreated for ℓ periods. Therefore, AVSQ_ℓ generalizes ATT_ℓ to non-binary and/or non-staggered designs.

Estimation of the non-normalized event-study effects. For every $\ell \in \{1, \dots, L\}$, let

$$\widehat{\text{AVSQ}}_\ell = \frac{1}{\#\mathcal{S}_\ell} \sum_{g \in \mathcal{S}_\ell} S_g \widehat{\text{AVSQ}}_{g,\ell}.$$

Theorem 21 implies that $\widehat{\text{AVSQ}}_\ell$ is unbiased for AVSQ_ℓ under Assumptions NA and PTNC.

AVSQ_ℓ is an average effect of being exposed to a weakly higher treatment for ℓ periods, where the magnitude and timing of treatment increments varies across groups.

While in binary and staggered designs AVSQ_ℓ reduces to ATT_ℓ, the interpretation of AVSQ_ℓ is more complicated in non-binary and/or non-staggered designs. For instance, if (8.1) holds ($D_{g,t} = 1\{E_g \geq t \geq F_g\}$), for all g such that $F_g - 1 + \ell > E_g$, group g has exited the treatment at $F_g - 1 + \ell$, and

$$\text{AVSQ}_{g,\ell} = E \left[Y_{g,F_g-1+\ell}(\mathbf{0}_{F_g-1}, \mathbf{1}_{E_g-(F_g-1)}, \mathbf{0}_{F_g-1+\ell-E_g}) - Y_{g,F_g-1+\ell}(\mathbf{0}_{F_g-1+\ell}) \right]$$

is the effect of having been treated for $E_g - (F_g - 1)$ periods, $F_g - 1 + \ell - E_g$ periods before the outcome is measured. Thus, the number and the recency of the treatment periods that generate AVSQ_{g,ℓ} vary across groups, complicating the interpretation of AVSQ_ℓ. Similarly, with three periods and three groups such that ($D_{1,1} = 0, D_{1,2} = 4, D_{1,3} = 0$), ($D_{2,1} = 0, D_{2,2} = 2, D_{2,3} = 3$), and ($D_{2,1} = 0, D_{2,2} = 0, D_{2,3} = 0$), AVSQ₂ is the average of $E(Y_{1,3}(0, 4, 0) - Y_{1,3}(0, 0, 0))$ and $E(Y_{2,3}(0, 2, 3) - Y_{2,3}(0, 0, 0))$. Thus, the magnitude and timing of the treatment increments generating AVSQ_{g,ℓ} varies across groups, which again complicates the interpretation of AVSQ_ℓ.

Treatment-path specific event-study effects? If the design is such that the number of treatment paths is low relative to G , then one may be able to precisely estimate the average of AVSQ_{g,ℓ} separately across all groups with the same path, thus yielding estimates of the average effects of specific treatment paths. The `did_multiplegt_dyn` command estimates treatment-path specific event-study effects when the `by_path` option is specified. For instance, if (8.1) holds ($D_{g,t} = 1\{E_g \geq t \geq F_g\}$) the number of treatment paths may often be low enough for this solution to be practical. But in more complicated designs, the number of paths may be too large for this solution to be practical, especially as ℓ increases. In such instances, we still recommend that researchers report the period-1-to- $F_g - 1 + \ell$ treatment paths and their distribution: this information may be helpful to interpret $\widehat{\text{AVSQ}}_\ell$. The `did_multiplegt_dyn` command reports the paths and their distribution when the `design` option is specified.

Alternative estimators. In a binary and staggered design, $\widehat{\text{AVSQ}}_1$ is numerically equivalent to the DID_M estimator in de Chaisemartin and D'Haultfœuille (2020), and for all ℓ $\widehat{\text{AVSQ}}_\ell$ is numerically equivalent to the event-study estimator of the effect of ℓ periods of exposure to

treatment of Callaway and Sant'Anna (2021), using the not-yet treated as controls. Outside of binary and staggered designs, when all groups are untreated at period one, $\widehat{\text{AVSQ}}_\ell$ is numerically equivalent to the estimator obtained by redefining the treatment as an indicator equal to one if group g 's treatment has ever changed at t , and then computing the event-study estimator of ℓ periods of exposure to treatment of Callaway and Sant'Anna (2021) with this binarized and staggerized treatment. This “binarize and staggerize” idea has for instance been used by Deryugina (2017) or Krolkowski (2018). When groups' period-one treatment varies, the two estimators are not equivalent:² $\widehat{\text{AVSQ}}_\ell$ only compares switchers and not-yet-switchers with the same period-one treatment, whereas the estimator of Callaway and Sant'Anna (2021) applied to this binarized and staggerized treatment compares switchers and non-switchers with different period-one treatments. Then, that estimator relies on (8.23), an assumption which, when combined with Assumption PT, essentially rules out effects of lagged treatments on the outcome, as discussed above.

8.3.2.2 Normalized actual-versus-status-quo event-study effects

Definition and estimation of the normalized actual-versus-status-quo effects. For any g such that $F_g \leq T_g$ and any $\ell \in \{1, \dots, T_g - (F_g - 1)\}$, let

$$\text{AVSQ}_{g,\ell}^D = \sum_{k=0}^{\ell-1} (D_{g,F_g+k} - D_{g,1})$$

be the incremental treatment doses received by g from F_g to $F_g - 1 + \ell$, with respect to the doses it would have received in the status-quo counterfactual. Then, let

$$\text{AVSQ}_{g,\ell}^n = \frac{\text{AVSQ}_{g,\ell}}{\text{AVSQ}_{g,\ell}^D}.$$

We refer to $\text{AVSQ}_{g,\ell}^n$ as the normalized AVSQ effect of group g at $F_g - 1 + \ell$. It follows directly from Theorem 21 that $\widehat{\text{AVSQ}}_{g,\ell}/\text{AVSQ}_{g,\ell}^D$ is unbiased for $\text{AVSQ}_{g,\ell}^n$.

²For instance, East, Miller, Page and Wherry (2023) consider designs where groups' period-one treatment varies, and binarize and staggerize the treatment and compute the event-study estimators of Callaway and Sant'Anna (2021).

AVSQ_{g,ℓ}ⁿ is a weighted average of the effects of the current and $\ell - 1$ first treatment lags. For $k \in \{0, \dots, \ell - 1\}$, let³

$$\begin{aligned} s_{g,\ell,k} = & E \left[Y_{g,F_g-1+\ell}(\mathbf{D}_{g,1,F_g-1}, D_{g,F_g}, \dots, D_{g,F_g-1+\ell-k-1}, \underline{D_{g,F_g-1+\ell-k}}, \mathbf{D}_{g,1,k}) \right. \\ & \left. - Y_{g,F_g-1+\ell}(\mathbf{D}_{g,1,F_g-1}, D_{g,F_g}, \dots, D_{g,F_g-1+\ell-k-1}, \underline{D_{g,1}}, \mathbf{D}_{g,1,k}) \right] / (D_{g,F_g-1+\ell-k} - D_{g,1}) \end{aligned}$$

be the slope of the expected potential outcome function of group g at $F_g - 1 + \ell$ with respect to its k th treatment lag (the underlined term), when that lag is switched from its status-quo counterfactual value $D_{g,1}$ to its actual value $D_{g,F_g-1+\ell-k}$, whereas all its previous treatments are held at their actual values, and all its subsequent treatments are held at their status-quo value.

For any $k \in \{0, \dots, \ell - 1\}$, let

$$w_{g,\ell,k} = \frac{D_{g,F_g-1+\ell-k} - D_{g,1}}{\text{AVSQ}_{g,\ell}^D}.$$

Theorem 22 For every (g, ℓ) such that $1 \leq \ell \leq T_g - (F_g - 1)$, $\text{AVSQ}_{g,\ell}^n = \sum_{k=0}^{\ell-1} w_{g,\ell,k} s_{g,\ell,k}$. Moreover, $\sum_{k=0}^{\ell-1} w_{g,\ell,k} = 1$ for all (g, ℓ) and under (8.25), $w_{g,\ell,k} \geq 0$.

Interpret Theorem 22.

Theorem 22 shows that $\text{AVSQ}_{g,\ell}^n$ is a weighted average of the slopes of g 's potential outcome at $F_g - 1 + \ell$ with respect to its $\ell - 1$ first treatment lags, where for $k \in \{0, \dots, \ell - 1\}$, the effect of the k th lag receives a weight proportional to the absolute value of the difference between g 's k th treatment lag and its status-quo treatment.

Theorem 22 in some specific designs. For concreteness, we rewrite the result in Theorem 22 in two specific designs. First, in binary-and-staggered designs, Theorem 22 reduces to

$$\text{AVSQ}_{g,\ell}^n = \frac{1}{\ell} \sum_{k=0}^{\ell-1} E \left[Y_{g,F_g-1+\ell}(\mathbf{0}_{F_g-1}, \mathbf{1}_{\ell-k-1}, 1, \mathbf{0}_k) - Y_{g,F_g-1+\ell}(\mathbf{0}_{F_g-1}, \mathbf{1}_{\ell-k-1}, 0, \mathbf{0}_k) \right].$$

³We use the convention that for $k = 0$ and any d , \mathbf{d}_k stands for the empty vector. Accordingly when $k = \ell - 1$, $(D_{g,F_g}, \dots, D_{g,F_g-1+\ell-k-1})$ also stands for the empty vector. We sometimes refer to a cell's current treatment as its 0th treatment lag, and use the convention $0/0=0$.

Then, $\text{AVSQ}_{g,\ell}^n$ is the simple average, across k ranging from 0 to $\ell - 1$, of the effect of switching the k th treatment lag from 0 to 1, holding previous treatments at 1 and subsequent treatments at 0. Second, if (8.2) holds ($D_{g,t} = I_g \{t \geq F_g\}$), Theorem 22 reduces to

$$\text{AVSQ}_{g,\ell}^n = \frac{1}{\ell} \sum_{k=0}^{\ell-1} \frac{E \left[Y_{g,F_g-1+\ell}(\mathbf{0}_{F_g-1}, \mathbf{I}_{g,\ell-k-1}, I_g, \mathbf{0}_k) - Y_{g,F_g-1+\ell}(\mathbf{0}_{F_g-1}, \mathbf{I}_{g,\ell-k-1}, 0, \mathbf{0}_k) \right]}{I_g}.$$

Thus, in staggered designs with group-specific treatment intensities, $\text{AVSQ}_{g,\ell}^n$ is the average, across k ranging from 0 to $\ell - 1$, of the effect of switching the k th lag from 0 to I_g , normalized by I_g .

Definition and estimation of the aggregated normalized AVSQ ES effects. Let

$$\text{AVSQ}_\ell^D = \frac{1}{\#\mathcal{S}_\ell} \sum_{g \in \mathcal{S}_\ell} |\text{AVSQ}_{g,\ell}^D|.$$

For $\ell \in \{1, \dots, L\}$, let

$$\text{AVSQ}_\ell^n = \frac{1}{\#\mathcal{S}_\ell} \sum_{g \in \mathcal{S}_\ell} \frac{|\text{AVSQ}_{g,\ell}^D|}{\text{AVSQ}_\ell^D} \text{AVSQ}_{g,\ell}^n.$$

Note the following relation between the non-normalized and normalized event-study effects:

$$\text{AVSQ}_\ell^n = \frac{\text{AVSQ}_\ell}{\text{AVSQ}_\ell^D}. \quad (8.27)$$

AVSQ_ℓ^n is a weighted average of all the $\text{AVSQ}_{g,\ell}^n$ that can be estimated, with weights proportional to $|\text{AVSQ}_{g,\ell}^D|$.⁴ It follows directly from Theorem 21 that

$$\widehat{\text{AVSQ}}_\ell^n := \frac{1}{\#\mathcal{S}_\ell} \sum_{g \in \mathcal{S}_\ell} \frac{|\text{AVSQ}_{g,\ell}^D|}{\text{AVSQ}_\ell^D} \frac{\widehat{\text{AVSQ}}_{g,\ell}}{\text{AVSQ}_{g,\ell}^D}$$

is unbiased for AVSQ_ℓ^n . As in (8.27), there is the following relationship between the normalized and non-normalized ES estimators:

$$\widehat{\text{AVSQ}}_\ell^n = \widehat{\text{AVSQ}}_\ell / \text{AVSQ}_\ell^D.$$

⁴In designs where some g s are such that $|\text{AVSQ}_{g,\ell}^D|$ is close to zero, the estimator of the unweighted average of the $\text{AVSQ}_{g,\ell}^n$ s will suffer from a small-denominator problem, similar to that we already discussed in Chapter 7 when we compared the ATT and the WATT. This is what leads us to consider a weighted average of the $\text{AVSQ}_{g,\ell}^n$.

AVSQ $_{\ell}^n$ is a weighted average of the effects of the current and $\ell - 1$ first treatment lags, with weights that can be computed. It follows from Theorem 22 that AVSQ $_{\ell}^n$ is a weighted average of the effects of groups' current and $\ell - 1$ first treatment lags on their outcome. The total weight assigned by AVSQ $_{\ell}^n$ to the effect of the k th-lag (for $0 \leq k \leq \ell - 1$) is equal to

$$w_{\ell,k} = \frac{1}{\#\mathcal{S}_{\ell}} \sum_{g \in \mathcal{S}_{\ell}} \frac{|D_{g,F_g-1+\ell-k} - D_{g,1}|}{\text{AVSQ}_{\ell}^D}.$$

Compute $w_{\ell,k}$ in designs where groups' treatment can only change once, meaning that $D_{g,t} = D_{g,F_g}$ for all $t \geq F_g$.

When groups' treatment can only change once, $\text{AVSQ}_{g,\ell}^D = \sum_{k=0}^{\ell-1} (D_{g,F_g+k} - D_{g,1}) = \ell(D_{g,F_g} - D_{g,1})$. Therefore,

$$\text{AVSQ}_{\ell}^D = \ell \frac{1}{\#\mathcal{S}_{\ell}} \sum_{g \in \mathcal{S}_{\ell}} |D_{g,F_g} - D_{g,1}|,$$

so $w_{\ell,k} = 1/\ell$. Then, AVSQ $_1^n$ is an effect of the current treatment on the outcome, AVSQ $_2^n$ is a weighted average of the effect of the current treatment and of the first treatment lag on the outcome with weights 1/2, AVSQ $_3^n$ is a weighted average of the effect of the current treatment and of the first and second treatment lag on the outcome with weights 1/3, etc. When groups' treatment can change more than once, we recommend reporting $k \mapsto w_{\ell,k}$, to document which lags contribute the most to AVSQ $_{\ell}^n$. The `did_multiplegt_dyn` command reports $k \mapsto w_{\ell,k}$ when the `normalized_weights` option is specified.

Estimating separately the effect of the current and lagged treatments.* Researchers estimating distributed-lag TWFE regressions seek to separately estimate the effect of the current and lagged treatments on the outcome. By estimating normalized AVSQ ES effects, they can estimate weighted averages of the effects of the current and lagged treatments on the outcome, thus fulfilling a related but different estimation goal. For instance, without further assumptions, one cannot use AVSQ $_2^n$ to tease out the effect of the current treatment and of its first lag on the outcome. In a working paper version of de Chaisemartin and D'Haultfœuille (2025) (see Section

4 of de Chaisemartin and D'Haultfœuille, 2021), it is shown that if one assumes that

$$Y_{g,t}(\mathbf{d}) = Y_{g,t}(\mathbf{0}_t) + \sum_{l=0}^K \gamma_g^l d_{t-l},$$

meaning that lags' effects are additively separable, linear, and constant over time, then there is an invertible linear system relating $(\gamma_g^l)_{l \in \{0, \dots, T_g - F_g\}}$ and $(AVSQ_{g,\ell})_{\ell \in \{1, \dots, T_g - (F_g - 1)\}}$. Therefore, $(\gamma_g^l)_{l \in \{0, \dots, T_g - F_g\}}$ can be unbiasedly estimated, and averages of the effects of the current and lagged treatments on the outcome across groups can also be unbiasedly estimated. Those estimators allow for heterogeneous effects across groups, unlike distributed-lag TWFE estimators, but they still rely on strong linearity and separability assumptions. We are not aware of a Stata or R package computing those estimators.

8.3.2.3 Cost-benefit analysis

In this section, we strengthen (8.25) by assuming that groups always have a weakly-larger treatment than their period-one treatment:

$$\forall g \in \{1, \dots, G\}, D_{g,t} \geq D_{g,1} \quad \forall t. \quad (8.28)$$

We impose (8.28) to reduce the notational burden. When it fails, one can just conduct the cost-benefit analysis separately for groups with $S_g = 1$ and for groups with $S_g = -1$.

Cost-benefit analysis. In this section, our target parameter is

$$ACE := \frac{\sum_{g: F_g \leq T_g} \sum_{\ell=1}^{T_g - (F_g - 1)} AVSQ_{g,\ell}}{\sum_{g: F_g \leq T_g} \sum_{\ell=1}^{T_g - (F_g - 1)} (D_{g,F_g-1+\ell} - D_{g,1})}.$$

As explained below, ACE corresponds to an average cumulative effect per unit of treatment, whence its name. To motivate this parameter, let us take the perspective of a planner, seeking to conduct a cost-benefit analysis comparing groups' actual treatments \mathbf{D} to the counterfactual "status-quo" scenario where they would have always kept their period-one treatment. Assume that the outcome is a measure of output, such as agricultural yields or wages, expressed in monetary units. Assume also that the treatment is costly, with a cost linear in dose, and known to the analyst. Then, let $c_{g,\ell} \geq 0$ denote the cost of administering one treatment unit in group g at period $F_g - 1 + \ell$. Assuming that the planner's discount factor is equal to 1, groups' actual

treatments are beneficial in monetary terms relative to the status quo, up to period T_g , if and only if

$$\sum_{g:F_g \leq T_g} \sum_{\ell=1}^{T_g-(F_g-1)} \text{AVSQ}_{g,\ell} - \sum_{g:F_g \leq T_g} \sum_{\ell=1}^{T_g-(F_g-1)} c_{g,\ell}(D_{g,F_g-1+\ell} - D_{g,1}) > 0 \Leftrightarrow \text{ACE} > c,$$

where

$$c = \frac{\sum_{g:F_g \leq T_g} \sum_{\ell=1}^{T_g-(F_g-1)} c_{g,\ell}(D_{g,F_g-1+\ell} - D_{g,1})}{\sum_{g:F_g \leq T_g} \sum_{\ell=1}^{T_g-(F_g-1)} (D_{g,F_g-1+\ell} - D_{g,1})}$$

is the average treatment cost, across all the incremental treatment doses received with respect to the status-quo counterfactual. Then, comparing the ACE to the per-unit treatment cost is sufficient to evaluate if changing groups' treatments from their initial treatments to \mathbf{D} was beneficial.

In binary-and-staggered designs, if no group is treated at period one and there are never-treated groups, the ACE reduces to a well-known treatment-effect parameter, which one?

In binary-and-staggered designs, the ACE reduces to the ATT. In binary-and-staggered designs, if no group is treated at period one and there are never-treated groups,

$$\text{ACE} = \frac{1}{N_1} \sum_{(g,t):D_{g,t}=1} \text{TE}_{g,t},$$

so the ACE reduces to the ATT. Thus, the ACE generalizes the ATT to non-binary and/or non-staggered designs.

The ACE is an average cumulative effect per unit of treatment. Let us first consider a simple example with two groups and four periods, such that $D_{1,1} = 0$, $D_{1,2} = 1$, $D_{1,3} = 1$, and $D_{1,4} = 0$, while group 2 is never treated. In this example, the formula for the ACE reduces to

$$\begin{aligned} \text{ACE} &= \frac{E[Y_{1,2}(0,1) - Y_{1,2}(0,0) + Y_{1,3}(0,1,1) - Y_{1,3}(0,0,0) + Y_{1,4}(0,1,1,0) - Y_{1,4}(0,0,0,0)]}{1+1+0} \\ &= \frac{1}{2} E[Y_{1,2}(0,1) - Y_{1,2}(0,0) + Y_{1,3}(0,1,0) - Y_{1,3}(0,0,0) + Y_{1,4}(0,1,0,0) - Y_{1,4}(0,0,0,0)] \\ &\quad + \frac{1}{2} E[Y_{1,3}(0,1,1) - Y_{1,3}(0,1,0) + Y_{1,4}(0,1,1,0) - Y_{1,4}(0,1,0,0)]. \end{aligned} \tag{8.29}$$

The first expectation in (8.29) is the cumulative effect produced by group 1's period-2 treatment, at periods 2, 3, and 4, relative to the situation where it would have always remained untreated. The second expectation in (8.29) is the cumulative effect produced by group 1's period-3 treatment, at periods 3 and 4, conditional on its period-2 treatment and relative to the situation where it would have been untreated at periods 3 and 4. Accordingly, ACE is the average effect of the two treatment doses received by group 1, cumulated across all periods after each of those two doses are received. A similar interpretation holds beyond this simple example.

For $k \in \{0, \dots, \ell - 1\}$, let $\text{AVSQ}_{g,\ell,k}$ be the numerator of the slope $s_{g,\ell,k}$ defined above. $\text{AVSQ}_{g,\ell,k}$ is the effect, on the expected potential outcome of group g at $F_g - 1 + \ell$, of switching its k th treatment lag from its status-quo to its actual value, whereas all its previous treatments are held at their actual values, and all its subsequent treatments are held at their status-quo value. As

$$\text{AVSQ}_{g,\ell} = \sum_{k=0}^{\ell-1} \text{AVSQ}_{g,\ell,k},$$

$$\text{ACE} = \frac{\sum_{g:F_g \leq T_g} \sum_{\ell=1}^{T_g-(F_g-1)} \sum_{k=0}^{\ell-1} \text{AVSQ}_{g,\ell,k}}{\sum_{g:F_g \leq T_g} \sum_{\ell=1}^{T_g-(F_g-1)} (D_{g,F_g-1+\ell} - D_{g,1})} = \frac{\sum_{g:F_g \leq T_g} \sum_{k=0}^{T_g-F_g} \sum_{\ell=k+1}^{T_g-(F_g-1)} \text{AVSQ}_{g,\ell,k}}{\sum_{g:D_{g,1}=0, F_g \leq T_g} \sum_{k=0}^{T_g-F_g} (D_{g,F_g+k} - D_{g,1})}.$$

$\sum_{\ell=k+1}^{T_g-(F_g-1)} \text{AVSQ}_{g,\ell,k}$ is the cumulative effect, from period $F_g + k$ to T_g , of switching g 's period $F_g + k$ treatment from $D_{g,1}$ to D_{g,F_g+k} . The sum of cumulative effects in the numerator of ACE is scaled by the sum of all the incremental treatments $D_{g,F_g+k} - D_{g,1}$ that generate those cumulative effects. Accordingly, ACE may be interpreted as an average cumulative effect per unit of treatment.

Average number of time periods over which the effect of a dose is cumulated. The effect of switching g 's period $F_g + k$ treatment from $D_{g,1}$ to D_{g,F_g+k} is cumulated from period $F_g + k$ to T_g , namely over $T_g - F_g - k + 1$ periods. To interpret ACE, one may compute

$$\frac{\sum_{g:F_g \leq T_g} \sum_{k=0}^{T_g-F_g} (D_{g,F_g+k} - D_{g,1})(T_g - F_g - k + 1)}{\sum_{g:D_{g,1}=0, F_g \leq T_g} \sum_{k=0}^{T_g-F_g} (D_{g,F_g+k} - D_{g,1})},$$

the average number of time periods over which the effect of a dose is cumulated, across all incremental doses received by switchers over the study period. One may then divide ACE by this average number of time periods, to get an average effect of being exposed to one dose of treatment for one period. The `did_multiplegt_dyn` command reports this average number of time periods.

Estimating ACE. It follows directly from Theorem 21 that

$$\widehat{\text{ACE}} := \frac{\sum_{g:F_g \leq T_g} \sum_{\ell=1}^{T_g - (F_g - 1)} \widehat{\text{AVSQ}}_{g,\ell}}{\sum_{g:F_g \leq T_g} \sum_{\ell=1}^{T_g - (F_g - 1)} (D_{g,F_g-1+\ell} - D_{g,1})}$$

is unbiased for ACE.

8.3.3 Inference

de Chaisemartin and D'Haultfœuille (2025) propose analytic confidence intervals (CIs) for the AVSQ_ℓ , AVSQ_ℓ^n , and ACE effects, based on asymptotic approximations where the number of groups goes to infinity, under the assumption that groups are independent. We recommend, as in Section 3.3.2, that researchers using those CIs with less than 40 switching groups or with less than 40 control groups perform simulations tailored to their data to assess their coverage rate.⁵ Those CIs can be conservative (overly wide) when there are values of $(D_{g,1}, F_g, D_{g,F_g})$ such that only one group has that value: with only one group, the variance of the outcome evolution across groups with that value cannot be unbiasedly estimated. If this issue comes from the fact that $(D_{g,1}, D_{g,F_g})$ takes many different values, then researchers may treat the treatment as a continuous treatment variable and use the estimators proposed in Section 8.3.4.2 below. If this issue comes from the fact that F_g takes many different values, researchers may coarsen their time variable (e.g. aggregate a daily panel at the weekly level), to ensure that most values of F_g have at least two groups. Finally, they may also use bootstrap instead of analytic CIs.

8.3.4 Extensions

8.3.4.1 Pre-trend estimators

de Chaisemartin and D'Haultfœuille (2025) propose pre-trend estimators one can use to test Assumptions NA and PTNC. For any $g : 3 \leq F_g \leq T_g$ and $\ell \in \{1, \dots, \min(T_g - (F_g - 1), F_g - 2)\}$,

⁵When estimating AVSQ_ℓ , the number of treated groups is $\#\mathcal{S}_\ell$. The number of control groups is the cardinality of the union of $\{g' : D_{g',1} = D_{g,1}, F_{g'} > F_g - 1 + \ell\}$ across all g for which $\text{AVSQ}_{g,\ell}$ can be estimated. In words, the number of control groups is the number of groups used as controls to estimate $\text{AVSQ}_{g,\ell}$ for at least one g .

let

$$\widehat{\text{AVSQ}}_{g,-\ell} = Y_{g,F_g-1-\ell} - Y_{g,F_g-1} - \frac{1}{\#\mathcal{C}_{g,\ell}} \sum_{g' \in \mathcal{C}_{g,\ell}} (Y_{g',F_g-1-\ell} - Y_{g',F_g-1}).$$

$\widehat{\text{AVSQ}}_{g,-\ell}$ mimicks $\widehat{\text{AVSQ}}_{g,\ell}$. Like $\widehat{\text{AVSQ}}_{g,\ell}$, it compares the outcome evolution of g to that of groups with the same baseline treatment as g , and that have not switched treatment yet at $F_g - 1 + \ell$. But unlike $\widehat{\text{AVSQ}}_{g,\ell}$, it compares those groups' outcome evolutions from period $F_g - 1$ to period $F_g - 1 - \ell$, namely before group g 's treatment changes for the first time. Accordingly, $\widehat{\text{AVSQ}}_{g,-\ell}$ assesses if g and its control groups experience the same evolution of their status-quo outcome over ℓ periods, the number of periods over which parallel trends has to hold for $\widehat{\text{AVSQ}}_{g,\ell}$ to be unbiased for $\text{AVSQ}_{g,\ell}$. One can show that under Assumptions NA and PTNC,

$$E(\widehat{\text{AVSQ}}_{g,-\ell}) = 0.$$

Then, let $\mathcal{S}_\ell^{\text{pl}} = \{g : 1 \leq F_g - 1 - \ell, F_g - 1 + \ell \leq T_g\}$ be the set of groups for which $\widehat{\text{AVSQ}}_{g,\ell}$ can be computed ($F_g - 1 + \ell \leq T_g$) and for which $\widehat{\text{AVSQ}}_{g,-\ell}$ can also be computed ($F_g - 1 - \ell \geq 1 \Leftrightarrow \ell \leq F_g - 2$), and let

$$\widehat{\text{AVSQ}}_{-\ell} = \frac{1}{\#\mathcal{S}_\ell^{\text{pl}}} \sum_{g \in \mathcal{S}_\ell^{\text{pl}}} S_g \widehat{\text{AVSQ}}_{g,-\ell}$$

be a pre-trends estimator averaging $\widehat{\text{AVSQ}}_{g,-\ell}$ across those groups. **XDH** Accordingly, we define the normalized pre-trend estimator as

$$\widehat{\text{AVSQ}}_{-\ell}^n = \widehat{\text{AVSQ}}_{-\ell} / \widehat{\text{AVSQ}}_\ell^D.$$

The $\widehat{\text{AVSQ}}_{-\ell}$ and $\widehat{\text{AVSQ}}_{-\ell}^n$ estimators are computed by the `did_multiplegt_dyn` command, when the `placebo(#)` option is specified.

8.3.4.2 Continuous treatment

The estimators above compare switchers and not-yet-switchers with the same period-one treatment. The challenge with a continuous treatment is that the sample does not contain switchers and not-yet-switchers with the same period-one treatment. Then, to estimate $E[Y_{g,t}(\mathbf{D}_{g,1,t}) - Y_{g,t-1}(\mathbf{D}_{g,1,t-1})]$, a switcher's outcome evolution in the status-quo counterfactual, we cannot just use the average $t - 1$ -to- t outcome evolution of not-yet-switchers with the same $D_{g,1}$. To circumvent this issue, de Chaisemartin and D'Haultfœuille (2025) propose to replace Assumption

PTNC, the parallel-trends assumption on the status-quo outcome, by the following, stronger condition: for all $t \geq 2$,

$$E[Y_{g,t}(\mathbf{D}_{g,1,t}) - Y_{g,t-1}(\mathbf{D}_{g,1,t-1})] = \sum_{k=0}^K \gamma_{k,t} D_{g,1}^k, \quad (8.30)$$

for some integer K . On top of assuming that groups with the same-period one treatment all have the same counterfactual outcome trends if their treatment does not change, (8.30) also assumes a functional form, namely a degree- K polynomial, for how those counterfactual outcome trends vary with $D_{g,1}$. Then, $(\gamma_{0,t}, \dots, \gamma_{K,t})$ can be unbiasedly estimated by regressing $Y_{g,t} - Y_{g,t-1}$ on $(1, D_{g,1}, \dots, D_{g,1}^K)_{k=0,\dots,K}$, in the sample of groups such that $F_g > t$. With those estimators in hand, one can use

$$\sum_{t=F_g}^{F_g-1+\ell} \sum_{k=0}^K \hat{\gamma}_{k,t} D_{g,1}^k$$

to estimate switchers' counterfactual $F_g - 1$ to $F_g - 1 + \ell$ outcome evolutions if their treatment had not switched. Finally, one can use

$$Y_{g,F_g-1+\ell} - Y_{g,F_g-1} - \sum_{t=F_g}^{F_g-1+\ell} \sum_{k=0}^K \hat{\gamma}_{k,t} D_{g,1}^k$$

to estimate their AVSQ $_{g,\ell}$ effect. The corresponding estimators are computed by the `did_multiplegt_dyn` command, when the `continuous(#)` option is specified. The option's argument is K , the polynomial degree assumed by the researcher in (8.30). Note that those estimators are closely related to the DID estimators with covariates discussed in Chapter 4, with the polynomial in the baseline treatment $(1, D_{g,1}, \dots, D_{g,1}^K)_{k=0,\dots,K}$ playing the role of the covariates. Importantly, and like some of the DID estimators with covariates discussed in Chapter 4, the estimators discussed in this section are parametric and rely on the researcher's choice of a functional form for groups' outcome evolution under the status-quo counterfactual. Proposing a non-parametric estimator could be done, leveraging ideas similar to those in de Chaisemartin et al. (2022), a paper discussed below: this is an interesting avenue for future research.

8.3.4.3 Testing if lagged treatments affect the outcome

A joint test that lagged treatments do not affect the outcome and that treatment effects do not change over time. In this paragraph, we assume that groups' treatment can only change once, meaning that $D_{g,t} = D_{g,F_g}$ for all $t \geq F_g$. If that condition is not met, the

result below still holds, restricting attention to (g, t) cells such that t is strictly before g 's second treatment change ($t < \min\{t' : t' > F_g, D_{g,t'} \neq D_{g,t'-1}\}$). If $D_{g,t} = D_{g,F_g}$ for all $t \geq F_g$,

$$\text{AVSQ}_{g,\ell} = E \left[Y_{g,F_g-1+\ell}(\mathbf{D}_{g,1,F_g-1}, \mathbf{D}_{g,F_g,\ell}) - Y_{g,F_g-1+\ell}(\mathbf{D}_{g,1,F_g-1+\ell}) \right].$$

Now, assume that Assumption ND holds: lagged treatments cannot affect the outcome. Then,

$$\text{AVSQ}_{g,\ell} = E \left[Y_{g,F_g-1+\ell}(D_{g,F_g}) - Y_{g,F_g-1+\ell}(D_{g,1}) \right].$$

Then, further assume that for all (g, t, d, d') ,

$$E [Y_{g,t}(d) - Y_{g,t}(d')] = \delta_{g,d,d'} : \quad (8.31)$$

the effect of the current treatment on the outcome does not depend on time. Then,

$$\text{AVSQ}_{g,\ell} = \delta_{g,D_{g,F_g},D_{g,1}},$$

which does not depend on ℓ . Finally, for any ℓ' and any $\ell \in \{1, \dots, \ell'\}$, let

$$\text{AVSQ}_{\ell,\ell'}^{bal} = \frac{1}{\#\mathcal{S}_{\ell'}} \sum_{g \in \mathcal{S}_{\ell'}} \text{AVSQ}_{g,\ell}$$

denote a version of AVSQ_ℓ , defined on the same subsample of groups as $\text{AVSQ}_{\ell'}$, thus ensuring that for every $\ell \in \{1, \dots, \ell'\}$ $\text{AVSQ}_{\ell,\ell'}^{bal}$ applies to the same set of groups. Then, as Assumption ND and (8.31) imply that $\text{AVSQ}_{g,\ell}$ does not depend on ℓ , those two assumptions also imply that $\text{AVSQ}_{\ell,\ell'}^{bal}$ does not depend on ℓ , a testable condition. The corresponding test is computed by the `did_multiplegt_dyn` command, when the `effects_equal` and `same_switchers` options are specified.

A test that lagged treatments do not affect the outcome, in some specific designs.

In designs with a binary treatment and where some groups leave the treatment after having been previously treated, Liu et al. (2024) propose another test of Assumption ND. Their test amounts to estimating the average treatment effect across previously treated groups, at time periods where those groups have left the treatment. Under Assumption ND, this average treatment effect should be equal to zero. This yields a test of Assumption ND alone, rather than a joint test of Assumption ND and (8.31), an advantage with respect to the test described in the previous paragraph. A disadvantage of the test of Liu et al. (2024) is that it can only be used in designs with a binary treatment and where some treated groups leave the treatment. Their test is implemented by the `fetc` Stata (Liu et al., 2022b) and R (Liu et al., 2022a) commands.

8.3.4.4 Estimators with control variables

de Chaisemartin and D'Haultfœuille (2025) also propose estimators relying on a conditional parallel-trends assumption, which extend the DID estimators with covariates reviewed in Section 4.1 to non-binary and/or non-staggered designs. Estimators relying on a conditional parallel-trends assumption with a linear functional form are computed by the `did_multiplegt_dyn` command, when the `controls` option is specified. Estimators relying on a non-parametric conditional parallel-trends assumption are computed when the `trends_nonparam` option is specified. Only time-invariant variables can be inputted to that option, and the interaction of those variables has to be coarser than the group variable. For instance, if one works with a panel of firms and one wants to allow for industry-specific trends, one should specify `trends_nonparam(industry)`. Finally, estimators allowing for group-specific linear trends are computed when the `trends_lin` option is specified.

8.3.4.5 Estimating heterogeneous treatment effects

Assume that one wants to assess if treatment effects are correlated with a $K \times 1$ vector of time-invariant covariates X_g , whose first coordinate is a constant. In this section, we extend the method described in Section 6.4.1 to non-binary and/or non-staggered designs.

Target parameter. Let $\beta_{\ell,X}$ be the coefficient in an infeasible regression of $S_g \text{AVSQ}_{g,\ell}$, the effect of having been exposed to a weakly higher treatment for ℓ periods in group g on X_g and indicators for all possible values of $(F_g, D_{g,1}, S_g)$, in the sample of groups such that $F_g - 1 + \ell \leq T_g$. $X_g^T \beta_{\ell,X}$ is the best linear predictor of $S_g \text{AVSQ}_{g,\ell}$ given X_g and indicators for all possible values of $(F_g, D_{g,1}, S_g)$. As in Chapter 6, the coefficient from a regression where indicators for all possible values of $(F_g, D_{g,1}, S_g)$ are not controlled for would arguably be a more natural target, but $\beta_{\ell,X}$ is easier to estimate.

Estimator. To estimate $\beta_{\ell,X}$, de Chaisemartin and D'Haultfœuille (2025) propose to use $\widehat{\beta}_{\ell,X}$, the coefficient on X_g in an OLS regression of $S_g(Y_{g,F_g-1+\ell} - Y_{g,F_g-1})$ on X_g and indicators for all possible values of $(F_g, D_{g,1}, S_g)$, in the sample of groups such that $F_g - 1 + \ell \leq T_g$. Let $\beta_{\ell,X}^{\Delta Y(\mathbf{D}_{g,1,t})}$ be the coefficient in a regression of $E[Y_{g,F_g-1+\ell}(\mathbf{D}_{g,1,F_g-1+\ell}) - Y_{g,F_g-1}(\mathbf{D}_{g,1,F_g-1})]$ on X_g and

indicators for all possible values of $(F_g, D_{g,1}, S_g)$, in the sample of groups such that $F_g - 1 + \ell \leq T_g$. If the covariates are non stochastic or conditioned upon, and if for all $k \in \{2, \dots, K\}$ $\beta_{k,\ell,X}^{\Delta Y(D_{g,1,t})} = 0$, meaning that the status-quo outcome evolutions of groups reaching ℓ periods of exposure to treatment before T_g are uncorrelated with the non-constant variables in X_g , then de Chaisemartin and D'Haultfoeuille (2025) show that $\hat{\beta}_{\ell,X,k}$ is unbiased for $\beta_{\ell,X,k}$. $\beta_{k,\ell,X}^{\Delta Y(D_{g,1,t})} = 0$ is placebo testable. To do so, we regress $S_g(Y_{g,F_g-1-\ell} - Y_{g,F_g-1})$ on X_g and indicators for all possible values of $(F_g, D_{g,1}, S_g)$, in the sample of groups such that $F_g - 1 + \ell \leq T_g, F_g - 1 - \ell \geq 2$. Then, we test if the coefficient on X_g is equal to zero. The `did_multiplegt_dyn` Stata and R packages have a `predict_het` option that can be used to compute $\hat{\beta}_{\ell,X}$ and placebo-test $\beta_{k,\ell,X}^{\Delta Y(D_{g,1,t})} = 0$.

8.3.4.6 Estimators with several treatments*

Binary and staggered designs with two consecutive treatments. In Section 3.2 of their web appendix, de Chaisemartin and D'Haultfoeuille (2023a) propose estimators for cases where one is interested in the effect of several, rather than one treatment. To simplify, they start by considering the case with two binary and staggered treatments, where groups can only start receiving the second treatment after they have received the first. This last restriction holds when the second treatment is a reinforcement of the first. For instance, one may want to separately estimate the effects of medical and recreational marijuana laws in the US: so far, states have passed the former before the latter (see Meinhofer, Witman, Hinde and Simon, 2021). Another example are voter ID laws in the US, where less-strict laws are typically passed before stricter ones (see Cantoni and Pons, 2021). Another example are anti-deforestation policies in the Amazon rainforest, where plots of lands are typically put into a concession, and then some concessions get certified (see Rico-Straffon, Wang, Panlasigui, Loucks, Swenson and Pfaff, 2023).

Estimating separately the effect of the first treatment. In such designs, estimating separately the effect of the first treatment is straightforward: one can for instance compute the estimators in Callaway and Sant'Anna (2021) or de Chaisemartin and D'Haultfoeuille (2025), restricting the sample to all (g, t) s that have not received the second treatment. In the marijuana laws example, to estimate the effect of medical marijuana laws, one can just restrict the sample

to all state \times year (g, t) such that state g has not passed a recreational law yet in year t . The horizon until which effects of the first treatment can be estimated will just be truncated by the second treatment.

Estimating the effect of receiving at least one of the two treatments. Estimating the effect of receiving at least one of the two treatments is also straightforward: one can just compute the estimators in Callaway and Sant'Anna (2021) or de Chaisemartin and D'Haultfoeuille (2025), redefining the treatment variable as equal to one for all (g, t) s that have received at least one treatment.

Estimating separately the effect of the second treatment. Estimating separately the effect of the second treatment is more challenging but can still be achieved, under the assumption that the effect of one additional period of exposure to the first treatment is the same in every group. [Intuitively, why is it necessary to impose that assumption, if one wants to use a DID to estimate the incremental effect of the second treatment?](#)

To understand why that assumption is needed, let us go back to the marijuana law example. Without that assumption, a state passing a recreational law may start experiencing a different outcome trend than other states that have only passed a medical law, either because of the recreational law, or because the additional effect of being exposed to the medical law for one more period is different in that state and in other states. Thus, that assumption is key to disentangle the effects of the two treatments. Though it is arguably strong, that assumption is partly testable: it implies that groups that start receiving the first treatment at the same time should have the same outcome evolutions until they adopt the second treatment. Under that assumption, one can estimate the additional effect of the second treatment using, say, the `did_multiplegt_dyn` command, restricting the sample to the (g, t) s that have received the first treatment, and including the adoption date of the first treatment in the `trends_nonparam` option. The resulting estimators compare the outcome evolution of groups that adopt/do not

adopt the second treatment, and that adopted the first treatment at the same date. When the number of groups is relatively low (e.g.: the 50 US states), there may not be any pair of groups receiving the first treatment at the same time period. Then, de Chaisemartin and D'Haultfœuille (2023a) propose two alternative estimators. First, instead of assuming that the effect of one more treatment period is homogeneous across groups, one may assume that the effect of the first treatment evolves linearly with the number of periods of exposure, with a slope that may differ across groups. Under that assumption, one should specify the `trends_lin` instead of the `trends_nonparam` option. Second, one may assume that the effect of one more treatment period is homogeneous across groups and across time periods. Under that assumption, one should include indicators for whether g has been exposed to the first treatment for *at least* 1, 2, etc. periods at period t in the `controls` option.

Separately estimating the effect of entering and leaving the treatment when groups can enter and leave treatment once. In Section 1.6 of their web appendix, de Chaisemartin and D'Haultfœuille (2025) show that similar ideas can be used to separately estimate the effect of entering and leaving the treatment if (8.1) holds ($D_{g,t} = 1\{E_g \geq t \geq F_g\}$), by relabeling entry and exit of treatment as two different treatments. To estimate the effect of joining the treatment, one can use the `did_multiplegt_dyn` command, restricting the sample to all (g, t) s that have not exited the treatment yet. To estimate the effect of leaving the treatment, one should restrict the sample to all (g, t) s such that g has already been treated at t , define the treatment variable as an indicator for having left the treatment, and either control non-parametrically for the date when groups' entered the treatment, or allow for group-specific linear trends, or control for indicators for whether g has been exposed to the first treatment for at least 1, 2, etc. periods at period t .

Extension to binary and staggered designs with two non-consecutive treatments. There may be applications with two binary and staggered treatments, but such that some groups receive treatment one first, other groups receive treatment two first, and other groups receive both treatments at the same time. Then, one can first restrict attention to the subsample of groups that only receive treatment one, or receive both treatments but receive the second one strictly after the first, or do not receive any treatment. In that subsample, one can estimate

the effects of receiving only the first treatment, using the `did_multiplegt_dyn` command and restricting the sample to (g, t) s that have not received the second treatment. One can then estimate the effect of receiving the second treatment when one has already received the first one, restricting the sample to the (g, t) s that have received the first treatment, and specifying the `trends_nonparam`, `trends_lin` or `controls` option as described above. Second, one can restrict attention to the subsample of groups that only receive treatment two, or receive both treatments but receive the first one strictly after the second, or do not receive any treatment. In that subsample, one can estimate the effect of receiving only the second treatment, and the effect of receiving the first treatment when one has already received the second one, using the same steps as above but reverting the roles of the first of second treatments. Finally, one can restrict attention to the subsample of groups that either receive both treatments at the same time or that do not receive any treatment, and estimate the effect of receiving both treatments at the same time in that subsample. Comparing these five sets of estimates may be indicative of whether the treatments are complements or substitutes, even though differences could also be driven by heterogeneous effects across the various subsamples.

Extension to non-binary absorbing treatments. The aforementioned estimators readily extend to absorbing treatments with variation in timing and dose ($D_{g,t} = I_g 1\{t \geq F_g\}$).

8.3.4.7 *The initial-conditions problem in designs where the treatment varies at period one**

When groups receive heterogeneous treatment doses at period one, this may suggest that they have experienced treatment changes before period one. If potential outcomes can depend on any treatment lag, even those before period one, unobserved treatment changes that took place before period one may still affect groups' outcome over the entirety of the study period, which could bias the DID estimators introduced above. Pre-trend tests can be used to assess if, in spite of dynamic effects of treatment changes that took place before the start of the study period, the parallel-trends assumption underlying those estimators remains plausible. Moreover, if one is ready to assume that groups' outcomes can only be affected by their first k treatment lags, then the estimators introduced above can be recomputed, restricting the sample to groups with a stable treatment from period 1 to $k + 1$ and to time periods $k + 1$ to T . In that subsample, those

estimators remain valid even if treatment changes before period one can affect the outcome. An issue with this strategy is that it leads to a reduced sample size, and may yield noisy estimators.

8.3.4.8 Instrumental-variable DID estimators*

To our knowledge, there does not exist yet heterogeneity-robust DID estimators for cases where one is only willing to make a parallel-trends assumption with respect to an instrument, and treatment lags can affect the outcome.

8.3.5 Application to the newspaper example

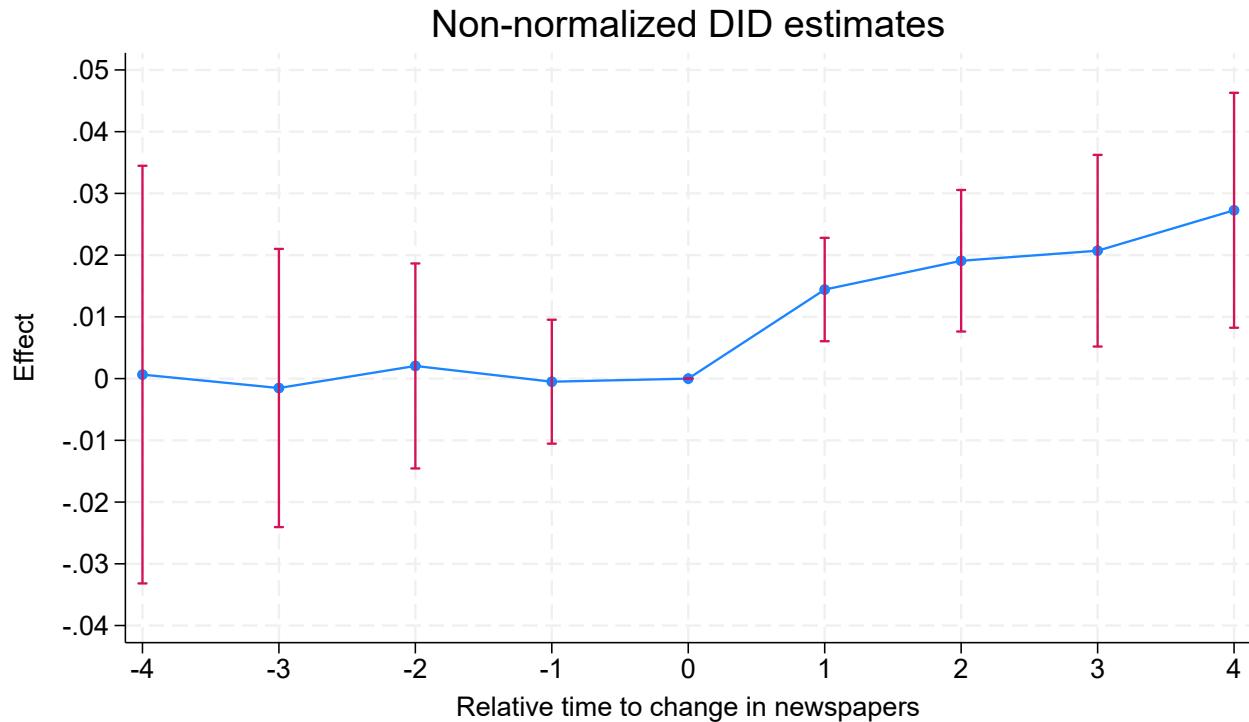
Using `gentzkowetal_didtextbook` and the `did_multiplegt_dyn` command, compute, for $\ell \in \{1, \dots, 4\}$, non-normalized event-study estimates $\widehat{\text{AVSQ}}_\ell$ of the effect of being exposed to a weakly larger number of newspapers for ℓ periods on turnout, as well as a test that effects are all equal. Also compute pre-trends estimates $\widehat{\text{AVSQ}}_{-\ell}$ for $\ell \in \{1, \dots, 4\}$. Interpret the results.

```
did_multiplegt_dyn prestout cnty90 year numdailies, effects(4)
placebo(4) effects_equal(all)
```

Non-normalized event-study and pre-trends estimates are shown in Figure 8.1 below. Being exposed to a weakly larger number of newspapers for one electoral cycle increases turnout by 1.44 percentage point, and the effect is statistically significant (s.e.=0.43 percentage point). That effect can be estimated for 1,119 out of the 1,195 counties in the data: 34 counties never experience a change in their number of newspapers, and 42 counties that do experience a change cannot be matched with a not-yet-switcher with the same number of newspapers at baseline. Being exposed to a weakly larger number of newspapers for two, three, and four electoral cycle also significantly increases turnout. Effects increase with exposure length, but one cannot reject the null that all effects are equal ($p\text{-value}=0.40$). As ℓ increases, effects mechanically apply to fewer and fewer counties, but the effect after four electoral cycles still applies to 917 counties. Pre-

trend estimates are small and individually and jointly insignificant. However, their confidence intervals are quite large. While the first pre-trend estimator applies to 906 of the 1,119 counties for which \widehat{AVSQ}_1 is estimated, the fourth pre-trend estimator only applies to 447 of the 917 counties for which \widehat{AVSQ}_4 is estimated. The confidence interval of \widehat{AVSQ}_{-4} is already quite large, but that of \widehat{AVSQ}_{-5} is substantially larger, so we have very little power to detect differential trends over more than five election cycles. This is why we only report four placebo and four event-study estimators.

Figure 8.1: Effects of being exposed to a weakly larger number of newspapers for ℓ periods on turnout



Note: This figure shows non-normalized DID estimates of the effect of being exposed to a weakly larger number of newspapers for ℓ periods on turnout, as well as pre-trends estimates, computed using the data of Gentzkow et al. (2011) and the `did_multiplegt_dyn` Stata command. Standard errors are clustered at the county level. 95% confidence intervals are shown in red.

Rerun the previous command, estimating one event-study effect and adding `design(0.8, console)`

at the end. What are the three most common “actual-versus-status-quo” comparisons averaged in $\widehat{\text{AVSQ}}_1$, the effect estimated by $\widehat{\text{AVSQ}}_1$? Rerun the previous command, estimating two event-study effects. What are the three most common “actual-versus-status-quo” comparisons averaged in $\widehat{\text{AVSQ}}_2$, the effect estimated by $\widehat{\text{AVSQ}}_2$? Rerun the previous command, estimating four event-study effects. What are the three most common “actual-versus-status-quo” comparisons averaged in $\widehat{\text{AVSQ}}_4$, the effect estimated by $\widehat{\text{AVSQ}}_4$?

```
did_multiplegt_dyn prestout cnty90 year numdailies, effects(1) design(0.8,console)
graph_off
did_multiplegt_dyn prestout cnty90 year numdailies, effects(2) design(0.8,console)
graph_off
did_multiplegt_dyn prestout cnty90 year numdailies, effects(4) design(0.8,console)
graph_off
```

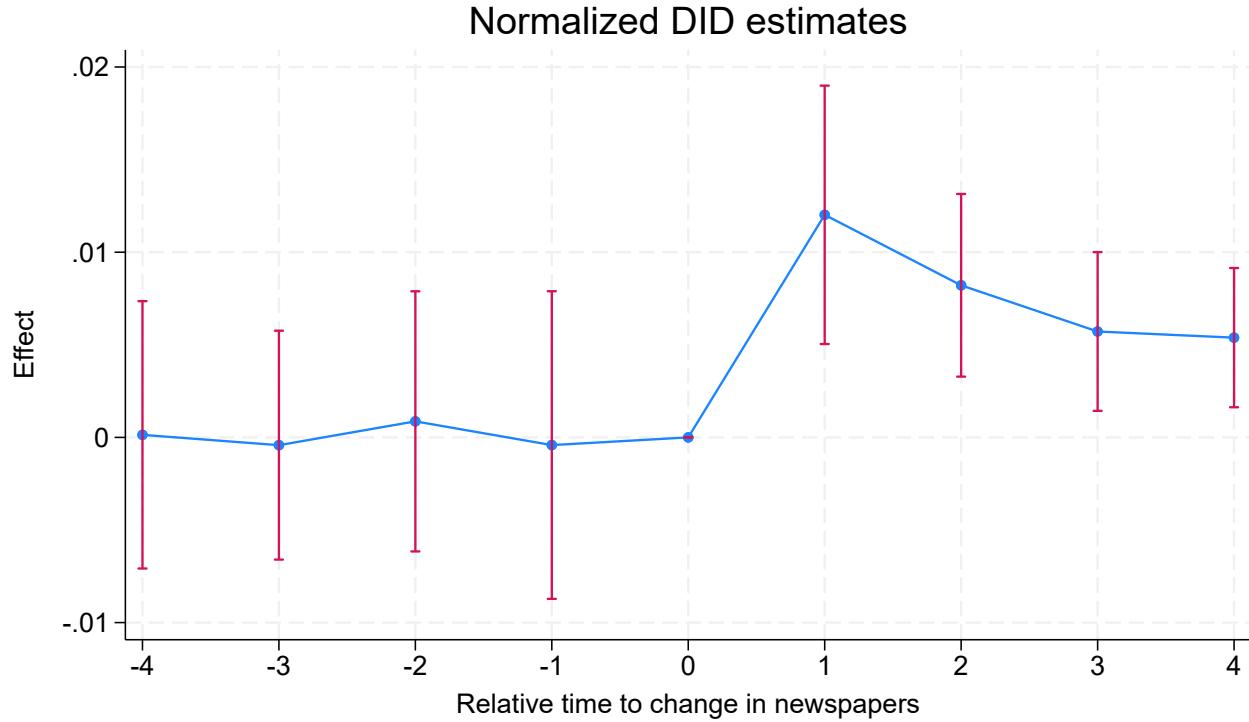
64% of the $\text{AVSQ}_{g,1}$ effects averaged in AVSQ_1 are effects of having one versus zero newspapers, 12% are effects of having two versus zero newspapers, and 5% are effects of having two versus one newspapers. 32% of the $\text{AVSQ}_{g,2}$ effects averaged in AVSQ_2 are effects of having $D_{g,F_g} = 1, D_{g,F_g+1} = 1$ instead of $D_{g,F_g} = 0, D_{g,F_g+1} = 0$, 18% are effects of having $D_{g,F_g} = 1, D_{g,F_g+1} = 0$ instead of $D_{g,F_g} = 0, D_{g,F_g+1} = 0$, and 12% are effects of having $D_{g,F_g} = 1, D_{g,F_g+1} = 2$ instead of $D_{g,F_g} = 0, D_{g,F_g+1} = 0$. Finally, 15% of the $\text{AVSQ}_{g,4}$ effects averaged in AVSQ_4 are effects of having $D_{g,F_g} = 1, D_{g,F_g+1} = 1, D_{g,F_g+2} = 1, D_{g,F_g+3} = 1$ instead of $D_{g,F_g} = 0, D_{g,F_g+1} = 0, D_{g,F_g+2} = 0, D_{g,F_g+3} = 0$, 14% are effects of having $D_{g,F_g} = 1, D_{g,F_g+1} = 0, D_{g,F_g+2} = 0, D_{g,F_g+3} = 0$ instead of $D_{g,F_g} = 0, D_{g,F_g+1} = 0, D_{g,F_g+2} = 0, D_{g,F_g+3} = 0$, and 5% are effects of having $D_{g,F_g} = 1, D_{g,F_g+1} = 2, D_{g,F_g+2} = 2, D_{g,F_g+3} = 2$ instead of $D_{g,F_g} = 0, D_{g,F_g+1} = 0, D_{g,F_g+2} = 0, D_{g,F_g+3} = 0$. As ℓ increases, AVSQ_ℓ averages the effects of more and more heterogeneous paths across groups, and the three most common paths account for a smaller fraction of all the $\text{AVSQ}_{g,\ell}$ effects averaged in AVSQ_ℓ . For most paths, too few groups have that path to estimate reasonably precisely a path-specific effect.

Using `gentzkowetal_didtextbook` and the `did_multiplegt_dyn` command, compute, for $\ell \in \{1, \dots, 4\}$, normalized event-study estimates $\widehat{\text{AVSQ}}_{\ell}^n$, as well as the weights $w_{\ell,k}$ that normalized effect ℓ puts on the effect of the k th treatment lag for $k \in \{0, \dots, \ell-1\}$. Compute also normalized pre-trends estimates $\widehat{\text{AVSQ}}_{-\ell}^n$ for $\ell \in \{1, \dots, 4\}$. Does it seem that the effect of lagged newspapers on turnout is larger or smaller than the effect of contemporaneous newspapers?

```
did_multiplegt_dyn prestout cnty90 year numdailies, effects(4) placebo(4) normalized
effects_equal(all)
```

Normalized event-study and pre-trends estimates are shown in Figure 8.2 below. Normalized event-study estimates are decreasing with ℓ , but one cannot reject the null that all effects are equal (p-value=0.17). $w_{1,0} = 1$: the first event-study estimate is an effect of contemporaneous newspapers on turnout. $w_{2,0} = 0.48$ and $w_{2,1} = 0.52$: the second normalized event-study estimate is a weighted average of the effects of contemporaneous newspapers and of the first lag of newspapers on turnout, with approximately equal weights. $w_{3,0} = 0.35$, $w_{3,1} = 0.31$, and $w_{3,2} = 0.33$: the third normalized event-study estimate is a weighted average of the effects of contemporaneous newspapers and of the first and second lag of newspapers, with approximately equal weights. Finally, $w_{4,0} = 0.28$, $w_{4,1} = 0.26$, $w_{4,2} = 0.23$, and $w_{4,3} = 0.24$: the fourth normalized event-study estimate is a weighted average of the effects of contemporaneous newspapers and of the first, second, and third lag of newspapers, again with approximately equal weights. Then, the fact that normalized event-study estimates are decreasing with ℓ may suggest that lagged newspapers have a smaller effect on turnout than contemporaneous newspapers.

Figure 8.2: Normalized DID estimates of effects of newspapers on turnout



Note: This figure shows normalized DID estimates of the effect of newspapers on turnout, as well as normalized pre-trends estimates, computed using the data of Gentzkow et al. (2011) and the `did_multiplegt_dyn` Stata command. Standard errors are clustered at the county level. 95% confidence intervals are shown in red.

Can you jointly test the null that the first lagged treatment has no effect on the outcome, and that treatment effects are constant over time?

As explained in Section 8.3.4.3, this test can be performed in a subsample of groups whose treatment does not change between F_g and $F_g + 1$, and for which both the first and second non-normalized event-study effects can be estimated. Then, the test amounts to assessing if the first and second non-normalized event-study effects are equal in that subsample. We implement

this test using the `did_multiplegt_dyn` command, restricting attention to (g, t) cells such that $t < F_g$ or $D_{g,F_g} = D_{g,F_g+1}$:

```
did_multiplegt_dyn prestout cnty90 year numdailies
if year<=first_change|same_treat_after_first_change==1,
effects(2) effects_equal(all) same_switchers graph_off
```

In that subsample of 512 counties, the estimates of the first and second non-normalized event-study effects are close and not significantly different ($p\text{-value}=0.83$). This suggests that the first lag of newspapers does not affect turnout.

8.4 Heterogeneity-robust estimators, ruling out dynamic effects

Allowing for dynamic effects is appealing, but then the previous section has shown that under a placebo-testable parallel-trends condition, one can only estimate event-study effects that average together effects of many different treatment paths, and may therefore be hard to interpret. Therefore, if the test proposed in Section 8.3.4.3 suggests that lagged treatments actually do not affect the outcome, one may consider assuming away dynamic effects. Then, one can estimate effects that are easier to interpret. Moreover, those effects can be estimated under a minimal parallel-trends assumption over consecutive periods rather than across multiple periods. Finally, with a non-absorbing treatment, allowing for dynamic effects makes it hard to separately estimate the effect of groups' first, second, etc. treatment changes. Instead, ruling out dynamic effects makes it possible to separately estimate the effect of each treatment change. This makes it easy to test for instance if the treatment effect changes over time, and this can also lead to more precise estimators. Therefore, throughout this section we impose Assumption ND. We start by describing the estimators proposed by de Chaisemartin and D'Haultfœuille (2020) and de Chaisemartin et al. (2022), before describing alternative estimators.

8.4.1 Design

In this section, we assume that $D_{g,t}$ takes values in $\mathcal{D} = \{0, \dots, \bar{d}\}$. If $\bar{d} = 1$, $D_{g,t}$ is binary, but we allow for a non-binary discrete treatment taking a finite number of values, as is for instance the case in the newspaper example. The estimators below can be computed whenever the following condition holds:

Design STAY-C (*Designs with stayers between consecutive periods*) $\exists(g, g')$ such that: (i) $D_{g,t-1} = D_{g',t-1}$; (ii) $D_{g,t} \neq D_{g',t-1}$ while $D_{g',t} = D_{g',t-1}$.

Design STAY-C requires that there exists a pair of consecutive time periods $(t-1, t)$ and a pair of groups (g, g') such that g and g' have the same treatment at $t-1$, g 's treatment changes from $t-1$ to t while g' 's treatment does not change. Hereafter, we refer to groups whose treatment changes from $t-1$ to t as $t-1$ -to- t switchers, while we refer to groups whose treatment does not change from $t-1$ to t as $t-1$ -to- t stayers. When $T = 2$, Design STAY-C is equivalent to Design STAY, the design in which the estimators in the previous section can be used. When $T > 2$, the two designs are no longer equivalent. Without dynamic effects, each pair of consecutive periods can be analyzed in isolation, because groups' outcomes at $t-1$ and t do not depend on their treatments at other periods. Then, Design STAY-C requires that one can match a $t-1$ -to- t switcher to a $t-1$ -to- t stayer with the same period- $t-1$ treatment, without taking into account their prior treatment histories. Instead, Design STAY requires that one can match a switcher to a not-yet-switcher with the same treatment history until the switcher switched. As in the previous section, (i) fails if the treatment is continuously distributed, but in Section 8.4.8.1 we will extend the estimators below to designs with a continuous treatment.

8.4.1.1 Parallel-trends assumption

$\forall t \in \{2, \dots, T\}$, let $\mathcal{D}_t^r = \{d : \exists(g, g') : D_{g,t-1} = D_{g',t-1} = D_{g',t} = d \neq D_{g,t}\}$ be the set of values of the lagged treatment $D_{g,t-1}$ such that at least one $t-1$ -to- t switcher and one $t-1$ -to- t stayer have $D_{g,t-1} = d$.

Assumption PTNC-C (*Parallel trends if groups' treatment does not change between consecutive periods, conditional on their lagged treatment*) $\forall t \in \{2, \dots, T\}, \forall(g, g')$, if $D_{g,t-1} = D_{g',t-1} \in$

\mathcal{D}_t^r , then

$$E[Y_{g,t}(D_{g,t-1}) - Y_{g,t-1}(D_{g,t-1})] = E[Y_{g',t}(D_{g,t-1}) - Y_{g',t-1}(D_{g,t-1})].$$

Interpret Assumption PTNC-C.

Assumption PTNC-C requires that if two groups have the same lagged treatment, then they have the same expected outcome evolution from period $t - 1$ to t , in the counterfactual where their treatment does not change from $t - 1$ to t . Importantly, Assumption PTNC-C only requires that some groups be on parallel trends over consecutive time periods, not over the entire duration of the panel. Specifically, because Assumption PTNC-C is conditional on $D_{g,t-1}$, it cannot be “chained” across pairs of time periods: for instance, under Assumption PTNC-C, two groups g and g' such that $D_{g,1} = 2, D_{g,2} = D_{g,3} = 3$ and $D_{g',1} = D_{g',2} = D_{g',3} = 2$ experience parallel trends from period one to two but not from period two to three, because they have the same treatment at period one but not at period two. As it restricts only one potential outcome per group, Assumption PTNC-C alone does not restrict groups’ treatment effects. When combined with (2.1), the standard parallel-trends assumption on the untreated outcome, does Assumption PTNC-C imply that the treatment effect should be constant over time?

No, but together the two conditions imply a parallel-trends condition on groups’ treatment effects:

$$E[Y_{g,t}(d) - Y_{g,t-1}(d)] = E[Y_{g',t}(d) - Y_{g',t-1}(d)]$$

and

$$E[Y_{g,t}(0) - Y_{g,t-1}(0)] = E[Y_{g',t}(0) - Y_{g',t-1}(0)],$$

for $d \in \mathcal{D}_t^r$ and (g, g') such that $D_{g,t-1} = D_{g',t-1} = d$. These two equalities imply that

$$E[Y_{g,t}(d) - Y_{g,t}(0) - (Y_{g,t-1}(d) - Y_{g,t-1}(0))] = E[Y_{g',t}(d) - Y_{g',t}(0) - (Y_{g',t-1}(d) - Y_{g',t-1}(0))].$$

Therefore, for all $d \in \mathcal{D}_t^r$, the effect of changing the treatment from 0 to d should follow the same evolution from $t - 1$ to t for all groups g such that $D_{g,t-1} = d$. On the other hand, if one were to assume that Assumption PTNC-C holds for all groups, as in (8.9), then together with (2.1) the two conditions would imply that some treatment effects are constant over time. Thus, as in the previous section, making the parallel-trends assumption conditional on groups' prior treatment substantially weakens the restrictions on treatment-effect heterogeneity implicitly imposed by those parallel-trends assumptions.

8.4.2 Target parameters

Let $\mathcal{S} = \{(g, t) : t \geq 2, D_{g,t} \neq D_{g,t-1}, \exists g' : D_{g',t-1} = D_{g',t} = D_{g,t-1}\}$ denote the set of $t - 1$ -to- t switchers with the same lagged treatment as at least one $t - 1$ -to- t stayer. For all $(g, t) \in \mathcal{S}$, let

$$\text{TE}_{g,t}^\Delta = \frac{E[Y_{g,t}(D_{g,t}) - Y_{g,t}(D_{g,t-1})]}{D_{g,t} - D_{g,t-1}}$$

denote the expectation of the slope of group g 's potential outcome function at period t , between its period- t and its period- $t - 1$ treatment. Recall that for any set A , $\#A$ denotes its number of elements, i.e. its cardinality. The first target parameter we consider is

$$\text{ATS} = E\left(\frac{1}{\#\mathcal{S}} \sum_{(g,t) \in \mathcal{S}} \text{TE}_{g,t}^\Delta\right).$$

ATS (Average Treatment effect of the Switchers) is the average of the slopes $\text{TE}_{g,t}^\Delta$ across all switchers in \mathcal{S} . In Design HAD and with i.i.d. groups, the ATS reduces to the ATT we considered in Chapter 7. The second target parameter we consider is a weighted average of switchers' slopes $\text{TE}_{g,t}^\Delta$:

$$\text{WATS} = E\left(\sum_{(g,t) \in \mathcal{S}} \frac{|D_{g,t} - D_{g,t-1}|}{\sum_{(g',t') \in \mathcal{S}} |D_{g',t'} - D_{g',t'-1}|} \text{TE}_{g,t}^\Delta\right).$$

In Design HAD and with i.i.d. groups, the WATS reduces to the WATT we considered in Chapter 7. If $D_{g,t}$ is binary $\text{ATS} = \text{WATS}$, but the two parameters can differ if $D_{g,t}$ is non-binary. One may also be interested in estimating separately the average treatment effect of switchers-in whose treatment increases ($D_{g,t} > D_{g,t-1}$), and of switchers-out whose treatment decreases. Accordingly, one can let $\mathcal{S}_+ = \{(g, t) : t \geq 2, D_{g,t} > D_{g,t-1}, \exists g' : D_{g',t-1} = D_{g',t} = D_{g,t-1}\}$,

$\mathcal{S}_- = \{(g, t) : t \geq 2, D_{g,t} < D_{g,t-1}, \exists g' : D_{g',t-1} = D_{g',t} = D_{g,t-1}\}$, and

$$\begin{aligned} \text{ATS}_+ &= E \left(\frac{1}{\#\mathcal{S}_+} \sum_{(g,t) \in \mathcal{S}_+} \text{TE}_{g,t}^\Delta \right) \\ \text{ATS}_- &= E \left(\frac{1}{\#\mathcal{S}_-} \sum_{(g,t) \in \mathcal{S}_-} \text{TE}_{g,t}^\Delta \right), \end{aligned}$$

and one can define WATS₊ and WATS₋ similarly.

8.4.3 Estimators

For all $(g, t) \in \mathcal{S}$, let

$$\widehat{\text{TE}}_{g,t}^\Delta = \frac{Y_{g,t} - Y_{g,t-1} - \frac{1}{\#\{g': D_{g',t-1} = D_{g',t} = D_{g,t-1}\}} \sum_{g': D_{g',t-1} = D_{g',t} = D_{g,t-1}} (Y_{g',t} - Y_{g',t-1})}{D_{g,t} - D_{g,t-1}}.$$

The numerator of $\widehat{\text{TE}}_{g,t}^\Delta$ is a DID estimator comparing the $t-1$ -to- t outcome evolution of switcher g to the average $t-1$ -to- t outcome evolutions of stayers with the same lagged treatment as g .

Theorem 23 *If Assumptions NA, ND, and PTNC-C hold, then for all $(g, t) \in \mathcal{S}$,*

$$E[\widehat{\text{TE}}_{g,t}^\Delta] = \text{TE}_{g,t}^\Delta.$$

Then, it directly follows from Theorem 23 that

$$\widehat{\text{ATS}} := \frac{1}{\#\mathcal{S}} \sum_{(g,t) \in \mathcal{S}} \widehat{\text{TE}}_{g,t}^\Delta$$

is unbiased for the ATS, while

$$\widehat{\text{WATS}} := \sum_{(g,t) \in \mathcal{S}} \frac{|D_{g,t} - D_{g,t-1}|}{\sum_{(g',t') \in \mathcal{S}} |D_{g',t'} - D_{g',t'-1}|} \widehat{\text{TE}}_{g,t}^\Delta$$

is unbiased for the WATS. One can follow similar steps to construct unbiased estimators of ATS_+ , ATS_- , WATS_+ , and WATS_- , respectively denoted $\widehat{\text{ATS}}_+$, $\widehat{\text{ATS}}_-$, $\widehat{\text{WATS}}_+$, and $\widehat{\text{WATS}}_-$.

Bibliographic notes. With a binary treatment, the multi-period DID estimator in Imai and Kim (2021) is numerically equivalent to $\widehat{\text{ATS}}_+$. $\widehat{\text{WATS}}$ is numerically equivalent to the DID_M estimator in de Chaisemartin and D'Haultfœuille (2020). With a discrete treatment, the AS estimator in de Chaisemartin et al. (2022) is numerically equivalent to $\widehat{\text{ATS}}$.

8.4.4 Pre-trend estimators

Let $\mathcal{S}^{\text{pl}} = \{(g, t) \in \mathcal{S} : D_{g,t-1} = D_{g,t-2}\}$ denote the subsample of $t - 1$ -to- t switchers that are also $t - 2$ -to- $t - 1$ stayers. For all $(g, t) \in \mathcal{S}^{\text{pl}}$, let

$$\widehat{\text{TE}}_{g,t}^{\Delta,\text{pl}} = \frac{Y_{g,t-1} - Y_{g,t-2} - \frac{1}{\#\{g': D_{g',t-2} = D_{g',t-1} = D_{g',t} = D_{g,t-1}\}} \sum_{g': D_{g',t-2} = D_{g',t-1} = D_{g',t} = D_{g,t-1}} (Y_{g',t-1} - Y_{g',t-2})}{D_{g,t} - D_{g,t-1}},$$

and let

$$\widehat{\text{ATS}}^{\text{pl}} = \frac{1}{\#\mathcal{S}^{\text{pl}}} \sum_{(g,t) \in \mathcal{S}^{\text{pl}}} \widehat{\text{TE}}_{g,t}^{\Delta,\text{pl}}.$$

$\widehat{\text{ATS}}^{\text{pl}}$ compares the $t - 2$ -to- $t - 1$ outcome evolutions of $t - 1$ -to- t switchers and stayers, restricting attention to $t - 2$ -to- $t - 1$ stayers. One can follow similar steps to define a placebo WATS estimator. If $\widehat{\text{ATS}}^{\text{pl}}$ is small when compared to $\widehat{\text{ATS}}$, differential trends between switchers and stayers are larger after switchers' treatment changes than before that change, which suggests that $\widehat{\text{ATS}}$ is unlikely to be strongly biased due to a violation of Assumption PTNC-C. [Why is it important to restrict the computation of the placebo estimator to \$t - 2\$ -to- \$t - 1\$ stayers?](#)

Otherwise, the placebo estimator could be contaminated by the treatment's effect: its expectation could differ from zero even if switchers and stayers are on parallel trends, if $t - 1$ -to- t switchers and stayers have different probabilities of being $t - 2$ -to- $t - 1$ switchers and/or different treatment effects.

8.4.5 Estimators robust to dynamic effects up to a pre-specified number of lags

$\widehat{\text{ATS}}$ and $\widehat{\text{WATS}}$ compare the outcome evolutions of $t - 1$ -to- t switchers and stayers. But there may be, say, $t - 1$ -to- t stayers whose treatment changed from $t - 2$ to $t - 1$. If lagged treatments affect units' current outcome, that change could still affect the $t - 1$ -to- t outcome evolution of those stayers. This could lead to a violation of the parallel-trends assumption underlying $\widehat{\text{ATS}}$ and $\widehat{\text{WATS}}$. To mitigate this concern, de Chaisemartin et al. (2022) propose the following

robustness check. One can recompute $\widehat{\text{ATS}}$ and $\widehat{\text{WATS}}$, restricting, for each pair of consecutive periods $(t - 1, t)$, the estimation sample to $t - 2$ -to- $t - 1$ stayers, as in the placebo analysis. They show that the resulting estimator is robust to dynamic effects up to one treatment lag. Similarly, if one wants to allow for effects of the first and second treatment lags on the outcome, one just needs to restrict the estimation sample to $t - 3$ -to- $t - 1$ stayers. However, the more robustness to dynamic effects one would like to have, the smaller the estimation sample becomes.

8.4.6 Inference

de Chaisemartin and D'Haultfoeuille (2020) propose confidence intervals for the WATS based on asymptotic approximations where the number of groups goes to infinity, under the assumption that groups are independent. We recommend, as in Section 3.3.2, that researchers using those confidence intervals with less than 40 switchers or with less than 40 stayers perform simulations tailored to their data to assess their coverage rate.⁶

8.4.7 Imputation estimator

Borusyak et al. (2024) and Liu et al. (2024) show that with a binary, non-absorbing treatment, their imputation estimator can still be used if one is ready to rule out dynamic effects. Then, under (2.1), the imputation estimator is unbiased for the average effect of the treatment across all treated (g, t) cells such that the fixed effect of group g and the fixed effect of period t can be estimated in the first-step TWFE regression of the outcome on group and time FEs in the sample of untreated (g, t) cells. $\widehat{\text{ATS}}$ and the imputation estimator are unbiased for average treatment effects across two different sets of (g, t) cells. This could lead the two estimators to differ if treatment effects are heterogeneous across those two sets. The two estimators also rely on different parallel-trends assumptions. In particular, the imputation estimator assumes that all groups experience the same evolution of their untreated outcome over the entire duration of

⁶The number of switchers is the number of groups g for which at least one cell (g, t) belongs to \mathcal{S} . The number of stayers is the number of groups g such that g belongs to the set of stayers attached to at least one cell (g', t) in \mathcal{S} .

the panel. Instead, with a binary treatment $\widehat{\text{ATS}}_+$ requires that untreated groups at period $t - 1$ experience the same evolution of their untreated outcome from $t - 1$ to t , a weaker assumption. The parallel-trends assumption underlying $\widehat{\text{ATS}}_-$ is neither weaker nor stronger than that underlying the imputation estimator, but again it only requires that some groups are on parallel trends over consecutive time periods.

8.4.8 Extensions

To simplify the exposition, in this section we assume that $T = 2$: if $T > 2$, one can just compute the estimators below for all pairs of consecutive time periods, and then take a weighted average across pairs of periods.

8.4.8.1 Continuous treatment

de Chaisemartin et al. (2022) extend the estimators in de Chaisemartin and D'Haultfœuille (2020) to designs with a continuously distributed treatment. With a discrete treatment, the estimators of de Chaisemartin and D'Haultfœuille (2020) compare switchers and stayers with the exact same baseline treatment. The challenge with a continuous treatment is that the sample does not contain switchers and stayers with the same period-one treatment. Then, to estimate $E[Y_{g,2}(D_{g,1}) - Y_{g,1}(D_{g,1})]$, a switcher's counterfactual outcome evolution if its treatment had not changed, we cannot just use the average outcome evolution of stayers with the same $D_{g,1}$. Instead, drawing some inspiration from the DID estimators with covariates we discussed in Chapter 4, propose another method to estimate $E[Y_{g,2}(D_{g,1}) - Y_{g,1}(D_{g,1})]$.

Intuitively, one can estimate non-parametrically $E[Y_{g,2}(D_{g,1}) - Y_{g,1}(D_{g,1})]$, and therefore the ATS and WATS, with a procedure similar to that described in Chapter 4 to control for covariates in DID estimation, with $D_{g,1}$ playing the role of the covariate. First, one estimates a non-parametric (e.g. series, kernel, lasso) regression of $Y_{g,2} - Y_{g,1}$ on $D_{g,1}$ among stayers. Second, one

computes switchers' predicted outcome evolution if their treatment had not changed, based on that regression. Finally, we subtract from switchers' outcome evolution their predicted outcome evolution without treatment, to recover their treatment effect, and we average across switchers. de Chaisemartin et al. (2022) derive doubly-robust moment conditions identifying the ATS and WATS, and they propose non-parametric doubly-robust estimators, with data-driven choices of the tuning parameters used in the first-step estimations. They show that the ATS estimator is \sqrt{G} -consistent if switchers cannot experience arbitrarily small treatment changes, while the WATS estimator is always \sqrt{G} -consistent. They also show that when switchers cannot experience arbitrarily small treatment changes, under some conditions the asymptotic variance of the WATS estimator is strictly lower than that of the ATS estimator.

8.4.8.2 Estimators with control variables

de Chaisemartin et al. (2022) also propose estimators relying on a conditional parallel-trends assumption, which extend the DID estimators with covariates reviewed in Section 4.1 to non-binary and/or non-staggered designs.

8.4.8.3 Estimating heterogeneous treatment effects

If one is interested in assessing heterogeneous treatment effects across covariates X_g taking a small number of values, one can just recompute the estimators above in subsamples of groups with the same value of X_g . de Chaisemartin et al. (2022) and de Chaisemartin and D'Haultfœuille (2020) do not propose methods to investigate heterogeneous effects along covariates taking a large number of values. With a binary treatment, such heterogeneity analyses can be conducted using the imputation estimator of Borusyak et al. (2024).

*8.4.8.4 Estimators with several treatments**

de Chaisemartin and D'Haultfœuille (2023a) propose an estimator generalizing that in de Chaisemartin and D'Haultfœuille (2020) to instances with several treatments. To isolate the effect of the first treatment, their estimator compares the $t-1$ -to- t outcome evolution, of switching groups (i) whose first treatment switches from $t-1$ to t while (ii) their other treatments do not change, to control groups (i') whose first treatment does not change, (ii') whose other treatments also do

not change, and (iii') that had the same treatments as the switching groups in period $t - 1$. (ii) and (ii') ensure that this estimator will not be subject to an issue affecting TWFE regressions with several treatments, which issue is that?

(ii) and (ii') ensure that the estimator of the effect of the first treatment is not contaminated by effects of other treatments, unlike the coefficient on the first treatment in a TWFE regressions with several treatments. Interestingly, this idea was present as early as in Snow (1856): to assess if cholera is transmitted by air or water, Snow found a treatment group whose water quality changed while its air quality did not change, and a control group whose water and air quality did not change. (i') ensures that the estimator is robust to heterogeneous effects of the first treatment across groups. Finally, (iii') ensures that the estimator is robust to heterogeneous effects over time of all treatments.

8.4.8.5 Instrumental-variable DID estimators*

de Chaisemartin et al. (2022) extend their estimators to the IV case, where one makes a parallel-trends assumption with respect to an instrument rather than the treatment. As discussed earlier, 2SLS-TWFE estimators may not estimate a LATE or a convex combination of treatment effects. Instead, de Chaisemartin et al. (2022) propose an IV-WATS estimator, defined as the ratio of a reduced-form WATS estimator with $Y_{g,t}$ as the outcome and the instrument $Z_{g,t}$ as the treatment, divided by a first-stage WATS estimator with $D_{g,t}$ as the outcome and $Z_{g,t}$ as the treatment. They show that under a monotonicity condition as in Imbens and Angrist (1994), and under parallel-trends conditions on $E(Y_{g,2}(D_{g,2}(Z_{g,1})) - Y_{g,1}(D_{g,1}(Z_{g,1})))$ and $E(D_{g,2}(Z_{g,1}) - D_{g,1}(Z_{g,1}))$, groups' outcome and treatment evolutions if their instrument had not changed, the IV-WATS estimator is consistent for a so-called IV-WATS effect. The IV-WATS effect is a weighted average of the slopes

$$\frac{Y_{g,2}(D_{g,2}(Z_{g,2})) - Y_{g,2}(D_{g,2}(Z_{g,1}))}{D_{g,2}(Z_{g,2}) - D_{g,2}(Z_{g,1})},$$

across “complier-switchers” such that $D_{g,2}(Z_{g,2}) \neq D_{g,2}(Z_{g,1})$. Those are the groups whose

instrument switches from period one to two (since $D_{g,2}(Z_{g,2}) \neq D_{g,2}(Z_{g,1})$ implies $Z_{g,2} \neq Z_{g,1}$), and whose treatment responds to that change, like the compliers in Imbens and Angrist (1994). The IV-WATS weights slopes proportionally to $D_{g,2}(Z_{g,2}) - D_{g,2}(Z_{g,1})$, compliers-switchers' first-stage effect. Similarly, a reduced-form ATS estimator divided by a first-stage ATS estimator is consistent for a weighted average of the same slopes, with weights proportional to $(D_{g,2}(Z_{g,2}) - D_{g,2}(Z_{g,1})) / (Z_{g,2} - Z_{g,1})$, the slope of compliers-switchers' first stage. Finally, de Chaisemartin et al. (2022) show that the “reduced-form” parallel-trends condition on $E(Y_{g,2}(D_{g,2}(Z_{g,1})) - Y_{g,1}(D_{g,1}(Z_{g,1})))$ restricts treatment effect heterogeneity over time and across units. Instead, a reduced-form parallel-trends condition conditional on $(Z_{g,1}, D_{g,1})$ no longer restricts treatment-effect heterogeneity over time, though it still restricts it across units. Accordingly, de Chaisemartin et al. (2022) recommend controlling for $(Z_{g,1}, D_{g,1})$ in the IV estimation.

8.4.9 Computation: Stata and R commands computing heterogeneity-robust estimators assuming away dynamic effects

Estimators of de Chaisemartin and D’Haultfœuille (2020) and de Chaisemartin et al. (2022). The `did_multiplegt_stat` Stata command (see de Chaisemartin, Ciccia, D’Haultfœuille, Knau and Sow, 2024c) can be used to compute those estimators. With a binary or discrete treatment, its syntax is:

```
did_multiplegt_stat outcome groupid timeid treatment, exact_match
```

That command can also compute the IV-WATS estimator. Then, with a binary or discrete instrument the syntax is:

```
did_multiplegt_stat outcome groupid timeid treatment instrument, estimator(iv-was)
exact_match
```

With a continuous treatment or instrument, one needs to drop the `exact_match` option. The command can also be used to compute pre-trends estimators, by specifying the `placebo` option. In R, the `didmultiplegt` (see Zhang and de Chaisemartin, 2020) command can be used to compute the WATS estimator with a binary or discrete treatment.

Imputation estimators. With a binary non-absorbing treatment, the `did_imputation` Stata command and the `didimputation` R command (see Butts, 2021b) compute the imputation estimators proposed by Borusyak et al. (2024). You can refer to Chapter 6 for the syntax of the Stata command. The `fetc` Stata (Liu et al., 2022b) and R (Liu et al., 2022a) command compute the estimators proposed by Liu et al. (2024).

8.4.10 Application to the newspaper example

Using `gentzkowetal_didtextbook` and the `did_multiplegt_stat` command, compute \widehat{ATS} , \widehat{WATS} , \widehat{ATS}^{pl} , and \widehat{WATS}^{pl} . Interpret the results. In particular, how do \widehat{ATS} and \widehat{WATS} compare to \widehat{AVSQ}_1^n ?

```
did_multiplegt_stat prestout cnty90 year numdailies, placebo(1) exact_match
 $\widehat{ATS} = 0.0061$ : across the 4,423 switching county  $\times$  year cells in  $\mathcal{S}$ , increasing the contemporaneous number of newspapers by one increases turnout by 0.61 percentage points on average, a significant effect at all conventional levels (s.e.=0.0016).  $\widehat{WATS} = 0.0058$  (s.e.=0.0015):7 the ATS and WATS estimators are close in this application. The  $\widehat{ATS}$  and  $\widehat{WATS}$  are more than twice smaller than  $\widehat{AVSQ}_1^n$ , which also estimates an average effect of increasing the contemporaneous number of newspapers by one on turnout. However,  $\widehat{AVSQ}_1^n$  estimates the average effect across first-time switchers, while  $\widehat{ATS}$  and  $\widehat{WATS}$  estimate the average effect across all switchers. 76.5% of first-time switchers had no newspapers at  $F_g - 1$ . Thus,  $\widehat{AVSQ}_1^n$  is, for the most part, averaging slopes between 0 and a strictly positive number of newspapers. Instead, only 23.2% of switchers had no newspapers at  $t - 1$ . Then, the fact that  $\widehat{AVSQ}_1^n$  is much larger than  $\widehat{ATS}$  and  $\widehat{WATS}$  could suggest that newspapers have a non-linear effect on turnout, and that going from
```

⁷ \widehat{WATS} slightly differs from the DID_M estimator in Table 3 of de Chaisemartin and D'Haultfœuille (2020), because it does not control for state-specific trends, and it does not group number of newspapers above three into one category.

zero to one newspaper has a larger effect than going from one to two, etc. Finally, \widehat{ATS}^{pl} is small and insignificant (-0.0011, s.e.=0.0025), though its confidence interval is quite wide. \widehat{WAT}^{pl} is also small and insignificant (-0.0000, s.e.=0.0023).

8.5 Heterogeneity-robust estimators in designs without stayers?

The heterogeneity-robust DID estimators reviewed earlier can be computed in designs with stayers. Such designs seem quite common. Of the 26 highly-cited AER papers estimating a TWFE regression in the survey of de Chaisemartin and D'Haultfœuille (2025), there are 12 for which we can replicate at least one TWFE regression estimated in the paper, without having to preprocess the publicly available data using a software we are not familiar with (e.g. ArcGIS). For papers for which we can replicate several TWFE regressions, we focus on the one reported first in the paper. For each of these 12 regressions, we assess if it has stayers: $\exists(g, t) : D_{g,t} = D_{g,t-1}$, or $\exists(g, t) : Z_{g,t} = Z_{g,t-1}$ for 2SLS regressions. For regressions with several treatments or instruments, we assess if at least one treatment or instrument has stayers. We find that 9 regressions have stayers. Of the three papers that do not have stayers for any treatment or instrument in their main regression, one is Pierce and Schott (2016), an heterogeneous adoption design with quasi-stayers. Another one is Fetzer (2019), who studies the effect of austerity in the UK on the propensity to vote for Brexit. While the main austerity measure in the paper's Table 1 Column (1) does not have stayers, other austerity measures in that same table have stayers. Overall, and though our sample is admittedly small, this suggests that we have now been able to propose heterogeneity-robust DID estimators that can be used in a majority of the cases where TWFE regressions are used. Yet, designs without stayers exist. They are for instance prominent in a very important field of research, which measures the impact of weather variables on economic or health outcomes, such as agricultural yields (see, e.g., Deschênes and Greenstone, 2007) or mortality. If $D_{g,t}$ is the amount of rainfall or the average temperature in location g and year t , all locations will experience different precipitations or temperatures in consecutive years: such treatments are continuously distributed across both g and t . We now review several alternatives to TWFEs in such cases. Throughout, we assume away dynamic

effects to simplify the exposition, though some of the estimators below can be extended to allow for dynamic effects. We also assume that $T = 2$, occasionally assuming the existence of a third period $t = 0$ when we discuss pre-trend estimators.

8.5.1 Using quasi-stayers

When there are quasi-stayers, namely groups that experience arbitrarily small treatment changes, a first solution is to use them as the control group. The researcher chooses a value h such that groups for which $|D_{g,2} - D_{g,1}| \leq h$ are considered as quasi-stayers. Then, the heterogeneity-robust DID estimators presented in the previous section can be computed as if the treatment of those cells had not changed, using for instance the `did_multiplegt_old` Stata command, with the `threshold_stable_treatment(#)` option. The option's argument is h , the bandwidth below which the researcher considers that a cell's treatment did not change. As in Chapter 7, results from non-parametric estimation can be used to choose that bandwidth optimally, up to the additional difficulty that in general designs heterogeneity-robust DID estimators control for the lagged treatment, and therefore the bandwidth has to be chosen conditional on a continuous control variable (de Chaisemartin, D'Haultfœuille and Vazquez-Bare, 2024). While the `did_had` Stata and R packages can be used to compute the optimal bandwidth in heterogeneous adoption designs, we are not aware of a Stata or R package that can be used to compute that bandwidth in a general design. In general designs, a further difficulty with this approach concerns the computation of pre-trend estimators. Remember that in the previous section, pre-trend estimators compared the average of $Y_{g,1} - Y_{g,0}$ between period-one-to-two switchers and stayers, restricting the sample to period-zero-to-one stayers. Here, this implies that pre-trend estimators have to be computed in the subsample of period-zero-to-one quasi-stayers. A second bandwidth h^{pl} needs to be chosen, and that second bandwidth depends on the first one h : the pre-trends estimator will compare $Y_{g,1} - Y_{g,0}$ between gs such that $|D_{g,2} - D_{g,1}| > h$ and gs such that $|D_{g,2} - D_{g,1}| \leq h$, restricting the sample to gs such that $g : |D_{g,1} - D_{g,0}| \leq h^{pl}$. We are not aware of results from non-parametric statistics that one could use to jointly choose those two bandwidths optimally.

8.5.2 Changing the treatment's definition

Another, “quick and easy” way of solving the problem is to round the treatment. In a weather application, with temperatures rounded, say, to the first digit, the control group becomes locations with the same rounded-to-the-first-digit temperatures at periods one and two. Thus, this approach is related to, but different from, using quasi-stayers as the control group. Letting $r(\cdot)$ denote the function mapping the raw treatment into its rounded value, this approach implicitly assumes that $Y_{g,t}(d) = Y_{g,t}(d')$ for all (d, d') such that $r(d) = r(d')$. For instance, with temperatures rounded to the first digit, agricultural yields can change when average temperatures change from 18.79 to 18.80 degrees, but they cannot change when temperatures change from 18.70 to 18.79, an assumption that may be hard to justify. In weather applications, there are sometimes more principled ways to change the treatment’s definition. For instance, there may be well-controlled laboratory evidence suggesting that a crop’s growth is only impaired when temperature goes above a threshold. Then, one can redefine the treatment as the number of days when temperatures exceeded that threshold in location g and year t . With this treatment definition, commonly used in the literature, we are back to a design with some stayers. Beyond agricultural yields, this approach may also be applicable to health outcomes. For instance, there may be a consensus in the medical literature that temperature variations within a range of “normal” temperatures have no effect on mortality. On the other hand, this approach may not be applicable to study the effect of temperatures on GDP. There, it seems harder to determine ex-ante a range of temperatures that would all lead to the same GDP level.

8.5.3 Functional-form assumptions

Alternatively, de Chaisemartin, D’Haultfœuille and Vazquez-Bare (2024) propose an estimator that allows for heterogeneous treatment effects, but relies on parametric functional-form assumptions on groups’ counterfactual outcome evolutions if their treatment had not changed, and on the treatment’s effect. A limitation of this approach is that those parametric functional-form assumptions cannot be tested via pre-trend tests, thus making it hard to assess their plausibility.

8.6 Appendix*

8.6.1 Proof of Theorem 18

Under (8.9), there exists a real number γ_2^Δ such that $E[Y_{g,2}(D_{g,1}) - Y_{g,1}(D_{g,1})] = \gamma_2^\Delta$ for all g .

One has that

$$\begin{aligned} E[\Delta Y_g] &= E[Y_{g,2}(D_{g,2}) - Y_{g,1}(D_{g,1})] \\ &= E[Y_{g,2}(D_{g,2}) - Y_{g,2}(D_{g,1}) + Y_{g,2}(D_{g,1}) - Y_{g,1}(D_{g,1})] \\ &= \gamma_2^\Delta + \Delta D_g \text{TE}_{g,2}^\Delta. \end{aligned} \quad (8.32)$$

Finally,

$$\begin{aligned} E[\hat{\beta}^{\text{fe}}] &= \frac{\sum_{g=1}^G (\Delta D_g - \Delta D.) E[\Delta Y_g]}{\sum_{g=1}^G (\Delta D_g - \Delta D.)^2} \\ &= \frac{\sum_{g=1}^G (\Delta D_g - \Delta D.) (\gamma_2^\Delta + \Delta D_g \text{TE}_{g,2}^\Delta)}{\sum_{g=1}^G (\Delta D_g - \Delta D.)^2} \\ &= \sum_{g=1}^G W_{g,t}^\Delta \text{TE}_{g,2}^\Delta. \end{aligned}$$

QED.

8.6.2 Proof of Theorem 19

We have

$$\begin{aligned} \frac{\text{cov}_u(\Delta D, \Delta Y)}{V_u(\Delta D)} &= \frac{\text{cov}_u(\Delta D, Y_2(0) - Y_1(0)) + \text{cov}_u(\Delta D, S_2 \times \Delta D) + \text{cov}_u(\Delta D, \Delta S \times D_1)}{\sum_{t'=1}^2 [V_u(D_{t'}) - \text{cov}_u(D_1, D_2)]} \\ &= \frac{V_u(\Delta D) E_u(S_2) + \text{cov}_u(\Delta D, D_1) E_u(\Delta S)}{\sum_{t'=1}^2 [V_u(D_{t'}) - \text{cov}_u(D_1, D_2)]} \\ &= \sum_{t=1}^2 \frac{V_u(D_t) - \text{cov}_u(D_1, D_2)}{\sum_{t'=1}^2 [V_u(D_{t'}) - \text{cov}_u(D_1, D_2)]} E_u(S_t). \end{aligned}$$

The first equality follows from (8.14). The second follows from (8.19). **QED.**

8.6.3 Proof of Theorem 20

By (8.22) and (6.13), $Y_{g,t} = \alpha_g + \gamma_t + \sum_{\ell=0}^K \gamma_{g,t}^\ell D_{g,t-\ell} + \varepsilon_{g,t}$, with $E[\varepsilon_{g,t}] = 0$. The rest of the proof is identical to that of Theorem 9, simply replacing $Y_{g,t}(\mathbf{0}_{t-\ell'}, \mathbf{1}_{\ell'}) - Y_{g,t}(\mathbf{0}_t)$ and $\mathbb{1}\{t = F_g - 1 + \ell\}$ by respectively $\gamma_{g,t}^\ell$ and $D_{g,t-\ell}$. **QED.**

8.6.4 Proof of Theorem 21

Let g be such that $F_g \leq T_g$. For any $\ell \in \{1, \dots, T_g - (F_g - 1)\}$,

$$\begin{aligned} E[\widehat{\text{AVSQ}}_{g,\ell}] &= E[Y_{g,F_g-1+\ell} - Y_{g,F_g-1}(\mathbf{D}_{g,1,F_g-1})] \\ &\quad - \frac{1}{\#\mathcal{C}_{g,\ell}} E \left[\sum_{g' \in \mathcal{C}_{g,\ell}} Y_{g',F_g-1+\ell}(\mathbf{D}_{g,1,F_g-1+\ell}) - Y_{g',F_g-1}(\mathbf{D}_{g,1,F_g-1}) \right] \\ &= \text{AVSQ}_{g,\ell} + E[Y_{g,F_g-1+\ell}(\mathbf{D}_{g,1,F_g-1+\ell}) - Y_{g,F_g-1}(\mathbf{D}_{g,1,F_g-1})] \\ &\quad - \frac{1}{\#\mathcal{C}_{g,\ell}} E \left[\sum_{g' \in \mathcal{C}_{g,\ell}} Y_{g',F_g-1+\ell}(\mathbf{D}_{g,1,F_g-1+\ell}) - Y_{g',F_g-1}(\mathbf{D}_{g,1,F_g-1}) \right] \\ &= \text{AVSQ}_{g,\ell}. \end{aligned}$$

The first equality follows from the definitions of F_g and $\widehat{\text{AVSQ}}_{g,\ell}$ and from Assumption NA. The second equality follows from adding and subtracting $Y_{g,F_g-1+\ell}(\mathbf{D}_{g,1,F_g-1+\ell})$ and from the definition of $\text{AVSQ}_{g,\ell}$. The third equality follows from Assumption PTNC and the definition of $N_{F_g-1+\ell}^g$. **QED.**

8.6.5 Proof of Theorem 22

For any $\ell \in \{1, \dots, T_g - (F_g - 1)\}$,

$$\delta_{g,\ell}^n = \frac{E[Y_{g,F_g-1+\ell}(\mathbf{D}_{g,1,F_g-1}, D_{g,F_g}, \dots, D_{g,F_g-1+\ell}) - Y_{g,F_g-1+\ell}(\mathbf{D}_{g,1,F_g-1+\ell}) | \mathbf{D}]}{\sum_{k=0}^{\ell-1} (D_{g,F_g-1+\ell-k} - D_{g,1})}$$

$$\begin{aligned}
&= \sum_{k=0}^{\ell-1} E \left[Y_{g,F_g-1+\ell}(\mathbf{D}_{g,1,F_g-1}, D_{g,F_g}, \dots, D_{g,F_g-1+\ell-k-1}, D_{g,F_g-1+\ell-k}, \mathbf{D}_{g,1,k}) \right. \\
&\quad \left. - Y_{g,F_g-1+\ell}(\mathbf{D}_{g,1,F_g-1}, D_{g,F_g}, \dots, D_{g,F_g-1+\ell-k-1}, D_{g,1}, \mathbf{D}_{g,1,k}) \mid \mathbf{D} \right] / \left(\sum_{k=0}^{\ell-1} (D_{g,F_g-1+\ell-k} - D_{g,1}) \right) \\
&= \sum_{k=0}^{\ell-1} w_{g,\ell,k} s_{g,\ell,k}.
\end{aligned}$$

QED.

8.6.6 Proof of Theorem 23

We have

$$\begin{aligned}
&E \left[Y_{g,t} - Y_{g,t-1} - \frac{1}{\#\{g' : D_{g',t-1} = D_{g',t} = D_{g,t-1}\}} \sum_{g': D_{g',t-1} = D_{g',t} = D_{g,t-1}} (Y_{g',t} - Y_{g',t-1}) \right] \\
&= E [Y_{g,t}(D_{g,t}) - Y_{g,t-1}(D_{g,t-1})] \\
&\quad - \frac{1}{\#\{g' : D_{g',t-1} = D_{g',t} = D_{g,t-1}\}} \sum_{g': D_{g',t-1} = D_{g',t} = D_{g,t-1}} (Y_{g',t}(D_{g,t-1}) - Y_{g',t-1}(D_{g,t-1})) \\
&= E [Y_{g,t}(D_{g,t}) - Y_{g,t}(D_{g,t-1})] \\
&\quad + E [Y_{g,t}(D_{g,t-1}) - Y_{g,t}(D_{g,t-1})] \\
&\quad - \frac{1}{\#\{g' : D_{g',t-1} = D_{g',t} = D_{g,t-1}\}} \sum_{g': D_{g',t-1} = D_{g',t} = D_{g,t-1}} (Y_{g',t}(D_{g,t-1}) - Y_{g',t-1}(D_{g,t-1})) \\
&= E [Y_{g,t}(D_{g,t}) - Y_{g,t}(D_{g,t-1})].
\end{aligned}$$

Dividing the previous display by $D_{g,t} - D_{g,t-1}$ proves the result. **QED.**

Chapter 9

Conclusion: practitioners' checklist

Let us close this textbook with some recommendations for practitioners wishing to leverage a potentially complex natural experiment and a DID-like estimator to learn the effect of a treatment on an outcome:

1. Lay out explicitly the causal effect you want to estimate.
2. Lay out explicitly the no-anticipation and parallel-trends (or factor-model) assumption underlying your identification.
3. Test and argue for your identifying assumption:
 - (a) Report pre-trend or placebo estimators of your identifying assumptions. For your results to be convincing, those pre-trend estimators should be smaller than your actual treatment-effect estimators, and ideally, they should allow you to rule out small differential trends, that cannot account for a large fraction of your treatment-effect estimators. See Chapter 3 for details.
 - (b) If your pre-trends are small and precisely estimated with a simple DID estimator, you may not need to use more complicated estimators, like a DID with control variables or a synthetic control or synthetic DID estimator. See Chapter 4 for details.
 - (c) If you use an interactive fixed effects, a synthetic control, or a synthetic DID estimator, we still recommend conducting a thorough pre-trends/placebo analysis. See Chapter 4 for details.

- (d) If you invoke an “as good as random” assumption (random treatment timing, random treatment dose), you need to conduct balancing checks to substantiate this assumption. For instance, you can regress the treatment timing or dose on all pre-treatment outcomes: pre-treatment outcomes should not predict the treatment timing or dose if those are as good as randomly assigned.
- (e) Even if those tests are conclusive, acknowledge that they remain suggestive, and discuss remaining threats to your identification. For instance, are there concomitant shocks or other policies that could explain why the treated and control groups start experiencing differential trends after the treated get treated?

4. Conduct your estimation:

- (a) Identify your design: is your treatment binary? Is it absorbing? Is there variation in treatment timing?
- (b) Depending on your design, use an appropriate heterogeneity-robust estimator.
 - i. Most treatments that social scientists are interested in are likely to have effects that vary across space and over time, so robustness to heterogeneity is a desirable feature.
 - ii. Heterogeneity-robust DID estimators are more transparent than TWFE estimators, which makes it easier to communicate a study’s methodology and findings outside of academic circles.
 - iii. Heterogeneity-robust estimators typically do not lead to large precision losses. The first large-scale meta-analysis comparing TWFE and heterogeneity-robust estimators shows that for the median paper, using a robust estimator increases standard errors by 10% (Chiu et al., 2023). See Chapter 6 for details.
- (c) If several heterogeneity-robust estimators are available for your research design, you do not need to compute and report all of them: typically, they tend to be close to each other. If they are not, this is evidence that the no-anticipation and parallel-trends assumptions underlying those estimators are violated, but pre-trend tests are a more direct way of testing those assumptions.

- (d) In a classical DID design, TWFE estimators are heterogeneity-robust estimators, so there is no need to use other estimators.
- (e) If you estimate event-study effects, do not report more effects than pre-trend estimators.
- (f) In a complicated design where the treatment is non-binary and/or non-absorbing, you may test whether lagged treatments affect the outcome. If there is no evidence of lagged effects, you may assume those effects away: doing so will allow you to estimate easy-to-interpret effects, under a minimal parallel-trends assumption. If there is evidence of lagged effects, you can estimate normalized event-study effects, equal to a weighted average of the effects of the current and lagged treatments on the outcome. See Chapter 8 for details.

5. Perform inference:

- (a) Cluster standard errors, either at the level at which the treatment is assigned, or at the most disaggregated level at which one can still construct a panel dataset. See starred Section 2.4 for details.
- (b) With at least 40 treated and 40 control groups, you can use confidence intervals relying on large-sample approximations.
- (c) With less than 40 treated or less than 40 control groups, we recommend that you conduct simulations based on your data to assess if, in your data, confidence intervals relying on large-sample approximations have close-to-nominal coverage. If those simulations suggest that they do not, see Section 3.3 for alternative inference procedures.

Bibliography

- Abadie, A. (2005), ‘Semiparametric difference-in-differences estimators’, *Review of Economic Studies* **72**(1), 1–19.
- Abadie, A. (2021), ‘Using synthetic controls: Feasibility, data requirements, and methodological aspects’, *Journal of Economic Literature* **59**(2), 391–425.
- Abadie, A., Athey, S., Imbens, G. W. and Wooldridge, J. M. (2023), ‘When should you adjust standard errors for clustering?’, *The Quarterly Journal of Economics* **138**(1), 1–35.
- Abadie, A., Diamond, A. and Hainmueller, J. (2010), ‘Synthetic control methods for comparative case studies: Estimating the effect of California’s tobacco control program’, *Journal of the American Statistical Association* **105**(490), 493–505.
- Abadie, A., Diamond, A. and Hainmueller, J. (2011), ‘SYNTH: Stata module to implement Synthetic Control Methods for Comparative Case Studies’, Statistical Software Components, Boston College Department of Economics.
- Abadie, A. and Gardeazabal, J. (2003), ‘The economic costs of conflict: A case study of the Basque country’, *American Economic Review* **93**(1), 113–132.
- Acemoglu, D., Autor, D., Dorn, D., Hanson, G. H. and Price, B. (2016), ‘Import competition and the great US employment sag of the 2000s’, *Journal of Labor Economics* **34**(S1), S141–S198.
- Ahrens, A., Chernozhukov, V., Hansen, C., Kozbur, D., Schaffer, M. and Wiemann, T. (2025), ‘An introduction to double/debiased machine learning’, *arXiv preprint arXiv:2504.08324*.
- Andrews, I., Roth, J. and Pakes, A. (2023), ‘Inference for linear conditional moment inequalities’, *Review of Economic Studies* **90**(6), 2763–2791.
- Angrist, J. D. (1998), ‘Estimating the labor market impact of voluntary military service using social security data on military applicants’, *Econometrica* **66**(2), 249–288.
- Angrist, J. D. and Pischke, J.-S. (2009), *Mostly harmless econometrics: An empiricist’s companion*, Princeton University Press.
- Arboleda Cárcamo, D. (2024), ‘Fitting a curve to the pre-trends’.

- Arellano, M. and Bond, S. (1991), ‘Some tests of specification for panel data: Monte carlo evidence and an application to employment equations’, *Review of Economic Studies* **58**(2), 277–297.
- Arellano, M. and Bonhomme, S. (2012), ‘Identifying distributional characteristics in random coefficients panel data models’, *Review of Economic Studies* **79**(3), 987–1020.
- Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W. and Wager, S. (2021), ‘Synthetic difference-in-differences’, *American Economic Review* **111**(12), 4088–4118.
- Arkhangelsky, D. and Imbens, G. (2024), ‘Causal models for longitudinal and panel data: A survey’, *The Econometrics Journal* **27**(3), C1–C61.
- Arkhangelsky, D., Imbens, G. W., Lei, L. and Luo, X. (2024), ‘Design-robust two-way-fixed-effects regression for panel data’, *Quantitative Economics* **15**, 999–1034.
- Armstrong, T. B., Weidner, M. and Zeleniev, A. (2022), Robust estimation and inference in panels with interactive fixed effects. arXiv preprint arXiv:2210.06639.
- Ashenfelter, O. (1978), ‘Estimating the effect of training programs on earnings’, *The Review of Economics and Statistics* **60**, 47–57.
- Athey, S. and Imbens, G. W. (2006), ‘Identification and inference in nonlinear difference-in-differences models’, *Econometrica* **74**(2), 431–497.
- Athey, S. and Imbens, G. W. (2022), ‘Design-based analysis in difference-in-differences settings with staggered adoption’, *Journal of Econometrics* **226**, 62–79.
- Autor, D. H., Dorn, D. and Hanson, G. H. (2013), ‘The china syndrome: Local labor market effects of import competition in the united states’, *American economic review* **103**(6), 2121–2168.
- Bach, L., Bozio, A., Guillouzouic, A. and Malgouyres, C. (2023), ‘Dividend taxes and the allocation of capital: Comment’, *American Economic Review* **113**(7), 2048–2052.
- Bai, J. (2009), ‘Panel data models with interactive fixed effects’, *Econometrica* **77**(4), 1229–1279.
- Bai, J. and Ng, S. (2021), ‘Matrix completion, counterfactuals, and factor analysis of missing data’, *Journal of the American Statistical Association* **116**(536), 1746–1763.
- Baker, A., Callaway, B., Cunningham, S., Goodman-Bacon, A. and Sant’Anna, P. H. (2025), ‘Difference-in-differences designs: A practitioner’s guide’, *arXiv preprint arXiv:2503.13323*.
- Bell, R. M. and McCaffrey, D. F. (2002), ‘Bias reduction in standard errors for linear regression with multi-stage samples’, *Survey Methodology* **28**(2), 169–182.

- Bellégo, C., Benatia, D. and Dortet-Bernardet, V. (forthc.), ‘The chained difference-in-differences’, *Journal of Econometrics* .
- Benatia, D., Bellégo, C., Cuerrier, J. and Dortet-Bernadet, V. (2025), ‘CDID: R module implementing the chained DID estimator’, CRAN.
- Bertrand, M., Duflo, E. and Mullainathan, S. (2004), ‘How much should we trust differences-in-differences estimates?’, *The Quarterly Journal of Economics* **119**(1), 249–275.
- Bester, C. A., Conley, T. G. and Hansen, C. B. (2011), ‘Inference with dependent data using cluster covariance estimators’, *Journal of Econometrics* **165**(2), 137–151.
- Bickel, P. J. (1969), ‘A distribution free version of the smirnov two sample test in the p-variate case’, *The Annals of Mathematical Statistics* **40**(1), 1–23.
- Blundell, R. and Costa-Dias, M. (2009), ‘Alternative approaches to evaluation in empirical microeconomics’, *Journal of human resources* **44**(3), 565–640.
- Blundell, R., Costa-Dias, M., Meghir, C. and Van Reenen, J. (2004), ‘Evaluating the employment impact of a mandatory job search program’, *Journal of the European Economic Association* **2**(4), 569–606.
- Bojinov, I., Rambachan, A. and Shephard, N. (2021), ‘Panel experiments and dynamic causal effects: A finite population perspective’, *Quantitative Economics* **12**, 1171–1196.
- Bonhomme, S. and Robin, J.-M. (2010), ‘Generalized non-parametric deconvolution with an application to earnings dynamics’, *Review of Economic Studies* **77**(2), 491–533.
- Borusyak, K. (2021), ‘DID_IMPUTATION: Stata module to perform treatment effect estimation and pre-trend testing in event studies’.
URL: <https://ideas.repec.org/c/boc/bocode/s458957.html>
- Borusyak, K., Hull, P. and Jaravel, X. (2022), ‘Quasi-experimental shift-share research designs’, *Review of Economic Studies* **89**(1), 181–213.
- Borusyak, K. and Jaravel, X. (2017), Revisiting event study designs. Working Paper.
- Borusyak, K., Jaravel, X. and Spiess, J. (2024), ‘Revisiting event-study designs: robust and efficient estimation’, *Review of Economic Studies* p. rdae007.
- Bravo, M. C., Roth, J. and Rambachan, A. (2022), ‘Honestdid: Stata module implementing the honestdid r package’.
URL: <https://EconPapers.repec.org/RePEc:boc:bocode:s459138>
- Brown, N. and Butts, K. (2023), Dynamic treatment effect estimation with interactive fixed effects and short panels, Technical report, Mimeo.

- Burgess, R., Jedwab, R., Miguel, E., Morjaria, A. and Padró i Miquel, G. (2015), ‘The value of democracy: evidence from road building in kenya’, *American Economic Review* **105**(6), 1817–51.
- Busch, A. and Girardi, D. (2023), ‘LPDID: Stata module implementing Local Projections Difference-in-Differences (LP-DiD) estimator’, Statistical Software Components, Boston College Department of Economics.
- Butts, K. (2021a), did2s: Two-stage difference-in-differences following gardner (2021), Technical report.
- URL:** <https://github.com/kylebutts/did2s/>
- Butts, K. (2021b), ‘didimputation: Imputation Estimator from Borusyak, Jaravel, and Spiess (2021) in R’.
- URL:** <https://cran.r-project.org/web/packages/didimputation/index.html>
- Butts, K. (2021c), Difference-in-differences estimation with spatial spillovers. arXiv preprint arXiv:2105.03737.
- Butts, K. and Gardner, J. (2021), did2s: Two-Stage Difference-in-Differences, Papers, arXiv.org.
- Caetano, C., Callaway, B., Payne, S. and Rodrigues, H. S. (2022), Difference in differences with time-varying covariates. arXiv preprint arXiv:2202.02903.
- Callaway, B. (2023), *qte: Quantile Treatment Effects*. R package version 1.4.0.
- Callaway, B., Goodman-Bacon, A. and Sant’Anna, P. H. (2021), Difference-in-differences with a continuous treatment. arXiv preprint arXiv:2107.02637.
- Callaway, B. and Sant’Anna, P. H. (2021), ‘Difference-in-differences with multiple time periods’, *Journal of Econometrics* **225**, 200–230.
- Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2018), ‘On the effect of bias estimation on coverage accuracy in nonparametric inference’, *Journal of the American Statistical Association* **113**(522), 767–779.
- Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2019), nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference. arXiv preprint arXiv:1906.00198.
- Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014), ‘Robust nonparametric confidence intervals for regression-discontinuity designs’, *Econometrica* **82**(6), 2295–2326.
- Cameron, A. C. and Miller, D. L. (2015), ‘A practitioner’s guide to cluster-robust inference’, *Journal of human resources* **50**(2), 317–372.
- Cantoni, E. and Pons, V. (2021), ‘Strict id laws don’t stop voters: Evidence from a us nationwide panel, 2008–2018’, *The Quarterly Journal of Economics* **136**(4), 2615–2660.

- Carter, A. V., Schnepel, K. T. and Steigerwald, D. G. (2017), ‘Asymptotic behavior of at-test robust to cluster heterogeneity’, *Review of Economics and Statistics* **99**(4), 698–709.
- Chabé-Ferret, S. (2015), ‘Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes’, *Journal of Econometrics* **185**(1), 110–123.
- Chang, N.-C. (2020), ‘Double/debiased machine learning for difference-in-differences models’, *The Econometrics Journal* **23**(2), 177–191.
- Chen, X., Christensen, T. and Kankanala, S. (2024), ‘Adaptive estimation and uniform confidence bands for nonparametric structural functions and elasticities’, *Review of Economic Studies*.
- Chernozhukov, V., Fernández-Val, I., Hahn, J. and Newey, W. (2013), ‘Average and quantile effects in nonseparable panel models’, *Econometrica* **81**(2), 535–580.
- Chernozhukov, V., Wüthrich, K. and Zhu, Y. (2021), ‘An exact and robust conformal inference method for counterfactual and synthetic controls’, *Journal of the American Statistical Association* **116**(536), 1849–1864.
- Chiu, A., Lan, X., Liu, Z. and Xu, Y. (2023), ‘What to do (and not to do) with causal panel analysis under parallel trends: Lessons from a large reanalysis study’, *arXiv preprint arXiv:2309.15983*.
- Clarke, D. (2017), ‘Estimating difference-in-differences in the presence of spillovers’.
- Conley, T. G. and Taber, C. R. (2011), ‘Inference with “difference in differences” with a small number of policy changes’, *The Review of Economics and Statistics* **93**(1), 113–125.
- Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2009), ‘Dealing with limited overlap in estimation of average treatment effects’, *Biometrika* **96**(1), 187–199.
- Daw, J. R. and Hatfield, L. A. (2018), ‘Matching and regression to the mean in difference-in-differences analysis’, *Health services research* **53**(6), 4138–4156.
- de Chaisemartin, C. (2010), A note on instrumented difference in differences. Working Paper.
- de Chaisemartin, C. (2011), Fuzzy differences in differences. Working Paper 2011-10, Center for Research in Economics and Statistics.
- de Chaisemartin, C., Ciccia, D., D'Haultfoeuille, X., Knau, F. and Sow, D. (2024), ‘Stute_test: Stata module to perform stute (1997) linearity test’.
- de Chaisemartin, C., Ciccia, D., D'Haultfoeuille, X. and Knau, F. (2024), Two-way fixed effects and differences-in-differences estimators in heterogeneous-adoption designs. *arXiv preprint arXiv:2405.04465*.

de Chaisemartin, C., Ciccia, D., D'Haultfoeuille, X., Knau, F., Malézieux, M. and Sow, D. (2024a), 'Didmultiplegtdyn: R module to estimate event-study difference-in-difference (did) estimators in designs with multiple groups and periods, with a potentially non-binary treatment that may increase or decrease multiple times'.

URL: <https://cran.r-project.org/web/packages/DIDmultiplegtdYN/index.html>

de Chaisemartin, C., Ciccia, D., D'Haultfoeuille, X., Knau, F., Malézieux, M. and Sow, D. (2024b), 'Did_multiplegtd_dyn: Stata module to estimate event-study difference-in-difference (did) estimators in designs with multiple groups and periods, with a potentially non-binary treatment that may increase or decrease multiple times'.

de Chaisemartin, C., Ciccia, D., D'Haultfoeuille, X., Knau, F., Malézieux, M. and Sow, D. (2024), 'Event-study estimators and variance estimators computed by the did_multiplegtd_dyn command'.

de Chaisemartin, C., Ciccia, D., D'Haultfoeuille, X., Knau, F. and Sow, D. (2024a), 'Didhad: R module to estimate the effect of a treatment on an outcome in a heterogeneous-adoption design with no stayers but some quasi stayers'.

URL: <https://cran.r-project.org/web/packages/YatchewTest/index.html>

de Chaisemartin, C., Ciccia, D., D'Haultfoeuille, X., Knau, F. and Sow, D. (2024b), 'Didhad: Stata module to estimate the effect of a treatment on an outcome in a heterogeneous-adoption design with no stayers but some quasi stayers'.

de Chaisemartin, C., Ciccia, D., D'Haultfoeuille, X., Knau, F. and Sow, D. (2024c), 'Did_multiplegtd_stat: Stata module to estimate event-study difference-in-difference (did) estimators in designs with multiple groups and periods, with a potentially non-binary treatment that may increase or decrease multiple times'.

de Chaisemartin, C., Ciccia, D., D'Haultfoeuille, X., Knau, F. and Sow, D. (2024d), 'Stutetest: Stata module to perform stute (1997) linearity test'.

URL: <https://cran.r-project.org/web/packages/StuteTest/index.html>

de Chaisemartin, C. and D'Haultfœuille, X. (2015), Fuzzy differences-in-differences. arXiv preprint arXiv:1510.01757v2.

de Chaisemartin, C. and D'Haultfœuille, X. (2018), 'Fuzzy differences-in-differences', *Review of Economic Studies* **85**(2), 999–1028.

de Chaisemartin, C. and D'Haultfœuille, X. (2020), 'Two-way fixed effects estimators with heterogeneous treatment effects', *American Economic Review* **110**(9), 2964–2996.

de Chaisemartin, C. and D'Haultfœuille, X. (2021), Difference-in-differences estimators of intertemporal treatment effects. arXiv preprint arXiv:2007.04267.

- de Chaisemartin, C. and D'Haultfœuille, X. (2023a), ‘Two-way fixed effects and differences-in-differences estimators with several treatments’, *Journal of Econometrics* **236**(2).
- de Chaisemartin, C. and D'Haultfœuille, X. (2023b), ‘Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey’, *The Econometrics Journal* **26**(3), C1–C30.
- de Chaisemartin, C. and D'Haultfœuille, X. (2024), Under the null of valid specification, pre-tests cannot make post-test inference liberal. arXiv e-prints arXiv:2407.03725.
- de Chaisemartin, C. and D'Haultfœuille, X. (2025), ‘Difference-in-differences estimators of intertemporal treatment effects’, *Review of Economics and Statistics*.
- de Chaisemartin, C., D'Haultfœuille, X. and Deeb, A. (2019), ‘twowayfeweights: Estimation of the Weights Attached to the Two-Way Fixed Effects Regressions in Stata’.
URL: <https://ideas.repec.org/c/boc/bocode/s458611.html>
- de Chaisemartin, C., D'Haultfœuille, X. and Guyonvarch, Y. (2019), ‘Fuzzy differences-in-differences with stata’, *Stata Journal* **19**(2), 435–458.
- de Chaisemartin, C., D'Haultfœuille, X., Pasquier, F., Sow, D. and Vazquez-Bare, G. (2022), Difference-in-differences for continuous treatments and instruments with stayers. arXiv preprint arXiv:2201.06898.
- de Chaisemartin, C., D'Haultfœuille, X. and Vazquez-Bare, G. (2024), Difference-in-difference estimators with continuous treatments and no stayers, in ‘AEA Papers and Proceedings’, Vol. 114, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, pp. 610–613.
- de Chaisemartin, C. and Lei, Z. (2021), Are bartik regressions always robust to heterogeneous treatment effects? Available at SSRN 3802200.
- de Chaisemartin, C. and Lei, Z. (2024), ‘Randomly assigned first differences?’, *arXiv preprint arXiv:2411.03208*.
- De Tocqueville, A. (1850), *La démocratie en Amérique*, Pagnerre.
- Deeb, A. and de Chaisemartin, C. (2019), Clustering and external validity in randomized controlled trials. arXiv preprint arXiv:1912.01052.
- Deryugina, T. (2017), ‘The fiscal cost of hurricanes: Disaster aid versus social insurance’, *American Economic Journal: Economic Policy* **9**(3), 168–98.
- Deschênes, O. and Greenstone, M. (2007), ‘The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather’, *American Economic Review* **97**(1), 354–385.

- Diamond, A. and Hainmueller, J. (2023), ‘SYNTH: R module to implement Synthetic Control Methods for Comparative Case Studies’, CRAN.
- DiCiccio, C. J. and Romano, J. P. (2017), ‘Robust Permutation Tests for Correlation and Regression Coefficients’, *Journal of the American Statistical Association* **112**, 1211–1220.
- Donald, S. G. and Lang, K. (2007), ‘Inference with difference-in-differences and other panel data’, *The review of Economics and Statistics* **89**(2), 221–233.
- Doudchenko, N. and Imbens, G. W. (2016), Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. National Bureau of Economic Research.
- Drukker, D. M. (2023), ‘Simultaneous tests and confidence bands for stata estimation commands’, *The Stata Journal* **23**(2), 518–544.
- Dube, A., Girardi, D., Jorda, O. and Taylor, A. M. (2023), A local projections approach to difference-in-differences event studies. National Bureau of Economic Research.
- East, C. N., Miller, S., Page, M. and Wherry, L. R. (2023), ‘Multigenerational impacts of childhood access to the safety net: Early life exposure to medicaid and the next generation’s health’, *American Economic Review* **113**(1), 98–135.
- Egerod, B. C. and Hollenbach, F. M. (2024), ‘How many is enough? sample size in staggered difference-in-differences designs’, *OSF Preprint*.
- Enikolopov, R., Petrova, M. and Zhuravskaya, E. (2011), ‘Media and political persuasion: Evidence from russia’, *American Economic Review* **101**(7), 3253–3285.
- Fabre, A. (2023), ‘Robustness of two-way fixed effects estimators to heterogeneous treatment effects’.
- Favara, G. and Imbs, J. (2015), ‘Credit supply and the price of housing’, *American Economic Review* **105**(3), 958–92.
- Ferman, B. (2019), Assessing inference methods. arXiv preprint arXiv:1912.08772.
- Ferman, B. (2021), ‘On the properties of the synthetic control estimator with many periods and many controls’, *Journal of the American Statistical Association* **116**(536), 1764–1772.
- Ferman, B. and Pinto, C. (2019), ‘Inference in differences-in-differences with few treated groups and heteroskedasticity’, *Review of Economics and Statistics* **101**(3), 452–467.
- Ferman, B., Pinto, C. and Possebom, V. (2020), ‘Cherry picking with synthetic controls’, *Journal of Policy Analysis and Management* **39**(2), 510–532.
- Fetzer, T. (2019), ‘Did austerity cause brexit?’, *American Economic Review* **109**(11), 3849–3886.

- Field, E. (2007), ‘Entitled to work: Urban property rights and labor supply in peru’, *The Quarterly Journal of Economics* **122**(4), 1561–1602.
- Flack, E. and Edward (2020), ‘bacondecomp: Goodman-Bacon Decomposition in R’.
URL: <https://cran.r-project.org/web/packages/bacondecomp/index.html>
- Fricke, H. (2017), ‘Identification based on difference-in-differences approaches with multiple treatments’, *Oxford Bulletin of Economics and Statistics* **79**(3), 426–433.
- Friedberg, L. (1998), ‘Did unilateral divorce raise divorce rates? evidence from panel data’, *The American Economic Review* **88**(3), 608–627.
- Frison, L. and Pocock, S. J. (1992), ‘Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design’, *Statistics in medicine* **11**(13), 1685–1704.
- Fuest, C., Peichl, A. and Siegloch, S. (2018), ‘Do higher corporate taxes reduce wages? micro evidence from germany’, *American Economic Review* **108**(2), 393–418.
- Gardner, J. (2021), Two-stage differences in differences. Working paper.
- Gardner, J., Thakral, N., Tô, L. T. and Yap, L. (2024), ‘Two-stage differences in differences’.
- Gentzkow, M., Shapiro, J. M. and Sinkinson, M. (2011), ‘The effect of newspaper entry and exit on electoral politics’, *American Economic Review* **101**(7), 2980–3018.
- Gobillon, L. and Magnac, T. (2016), ‘Regional policy evaluation: Interactive fixed effects and synthetic controls’, *Review of Economics and Statistics* **98**(3), 535–551.
- Goldsmith-Pinkham, P., Sorkin, I. and Swift, H. (2020), ‘Bartik instruments: What, when, why, and how’, *American Economic Review* **110**(8), 2586–2624.
- Goodman-Bacon, A. (2021), ‘Difference-in-differences with variation in treatment timing’, *Journal of Econometrics* **225**, 254–277.
- Goodman-Bacon, A., Goldring, T. and Nichols, A. (2019), ‘BACONDECOMP: Stata module to perform a Bacon decomposition of difference-in-differences estimation’.
URL: <https://ideas.repec.org/c/boc/bocode/s458676.html>
- Graham, B. S. and Powell, J. L. (2012), ‘Identification and estimation of average partial effects in “irregular” correlated random coefficient panel data models’, *Econometrica* **80**(5), 2105–2152.
- Harmon, N. (2024), ‘DID_STEPWISE: Stata module implementing the chained DID estimator’.
- Harmon, N. A. (2022), ‘Difference-in-differences and efficient estimation of treatment effects’.
- Hatamyar, J., Kreif, N., Rocha, R. and Huber, M. (2023), Machine learning for staggered difference-in-differences and dynamic treatment effect heterogeneity. arXiv preprint arXiv:2310.11962.

- Heckman, J. J., Ichimura, H. and Todd, P. E. (1997), ‘Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme’, *Review of Economic Studies* **64**(4), 605–654.
- Hirshberg, D. (2023), ‘SYNTHDID: R module to perform synthetic difference-in-differences estimation, inference, and visualization’.
- Hoehn-Velasco, L., Penglase, J., Pesko, M. and Shahid, H. (2024), ‘The California effect: The challenges of identifying the impact of social policies during an era of social change’, Available at SSRN 4802701 .
- Hsiao, C., Ching, H. S. and Ki Wan, S. (2012), ‘A panel data approach for program evaluation: measuring the benefits of political and economic integration of Hong Kong with mainland China’, *Journal of Applied Econometrics* **27**(5), 705–740.
- Hudson, S., Hull, P. and Liebersohn, C. (2015), Interpreting instrumented difference-in-differences, Technical report, Working Paper (available upon request).
- Imai, K. and Kim, I. S. (2021), ‘On the use of two-way fixed effects regression models for causal inference with panel data’, *Political Analysis* **29**(3), 405–415.
- Imbens, G., Kallus, N. and Mao, X. (2021), Controlling for unmeasured confounding in panel data using minimal bridge functions: From two-way fixed effects to factor models. arXiv preprint arXiv:2108.03849.
- Imbens, G. and Kalyanaraman, K. (2012), ‘Optimal bandwidth choice for the regression discontinuity estimator’, *Review of Economic Studies* **79**(3), 933–959.
- Imbens, G. W. and Angrist, J. D. (1994), ‘Identification and estimation of local average treatment effects’, *Econometrica* **62**(2), 467–475.
- Imbens, G. W. and Kolesar, M. (2016), ‘Robust standard errors in small samples: Some practical advice’, *Review of Economics and Statistics* **98**(4), 701–712.
- Imbens, G. W., Rubin, D. B. and Sacerdote, B. I. (2001), ‘Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players’, *American Economic Review* **91**(4), 778–794.
- Imbens, G. and Xu, Y. (2024), Lalonde (1986) after nearly four decades: Lessons learned. arXiv preprint arXiv:2406.00827.
- Ishimaru, S. (2021), ‘What do we get from two-way fixed effects regressions? implications from numerical equivalence’, arXiv preprint arXiv:2103.12374 .
- Jordà, Ò. (2005), ‘Estimation and inference of impulse responses by local projections’, *American Economic Review* **95**(1), 161–182.

- Kahn-Lang, A. and Lang, K. (2020), ‘The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications’, *Journal of Business & Economic Statistics* **38**(3), 613–620.
- Kim, D. (2024), ‘DDID: STATA ado package for “Difference-in-differences Estimator of Quantile Treatment Effect on the Treated”’.
- Kim, D. and Wooldridge, J. M. (2024), ‘Difference-in-differences estimator of quantile treatment effect on the treated’, *Journal of Business & Economic Statistics* pp. 1–12.
- Kiviet, J. F. (1995), ‘On bias, inconsistency, and efficiency of various estimators in dynamic panel data models’, *Journal of Econometrics* **68**(1), 53–78.
- Klosin, S. (2024), Dynamic biases of static panel data estimators. arXiv preprint arXiv:2410.16112.
- Koenker, R. (2005), *Quantile regression*, Vol. 38, Cambridge university press.
- Kolesar, M. (2023), ‘dfadjust: Degrees of Freedom Adjustment for Robust Standard Errors’. URL: <https://cran.r-project.org/web/packages/dfadjust/index.html>
- Kranker, K. (2019), ‘CIC: Stata module to implement the Athey and Imbens (2006) Changes-in-Changes model’, Statistical Software Components, Boston College Department of Economics.
- Krolikowski, P. (2018), ‘Choosing a control group for displaced workers’, *ILR Review* **71**(5), 1232–1254.
- Lechner, M. (2011), ‘The estimation of causal effects by difference-in-difference methods’, *Foundations and Trends® in Econometrics* **4**(3), 165–224.
- Lee, C. H. and Steigerwald, D. G. (2018), ‘Inference for clustered data’, *The Stata Journal* **18**(2), 447–460.
- Leeb, H. and Pötscher, B. M. (2003), ‘The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations’, *Econometric Theory* **19**(1), 100–142.
- Liu, L., Wang, Y. and Xu, Y. (2024), ‘A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data’, *American Journal of Political Science* **68**(1), 160–176.
- Liu, L., Wang, Y., Xu, Y., Liu, Z. and Liu, S. (2022a), ‘Fect: Companion r package to “a practical guide to counterfactual estimators for causal inference with time-series cross-sectional data”’.
- Liu, L., Wang, Y., Xu, Y., Liu, Z. and Liu, S. (2022b), ‘Fect: Companion stata package to “a practical guide to counterfactual estimators for causal inference with time-series cross-sectional data”’.

- Lu, C., Nie, X. and Wager, S. (2019), Robust nonparametric difference-in-differences estimation. arXiv e-prints.
- MacKinnon, J. G., Nielsen, M. Ø. and Webb, M. D. (2023), ‘Fast and reliable jackknife and bootstrap methods for cluster-robust inference’, *Journal of Applied Econometrics* **38**(5), 671–694.
- MacKinnon, J. G. and Webb, M. D. (2020), ‘Randomization inference for difference-in-differences with few treated clusters’, *Journal of Econometrics* **218**(2), 435–450.
- Manski, C. F. and Pepper, J. V. (2018), ‘How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions’, *Review of Economics and Statistics* **100**(2), 232–244.
- McKenzie, D. (2012), ‘Beyond baseline and follow-up: The case for more t in experiments’, *Journal of development Economics* **99**(2), 210–221.
- McNemar, Q. (1940), ‘Sampling in psychological research.’, *Psychological Bulletin* **37**(6), 331.
- Meinhofer, A., Witman, A. E., Hinde, J. M. and Simon, K. (2021), ‘Marijuana liberalization policies and perinatal health’, *Journal of health economics* **80**, 102537.
- Meyer, B. D., Viscusi, W. K. and Durbin, D. L. (1995), ‘Workers’ compensation and Injury duration: Evidence from a natural experiment’, *The American Economic Review* **85**(3), 322–340.
- Montiel Olea, J. L. and Plagborg-Møller, M. (2019), ‘Simultaneous confidence bands: Theory, implementation, and an application to svards’, *Journal of Applied Econometrics* **34**(1), 1–17.
- Mora, R. and Reggio, I. (2019), ‘Alternative diff-in-diffs estimators with several pretreatment periods’, *Econometric Reviews* **38**(5), 465–486.
- Moser, P. and Voena, A. (2012), ‘Compulsory licensing: Evidence from the trading with the enemy act’, *American Economic Review* **102**(1), 396–427.
- Nagengast, A. J. and Yotov, Y. V. (2025), ‘Staggered difference-in-differences in gravity settings: Revisiting the effects of trade agreements’, *American Economic Journal: Applied Economics* **17**(1), 271–296.
- Neyman, J., Dabrowska, D. M. and Speed, T. P. (1990), ‘On the application of probability theory to agricultural experiments. essay on principles. section 9.’, *Statistical Science* **5**, 465–472.
- Neyman, J. and Scott, E. L. (1948), ‘Consistent estimates based on partially consistent observations’, *Econometrica: Journal of the Econometric Society* pp. 1–32.
- Nickell, S. (1981), ‘Biases in dynamic models with fixed effects’, *Econometrica* **49**, 1417–1426.

- Pailañir, D., Clarke, D. and Ciccia, D. (2022), ‘SDID: Stata module to perform synthetic difference-in-differences estimation, inference, and visualization’, Statistical Software Components, Boston College Department of Economics.
- Pierce, J. R. and Schott, P. K. (2016), ‘The surprisingly swift decline of us manufacturing employment’, *American Economic Review* **106**(7), 1632–62.
- Poterba, J. M., Venti, S. F. and Wise, D. A. (1995), ‘Do 401 (k) contributions crowd out other personal saving?’, *Journal of Public Economics* **58**(1), 1–32.
- Puhani, P. A. (2012), ‘The treatment effect, the cross difference, and the interaction term in nonlinear “difference-in-differences” models’, *Economics Letters* **115**(1), 85–87.
- Rambachan, A. (2022), ‘Robust inference in difference-in-differences and event study designs’.
URL: <https://github.com/asheshrambachan/HonestDiD>
- Rambachan, A. and Roth, J. (2023), ‘A more credible approach to parallel trends’, *Review of Economic Studies* **90**(5), 2555–2591.
- Rico-Straffon, J., Wang, Z., Panlasigui, S., Loucks, C. J., Swenson, J. and Pfaff, A. (2023), ‘Forest concessions and eco-certifications in the peruvian amazon: Deforestation impacts of logging rights and logging restrictions’, *Journal of Environmental Economics and Management* **118**, 102780.
- Rios-Avila, F., Sant’Anna, P. and Callaway, B. (2021), ‘Csdid: Stata module for the estimation of difference-in-difference models with multiple time periods’.
URL: <https://EconPapers.repec.org/RePEc:boc:bocode:s458976>
- Rios-Avila, F., Sant’Anna, P. H. and Naqvi, A. (2021), ‘DRDID: Stata module for the estimation of Doubly Robust Difference-in-Difference models’, Statistical Software Components, Boston College Department of Economics.
URL: <https://ideas.repec.org/c/boc/bocode/s458977.html>
- Robins, J. (1986), ‘A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect’, *Mathematical modelling* **7**(9-12), 1393–1512.
- Roth, J. (2022), ‘Pretest with caution: Event-study estimates after testing for parallel trends’, *American Economic Review: Insights* **4**(3), 305–22.
- Roth, J. and Sant’Anna, P. H. (2023), ‘When is parallel trends sensitive to functional form?’, *Econometrica* **91**(2), 737–747.
- Roth, J. and Sant’Anna, P. H. (2023), ‘Efficient estimation for staggered rollout designs’, *Journal of Political Economy Microeconomics* **1**(4), 669–709.

- Royen, T. (2014), ‘A simple proof of the gaussian correlation conjecture extended to multivariate gamma distributions’, *Far Eastern Journal of Theoretical Statistics* **48**, 139—145.
- Rubin, D. (1974), ‘Estimating causal effects of treatments in randomized and nonrandomized studies’, *Journal of Educational Psychology* **66**(5).
- Rubin, D. B. (1978), ‘Bayesian inference for causal effects: The role of randomization’, *The Annals of Statistics* **6**, 34–58.
- Sant’Anna, P. and Callaway, B. (2021), ‘did: Treatment effects with multiple periods and groups in R’.
- URL:** <https://cran.r-project.org/web/packages/did/index.html>
- Sant’Anna, P. H. C. and Zhao, J. (2022), ‘DRDID: Doubly Robust Difference-in-Differences Estimators in R’.
- URL:** <https://cran.r-project.org/web/packages/DRDID/index.html>
- Sant’Anna, P. H. and Zhao, J. (2020), ‘Doubly robust difference-in-differences estimators’, *Journal of Econometrics* **219**(1), 101–122.
- Sasaki, Y. and Ura, T. (2021), Slow movers in panel data. arXiv preprint arXiv:2110.12041.
- Schmidheiny, K. and Siegloch, S. (2023), ‘On event studies and distributed-lags in two-way fixed effects models: Identification, equivalence, and generalization’, *Journal of Applied Econometrics* **38**(5), 695–713.
- Semmelweis, I. F. (1983), *The etiology, concept, and prophylaxis of childbed fever*, number 2, Univ of Wisconsin Press.
- Shahn, Z. (2023), Subgroup difference in differences to identify effect modification without a control group. arXiv preprint arXiv:2306.11030.
- Silva, J. S. and Tenreyro, S. (2006), ‘The log of gravity’, *The Review of Economics and Statistics* **88**(4), 641–658.
- Small, D. S., Tan, Z., Ramsahai, R. R., Lorch, S. A. and Brookhart, M. A. (2017), Instrumental variable estimation with a stochastic monotonicity assumption. Working paper.
- Snow, J. (1856), ‘On the mode of communication of cholera’, *Edinburgh medical journal* **1**(7), 668.
- Solon, G., Haider, S. J. and Wooldridge, J. M. (2015), ‘What are we weighting for?’, *Journal of Human resources* **50**(2), 301–316.
- Stevenson, B. and Wolfers, J. (2006), ‘Bargaining in the shadow of the law: Divorce laws and family distress’, *The Quarterly Journal of Economics* **121**(1), 267–288.

- Stute, W. (1997), ‘Nonparametric model checks for regression’, *The Annals of Statistics* **25**, 613–641.
- Stute, W., Manteiga, W. G. and Quindimil, M. P. (1998), ‘Bootstrap approximations in model checks for regression’, *Journal of the American Statistical Association* **93**(441), 141–149.
- Sun, L. (2020), ‘EVENTSTUDYWEIGHTS: Stata module to estimate the implied weights on the cohort-specific average treatment effects on the treated (CATTs) (event study specifications)’. **URL:** <https://ideas.repec.org/c/boc/bocode/s458833.html>
- Sun, L. (2021), ‘EVENTSTUDYINTERACT: Stata module to implement the interaction weighted estimator for an event study’. **URL:** <https://ideas.repec.org/c/boc/bocode/s458978.html>
- Sun, L. and Abraham, S. (2021), ‘Estimating dynamic treatment effects in event studies with heterogeneous treatment effects’, *Journal of Econometrics* **225**, 175–199.
- Ujhelyi, G. (2014), ‘Civil service rules and policy choices: evidence from us state governments’, *American Economic Journal: Economic Policy* **6**(2), 338–380.
- van der Vaart, A. and Wellner, J. A. (2023), *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Nature.
- Weiss, A. (2024), How much should we trust modern difference-in-differences estimates? Center for Open Science working paper.
- Wolfers, J. (2006), ‘Did unilateral divorce laws raise divorce rates? a reconciliation and new results’, *American Economic Review* **96**(5), 1802–1820.
- Wooldridge, J. (2021), Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators. Available at SSRN 3906345.
- Wooldridge, J. M. (2007), ‘Inverse probability weighted estimation for general missing data problems’, *Journal of Econometrics* **141**(2), 1281–1301.
- Wooldridge, J. M. (2023), ‘Simple approaches to nonlinear difference-in-differences with panel data’, *The Econometrics Journal* **26**(3), C31–C66.
- Xu, Y. (2017), ‘Generalized synthetic control method: Causal inference with interactive fixed effects models’, *Political Analysis* **25**(1), 57–76.
- Yitzhaki, S. (1996), ‘On using linear regressions in welfare economics’, *Journal of Business & Economic Statistics* **14**(4), 478–486.
- Zhang, S. and de Chaisemartin, C. (2020), ‘did_multiplegt: DID Estimation with Multiple Groups and Periods in R’. **URL:** <https://cran.r-project.org/web/packages/DIDmultiplegt/index.html>

Zhang, S. and de Chaisemartin, C. (2021), ‘TwoWayFEWeights: Estimation of the Weights Attached to the Two-Way Fixed Effects Regressions in R’.

URL: <https://cran.r-project.org/web/packages/TwoWayFEWeights/index.html>