The Econometrics Journal

Econometrics Journal (2024), volume **27**, pp. C1–C61.
https://doi.org/10.1093/ectj/utae014

Causal models for longitudinal and panel data: a survey

DMITRY ARKHANGELSKY[†] AND GUIDO IMBENS[‡]

[†]*CEMFI, Casado del Alsisal 5, 28014 Madrid, Spain.*
Email: darkhangel@cemfi.es

[‡]*Graduate School of Business and Department of Economics, Stanford University, 655 Knight Way, Stanford, CA 94305, USA.*
Email: imbens@stanford.edu

First version received: 3 August 2023; final version accepted: 22 April 2024.

Summary: In this survey we discuss the recent causal panel data literature. This recent literature has focused on credibly estimating causal effects of binary interventions in settings with longitudinal data, emphasising practical advice for empirical researchers. It pays particular attention to heterogeneity in the causal effects, often in situations where few units are treated and with particular structures on the assignment pattern. The literature has extended earlier work on difference-in-differences or two-way fixed effect estimators. It has more generally incorporated factor models or interactive fixed effects. It has also developed novel methods using synthetic control approaches.

Keywords: *Causal effects*, *difference in differences*, *factor models*, *panel data*, *synthetic control methods*, *two-way fixed effects*.

JEL codes: *C21*, *C22*, *C23*.

1. INTRODUCTION

In recent years, there has been a fast-growing and exciting body of research on new methods for estimating causal effects in panel or longitudinal data settings where we observe outcomes for a number of units repeatedly over time. This literature has taken some of the elements of the earlier panel data literature and combined them with insights from the causal inference literature. It has largely focused on the case with binary treatments, although the insights obtained in this body of work extend beyond that setting. Much of this work focuses on settings where traditionally difference-in-differences (DID) and two-way fixed effect (TWFE) methods (we largely use the two terms interchangeably, primarily using the TWFE acronym) have been popular among empirical researchers. In this survey, we review some of the methodological research and make connections to various other parts of the panel data literature.

Although we intend to make this survey of interest to empirical researchers, it is not primarily a guide with recommendations for specific cases. Rather, we intend to lay out our views on this literature in order that practitioners can decide which of the methods they wish to use in particular settings. In line with most of the literature, we see the models and assumptions used in this literature not as either holding exactly or not holding, but as approximations that may be useful in particular settings. For example, the TWFE setup has been criticised as making assumptions that are too strong. At some level, that is true almost by definition: parallel trends are unlikely to hold over any extended period of time for a large number of units. Similarly, assuming the absence of dynamic effects is unlikely to ever hold exactly, and treatment effects are surely heterogeneous. Nevertheless, in many cases, fixed effect models with time-invariant constant treatment effects may be effective baseline models. Understanding when those models are adequate, when relaxing them is likely to improve estimation, and what useful generalisations to relax their underlying assumptions are, is what we intend to do in this survey.

Ultimately, and this is perhaps our strongest statement on the relative merits of the various methods, we recommend against the current routine use of the standard TWFE estimator or related estimators. These methods have been very popular in empirical work and, in fact, continue to increase in popularity. See Goldsmith-Pinkham (2024) for a discussion tracing trends in their usage in recent years. Nevertheless, there are now many methods that generalise this estimator that we view as more attractive in practice. Some of these are based on strictly more general models for the potential outcomes (in particular factor models). Others use local versions of the TWFE estimator through weights on the cross-sectional and/or time dimension (building on the synthetic control literature). Both approaches often use regularisation to ensure good performance by avoiding overfitting even when the simpler methods (e.g., based on the standard TWFE model) are adequate. Although both approaches, more general outcome models and weights, are in our view superior to the standard TWFE methods, their relative performance varies by context and the relative merits are the subject of ongoing research. We therefore offer no specific recommendations for any one particular method. In addition, the more general methods still share some of the unattractive features of the TWFE estimator. In particular, they often pay little explicit attention to dynamics and time-series structure in potential outcomes. In addition many of the methods pay limited attention to the assignment mechanism. Developing methods that address these concerns is a promising and practically relevant area of future research.

A second recommendation concerns some of the specific issues raised in the recent TWFE literature. This literature has generated valuable new insights into the complications raised by the presence of heterogeneity in treatment effects. These insights have improved our understanding of the challenges with panel data. The new estimators developed in this literature do not, however, in our view, fully address all the practical challenges. On the positive side, they allow for much more heterogeneity in treatment effects than the earlier panel literature. On the negative side, like the earlier TWFE literature, the proposed estimators rely on unrealistically strong additivity and linearity assumptions on the potential outcomes, limiting the credibility of the estimates of counterfactuals. In our view there needs to be more balance in the richness of the models for the control potential outcomes and the richness of the models for the treatment effects. Putting no structure on the heterogeneity of the treatment effects is at odds with the typical goal of predicting the effect of implementations of the new policies to other locations, populations, or time periods rather than simply evaluating those policies on the currently exposed populations.

The paper is organised as follows. After the introduction, we first discuss in Section 2 some of the earlier econometric panel data literature. This serves both to set the stage for the framing of the questions of the current literature as well as to clarify differences in emphasis between the traditional and new literature. We also point out that some important conceptual issues that had been raised in the earlier literature have received less attention recently and that some are even in danger of being entirely ignored in the current literature.

Next, in Section 3, we discuss three ways of organising the panel data literature. First, we consider a classification by types of data available, e.g., proper panel data, repeated cross-sections, or row and column exchangeable data. (The latter refers to a matrix of data where both rows and columns are exchangeable, similar to panel data without any time-series structure.)

Second, we discuss an organisation by shapes of the data frame, e.g., many units or many periods. Finally, we discuss a classification based on the assignment for the causal variable of interest, e.g., block assignment, single treated unit, single treated period, or staggered adoption. We find these classifications useful because they matter for the relevance of various methods that have been proposed and they help organise them. Although the earlier econometric panel data literature also stressed the importance of the relative magnitude of the time and unit dimension as we do in our second classification, the realisation that the structure of the assignment process is important is a more recent insight. Many of the recent papers focus on particular parts of the general space of panel data inferential problems. For example, the vast literature assuming unconfoundedness in panel data settings has focused largely on the setting with a large number of units and relatively few time periods, and a subset of the units treated in the last period. In contrast, the synthetic control (SC) literature has primarily focused on the setting where the cross-section and time-series dimension are comparable in size, and where one or few units are treated from some period onwards. The recent DID/TWFE literature has paid particular attention to the setting with staggered adoption patterns in the assignment. The singular focus of some of the literature has helped in advancing them more rapidly, but, occasionally, insights from related settings have been overlooked.

In Section 4 we introduce some of the notation and estimands. We use, as in much of the causal inference literature, the potential outcome notation that makes explicit the causal nature of the questions.

In Section 5 we introduce the standard DID/TWFE setup as a stepping stone to the discussion of the recent developments in causal panel data literature. We see four main threads in the new causal panel literature, which we discuss in Sections 6 through 10.

First, in Section 6, we discuss the staggered adoption case. Much of the earlier TWFE literature concentrated on the case with a common adoption date. In contrast, one strand of the recent literature has focused on the setup where different groups adopt treatments at different points in time. In this staggered adoption case, recent research has highlighted some specific concerns with the standard TWFE estimator. In particular, in cases with general treatment effect heterogeneity, the implicit negative weights on the building blocks of the TWFE estimator have been argued to be unattractive, and alternatives have been proposed. We argue that these concerns have perhaps been exaggerated.

Second, as discussed in Section 7, the recent literature has generalised the popular TWFE structure to factor models. An important part of this literature is the SC approach developed in a series of influential papers by Alberto Abadie and co-authors (Abadie and Gardeazabal, 2003; Abadie et al., 2010). Although this literature shares key features with the TWFE literature, it has largely developed separately, ignoring some of the gains that can arise from combining the insights from each of them.

In the third strand, we consider in Section 8 a different class of generalisations of the TWFE setup, allowing for nonlinear models.

Fourth, as discussed in Section 9, the modern causal panel literature has sometimes taken a design-based approach to inference where the focus is on uncertainty arising from the assignment mechanism rather than a model-based or sampling-based perspective that is common in the earlier literature.

In Section 10, we discuss open questions in the causal panel data literature which we view as exciting avenues for future research.

Finally, in Section 11, we discuss some recommendations for empirical practice.

Table 1. Acronyms.

DID	Difference in differences
TWFE	Two way fixed effect
SC	Synthetic control
GRCS	Grouped repeated cross-section
RCED	Row column exchangeable data
SDID	Synthetic difference in differences
CIC	Changes in changes
NNMC	Nuclear norm matrix completion

There are some excellent recent discussions of the new DID/TWFE and causal panel data literature that are complementary to this survey. They differ in their focus and in the perspectives of the authors and complement ours in various ways. Some of these surveys (De Chaisemartin and d'Haultfoeuille, 2023; Roth et al., 2023) focus more narrowly on the DID/TWFE setting with heterogeneous treatment effects. They do not stress the connections with the synthetic control methods and factor models that we view as an important feature of the current panel data literature. In contrast, Abadie (2021) focuses primarily on synthetic control methods. In the current survey, we stress deeper linkages between these ideas and the TWFE literature as well as the potential benefits of combining them. In recent surveys in the political science literature and more in line with the current survey, Xu (2023) and Liu et al. (2024) also discuss the connections between synthetic control, unconfoundedness, and TWFE approaches.

In this discussion, we use a number of acronyms. For reference, we list those that we use regularly in Table 1.

2. THE ECONOMETRICS PANEL DATA LITERATURE

Although the new panel literature ostensibly focuses on different estimands and settings and emphasises different concerns about internal and external validity, many of the methods are closely related to those discussed in the earlier econometric panel data literature. Here we discuss at a high level some of the key insights from the earlier literature, in so far as they relate to the current literature, and some marked differences between the two. We come back to some of the specific areas of overlap in later sections. We do not attempt to review the earlier econometric literature, partly because that is a vast literature in itself, but mainly because there are many excellent surveys and textbooks, including Arellano and Honoré (2001), Arellano (2003), Baltagi (2008), Wooldridge (2010), Arellano and Bonhomme (2011b), and Hsiao (2022).

First of all, by the econometric panel data literature, we mean primarily the literature from the 1980s to the early 2000s, as, for example, reviewed in the surveys and textbooks, including Chamberlain (1982; 1984), Arellano and Honoré (2001), Arellano (2003), Baltagi (2008), Wooldridge (2010), Arellano and Bonhomme (2011b), and Hsiao (2022). This literature was initially motivated by the increased availability of various large public longitudinal data sets starting in the 1960s. These data sets included the Panel Study of Income Dynamics, the National Longitudinal Survey of Youth, which are proper panels where individuals are followed over fairly long periods of time, and the Current Population Survey, which, although primarily a repeated cross-section data set, has some short-term panel features, and at the state level can be viewed as a longer panel data set. These data sets vary substantially in the length of the time-series component, motivating different methods that could account for such data configurations.

The primary focus of the econometric literature has been on estimating invariant or structural parameters in the sense of Goldberger (1991). Part of the literature analysed fully parametric models, but more often semiparametric settings were considered. The parameters of interest could be causal in the modern sense, but the term itself would rarely be used explicitly. A major concern in this literature has been the presence of time-invariant unit-specific components. The literature distinguished between two types of such components: first, the so-called fixed effects and random effects. Fixed effects were conditioned on in the analyses and were modelled as unrestricted in their correlation with other variables. Random effects were treated as stochastic and often assumed to be uncorrelated with observed covariates (though not always; see the correlated random effects discussion in Chamberlain, 1984).[1] See for a general discussions Bell and Jones (2015) and Hsiao (2022). This distinction between fixed and random effects was often used as an organising principle for the panel data literature, in combination with the reliance on fixed T *versus* large T asymptotic approximations. A substantial literature was devoted to identification and inference results in settings with fixed effects leading to various forms of what Neyman and Scott (1948) labelled the incidental parameter problem. Especially when the fixed effects entered in nonadditive and nonlinear ways in short (with asymptotic approximations based on fixed length) panels, with limited dependent or discrete outcomes, this led to challenging identification problems, e.g., Chamberlain (1980), Honoré (1992), Magnac (2004), and Bonhomme (2012). In cases where identification in fixed length settings was not feasible, the literature introduced various methods for bias-correction (see Arellano and Hahn, 2007, for a survey) or developed bounds analyses (e.g., Honoré and Tamer, 2006). More recently, these bias-reduction ideas have been extended to nonlinear two-way models (e.g., Fernández-Val and Weidner, 2016; 2018).

The earlier econometric panel data literature paid close attention to the dynamics in the outcome process, arising from substantive questions such as the estimation of structural models for production functions and dynamic labour supply. Motivated by these questions, this literature distinguished between state dependence and unobserved heterogeneity (e.g., Heckman, 1981; Chamberlain, 1984) and various dynamic forms of exogeneity (e.g., weak, strong and strict exogeneity, and predeterminedness; see Engle et al., 1983; Arellano and Bond, 1991). These issues have not received as much attention yet in the current literature. The earlier literature also studied models that combined the presence of unit fixed effects with lagged dependent variables, leading to concerns about biases of least squares estimators in short panels—the so-called Nickell bias (Nickell, 1981) and the use of instrumental variable approaches (Nickell, 1981; Arellano and Bond, 1991; Blundell and Bond, 1998; Hahn and Kuersteiner, 2002; Alvarez and Arellano, 2003). This literature had a huge impact on empirical work in social sciences, but the recent literature has not connected much to these issues.

In contrast, an important theme in the current literature that was not discussed as much in the earlier literature concerns the presence of general heterogeneity in causal effects, both over time and across units, associated with observed as well as unobserved characteristics. The recognition of the importance of heterogeneity has led to findings that previously popular estimators are sensitive to the presence of such heterogeneity and to the development of more robust alternatives. These results are related to a subset of the econometric panel data literature, e.g., Chamberlain (1992), Arellano and Bonhomme (2011a), Graham and Powell (2012), and Chernozhukov et al.

[1] The terms fixed effects and random effects are not ideal and have led to some confusion, but they are by now so widely used that we use them as well.

(2013), which modelled heterogeneity in a way that is more in line with the current literature. We discuss this connection in detail in Section 6.

3. SETUP AND DATA CONFIGURATIONS

In this section, we consider three classifications of the literature. The first is based on different types of data. The second, in terms of the relative size of the cross-section and time-series dimensions, is familiar from the earlier literature. The third, in terms of the assignment mechanism, is original to the current literature. In the earlier literature, there was an additional classification that made a distinction that depended on the heterogeneity between cross-section units being modelled as fixed effects or random effects, e.g., Chamberlain (1984). This distinction plays less of a role in the current literature, although it is relevant for the design-based literature that we discuss in Section 9. All three classifications are helpful in understanding which specific methods may be useful and what type of asymptotic approximations for inference are credible. In addition, they allow us to place the individual papers, which often focus on particular settings, in context.

To put the following discussions into context, it is also helpful to remember that most of the recent literature has focused on the average causal effect of some intervention on the outcomes for the treated units during the periods they were treated. We do so here too, but one should keep in mind that one might be interested in an average effect beyond the study sample, or in an effect over time periods beyond the sample period. Later, we are more precise about the exact estimands we focus on and, in particular, how some of the assumptions, such as the absence of dynamic effects, affect both the choice of estimand and its interpretation.

3.1. Data Types

Although we focus in this paper mostly on the proper panel data setting where we observe outcomes for a number of units over a number of time periods, we also consider some other settings with observations at different points in time that we collectively refer to as panel data. Here we want to clarify the distinction and be precise about the notation.

3.1.1. Panel data. In the proper panel data case we have observations on N units, indexed by $i = 1, \ldots, N$, over T periods, indexed by $t = 1, \ldots, T$. The outcome of interest is denoted by Y_{it}, and the treatment is denoted by W_{it}, both doubly indexed by the unit and time indices. These observations may themselves consist of averages over more basic units as in the grouped repeated cross-section case from Section 3.1.2. We collect the outcomes and treatment assignments into two $N \times T$ matrices:

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & Y_{12} & Y_{13} & \ldots & Y_{1T} \\ Y_{21} & Y_{22} & Y_{23} & \ldots & Y_{2T} \\ Y_{31} & Y_{32} & Y_{33} & \ldots & Y_{3T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & Y_{N2} & Y_{N3} & \ldots & Y_{NT} \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} W_{11} & W_{12} & W_{13} & \ldots & W_{1T} \\ W_{21} & W_{22} & W_{23} & \ldots & W_{2T} \\ W_{31} & W_{32} & W_{33} & \ldots & W_{3T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ W_{N1} & W_{N2} & W_{N3} & \ldots & W_{NT} \end{pmatrix},$$

with the rows corresponding to units and the columns corresponding to time periods.

We may also observe other exogenous variables, denoted by X_{it} or X_i, depending on whether they vary over time or only by unit. Typically, we focus on a balanced panel where for all units $i = 1, \ldots, N$ we observe outcomes for all $t = 1, \ldots, T$ periods. In practice, concerns can arise from the panel being unbalanced either because we observe units for different lengths of time or because data is missing for some of them. We ignore both complications in the current discussion.

Classic examples of this proper panel setting include Ashenfelter (1978) with information on earnings for over 90,000 individuals for 11 years, and Abowd and Card (1989) with information on wages for 1,448 individuals also for 11 years. Another classic example is Card and Krueger (1994) with data for two periods and 399 fast-food restaurants.

3.1.2. Grouped repeated cross-section data. In a grouped repeated cross-section (GRCS) data setting, we have observations on N units. Each unit is observed only once, in period T_i for unit i, with the time period indexed by i to account for the fact that different units may be observed at different points in time. Typically T_i takes on only a few values (the repeated cross-sections) relative to the number of units, e.g., often just two or three, with many units sharing the same value for T_i. For some of the methods this is formally not required. The outcome and treatment received for unit i are denoted by Y_i and W_i respectively, both indexed just by the unit index i.[2] The set of units is partitioned into two or more groups, with the group that unit i belongs to denoted by $G_i \in \mathcal{G} = \{1, 2, \ldots, G\}$.

Define the average outcome for each group/time-period pair:

$$\overline{Y}_{gt} \equiv \sum_{i=1}^{N} \mathbf{1}_{G_i=g,T_i=t} Y_i \bigg/ \sum_{i=1}^{N} \mathbf{1}_{G_i=g,T_i=t},$$

and similar for $\overline{W}_{gt}$. If we view the $G \times T$ group averages $\overline{Y}_{gt}$, instead of the original Y_i, as the unit of observation, this grouped repeated cross-section setting is just like a panel as in Section 3.1.1, immediately allowing for methods that require repeated observations on the same unit. This was pointed out in Deaton (1985) and Wooldridge (2010). Many methods in the GRCS literature do not use the data beyond the group/time averages, and so the formal distinction between the grouped repeated cross-section and proper panel case becomes moot. However, in practice, empirical applications with grouped repeated cross-section data have typically many fewer groups than proper panel data have units, sometimes as few as two or three, limiting the scope for high-dimensional parametric models and raising concerns about the applicability of large N asymptotics.

In a seminal application of DID estimation with repeated cross-section data, with two groups and two periods (Meyer et al., 1995), the units are individuals getting injured on the job, and we observe individuals getting injured at most once. The time periods correspond to the year the individuals are injured, with data available for two years. Similarly, in Eissa and Liebman (1996) the units are different taxpayers in two different years, with the number of groups again equal to two. The case with more than two groups is studied in Bertrand et al. (2004) and in countless other studies, often with the groups corresponding to states, and the treatment regulations implemented at the state level.

3.1.3. Row and column exchangeable data. One data type that has not received as much attention as either panel or repeated cross-section data corresponds to what we refer to as row–column exchangeable data (RCED) (Aldous, 1981; Lynch, 1984). Like proper panel data, these data are doubly indexed, with outcomes denoted by Y_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, J$. The difference with

[2] Some empirical studies continue to use the panel notation that includes two indices for the outcomes and treatments in the repeated cross-section case, but that is confusing because Y_{it} and $Y_{it'}$ do not refer to the same unit i in the repeated cross-section case.

panel data is that there is no ordering for the second index (time in the proper panel case). An example of such a data type is supermarket shopping data, where we observe expenditures on item j for shopper i, or data from a ride-share company, where we observe outcomes for trips involving customer i and driver j, or a customer/product setting for an online retailer (Abadie et al., 2024). Although this is not a particularly common data configuration, it is useful to contrast it explicitly with proper panel and cross-section data. Proper panel data differs in two aspects from cross-section data: the double indexing and the time ordering: the RCED setting is in between the cross-section and proper panel case, with the double indexing, but no time ordering.

In this case, where the second index is not time, it is natural to model both units $i = 1, \dots, N$ and $j = 1, \dots, J$ as exchangeable, whereas with proper panel data, the exchangeability of the time periods is typically implausible. It is interesting to note that many, but not all, methods ostensibly developed for use with panel data are also applicable in this RCED setting. For example, TWFE methods, factor models, and many SC estimators, all discussed in more detail below, can be used with such data. The fact that those methods can be used in the RCED setting directly means that such estimators do not place any value on knowledge of the time-series ordering of the data. If *ex ante* one believes such information is valuable, one may wish to use methods that exploit it.

A related but even more general data type involves RCED with repeated observations. An example of such a data frame is a panel of matched employer–employee data (e.g., Abowd et al., 1999; Card et al., 2022). See Bonhomme (2020) for a recent survey of the relevant methods.

3.2. *Shapes of Data Frames*

Our second classification of the panel data literature is organised by the shape of the data frame. This is not an exact classification, and which category a particular data set fits and which methods are appropriate in part depend on the magnitude of the cross-section and time-series correlations and not just on the magnitude of N and T. Nevertheless, it is useful to reflect on the relative magnitude of the cross-section and time-series dimensions as it has implications for the properties of statistical methods for the analysis of such data. In particular, it often motivates the choice of asymptotic approximations based on large N and fixed T, or large N and large T.

3.2.1. Thin data frames: many units, few time periods ($N \gg T$). Much of the traditional panel data case considers the setting where the number of cross-section units is large relative to the number of time periods:

$$\underset{(N \gg T)}{\mathbf{Y}^{\text{thin}} =} \begin{pmatrix} Y_{11} & Y_{12} & Y_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ Y_{31} & Y_{32} & Y_{33} \\ Y_{41} & Y_{42} & Y_{43} \\ Y_{51} & Y_{52} & Y_{53} \\ Y_{61} & Y_{62} & Y_{63} \\ \vdots & \vdots & \vdots \\ Y_{N1} & Y_{N2} & Y_{N3} \end{pmatrix}.$$

This is a common setting when the units are individuals and it is challenging or expensive to get repeated observations for many periods for the same individual. The Panel Study of Income Dynamics (PSID) and National Longitudinal Surveys (NLS) panel data fit this setting, with often thousands of units. In this case inferential methods often rely on asymptotic approximations based on large N for fixed T. Incidental parameter problems of the type considered by Neyman and Scott (1948) are particularly relevant (see Lancaster, 2000, for a modern discussion). Specifically, if there are unit-specific parameters, e.g., fixed effects, it is not possible to estimate those parameters consistently. This does not necessarily imply that one cannot estimate the target parameters consistently, and the traditional literature developed many procedures that allowed for the elimination of these fixed effects, even if they enter nonlinearly, e.g., Honoré (1992), Chamberlain (2010), and Bonhomme (2012). However, the fact that the time-series dimension is small or modest does mean that random effect assumptions are potentially powerful because they place a stochastic structure on the individual components so that these individual components can be integrated out.

3.2.2. Fat data frames: few units, many time periods ($N \ll T$). The second setting is one where the number of time periods is large relative to the number of cross-section units:

$$\underset{(N \ll T)}{\mathbf{Y}^{\text{fat}}} = \begin{pmatrix} Y_{11} & Y_{12} & Y_{13} & Y_{14} & Y_{15} & Y_{16} & Y_{17} & Y_{18} & \dots & Y_{1T} \\ Y_{21} & Y_{22} & Y_{23} & Y_{24} & Y_{25} & Y_{26} & Y_{27} & Y_{28} & \dots & Y_{2T} \\ Y_{31} & Y_{32} & Y_{33} & Y_{34} & Y_{35} & Y_{36} & Y_{37} & Y_{38} & \dots & Y_{3T} \\ Y_{41} & Y_{42} & Y_{43} & Y_{44} & Y_{45} & Y_{46} & Y_{47} & Y_{48} & \dots & Y_{4T} \end{pmatrix}.$$

This setting is more common when the cross-section units are aggregates, e.g., states or countries, for which we have observations over many time periods, say output measures for quarters, or unemployment rates per month.

This setting is closely related to the traditional time-series literature, but the insights from that literature have not always been fully appreciated in the modern causal panel literature. There are some exceptions that take more of a time-series approach to this type of panel data, e.g., Brodersen et al. (2015) and Ben-Michael et al. (2023). The work on inference using conformal methods is also in this spirit, e.g., Chernozhukov et al. (2021).

3.2.3. Square data frames: comparable number of units and time periods: ($N \approx T$). In the third case the number of time periods and cross-section units is roughly comparable:

$$\underset{(N \approx T)}{\mathbf{Y}^{\text{square}}} = \begin{pmatrix} Y_{11} & Y_{12} & Y_{13} & \dots & Y_{1T} \\ Y_{21} & Y_{22} & Y_{23} & \dots & Y_{2T} \\ Y_{31} & Y_{32} & Y_{33} & \dots & Y_{3T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & Y_{N2} & Y_{N3} & \dots & Y_{NT} \end{pmatrix}.$$

A common example is that where the units are states and the time periods are years or quarters. We may have observations on 50 states for 30 years or for 80 quarters. This is a particularly challenging case and, at the same time, increasingly common in practice. Many empirical studies using DID/TWFE, SC, or related estimators fit into this setting.

Whether in this case asymptotic approximations based on large N and fixed T, or large N and large T, or neither, are appropriate is not always obvious. Simply looking at the magnitudes of the time-series and cross-section dimension itself is not sufficient to make that determination because the appropriate approximations also depend on the magnitude of cross-section and time-series correlations. There is an important lesson in this regard in the weak instrument literature. In the influential Angrist–Krueger analysis of the returns to schooling, Angrist and Krueger (1991) report results based on over 300,000 units and 180 instruments. Because of the relative magnitude of the number of units and instruments, one might have expected that asymptotic approximations based on a fixed number of instruments and an increasing number of units would be appropriate. Nevertheless, it turned out that the Bekker asymptotic approximation, developed by Bekker (1994) and based on letting the number of instruments increase proportionally to the number of units, is substantially more accurate because of the weak correlation between the instruments and the endogenous regressor (years of education in the Angrist–Krueger study).

The earlier econometric panel data literature discusses the trade-offs between various asymptotic approximations for the analysis of dynamic linear models, e.g., see Hahn and Kuersteiner (2002) and Alvarez and Arellano (2003). In dynamic models the fixed effect estimator is inconsistent in short panels. Alternative estimators have been proposed using lagged outcomes as instruments (Arellano and Bond, 1991; Blundell and Bond, 1998). As the panel becomes longer, the number of instruments grows. However, the more distant lags are often only weakly correlated with the endogenous regressors, leading to many weak instruments problems. One important aspect of the panel data analysis is that the fixed effect estimator is consistent in the large T limit, but not in the fixed T setting.

3.3. *Assignment Mechanisms*

The third classification for panel data methods we consider is based on features of the assignment process for the treatment. As in the classification based on the relative magnitudes of the components of the data frame, features of the assignment process are important for determining which statistical methods and which asymptotic approximations are reasonable. This classification is not present in the earlier panel data literature, but features prominently in the current literature. This reflects the more explicit focus on causal effects in general in the econometric literature of the last three decades. It should be noted that this classifications is not so much based on assumptions such as endogeneity or exogeneity, as it is about facts, regarding the assignment process.

One feature that is common to many applications of the methods is that the fraction treated unit/time-period pairs is small. This has two implications. First, the focus is typically on the average effect on the treated unit/time-period pairs. This may be for substantive reasons, but it is also motivated by the fact that if the fraction of treated pairs is small, the precision of estimates for the overall average effect will be considerably lower than the precision of estimates for the average effect for the treated pairs. Second, building statistical models for the treated outcomes will be of low value, as such models will not increase the precision of standard estimators. Thus, important modelling questions are about the control outcomes.

3.3.1. The general case. In the most general case the treatment may vary both across units and over time, with units switching in and out of the treatment group:

$$\underset{\text{(general)}}{\mathbf{W}^{\text{gen}}} = \begin{pmatrix} 1 & 1 & 0 & 0 & \dots & 1 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & 1 & \dots & 0 \\ 1 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 1 & 0 & \dots & 0 \end{pmatrix}. \tag{3.1}$$

With this type of data, we can use variation of the treatment within units and variation of the treatment within time periods to identify causal effects. Especially in settings without dynamic effects, the presence of both types of variation may improve the credibility of estimators for causal effects. This setting is particularly relevant for the RCED configurations, but it is less common in proper panel data settings. Some examples include marketing settings with the units corresponding to products and the treatment corresponding to promotions or discounts.

In this setting, assumptions about the absence or presence of dynamic treatment effects are particularly important. In applications where dynamic treatment effects are present, many commonly used methods assuming their absence lead to results that are difficult to interpret.

3.3.2. Single treated period. One important special case arises when a substantial number of units is treated, but these units are only treated in the last period.

$$\underset{\text{(last period)}}{\mathbf{W}^{\text{last}}} = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

In settings where the number of time periods is relatively small, this case is often analysed as a cross-section problem. The lagged outcomes are simply used as exogenous covariates or pre-treatment variables that should be adjusted for in treatment-control comparisons based on an unconfoundedness assumption (Rosenbaum and Rubin, 1983). A classic example in the economics literature is the LaLonde–Dehejia–Wabha data originally collected in LaLonde (1986) with the data set now commonly used constructed and analysed in Dehejia and Wahba (1999). This data set has served as a valuable playground for assessing new methodological advances in the literature on unconfoundedness. In that case, there are three periods of outcome data (earnings), but only one post-treatment outcome. The original study of LaLonde (1986) reported results for a variety of models, including some two-way fixed effect regressions. Much of the subsequent literature since Dehejia and Wahba (1999; 2002) has focused more narrowly on methods relying on unconfoundedness, sometimes in combination with functional form assumptions. See Imbens and Xu (2024) for a recent discussion. Asymptotics are typically, and appropriately so, based on large N and fixed T.

Given that the treatment is observed only in the last period, the presence of dynamic effects is not testable, and dynamics do not really matter in the sense that their presence only leads to a minor change in the interpretation of the estimand, typically the average effect for the treated units and time periods. Because of the shortness of the panel, these are obviously short-term effects, with little evidence regarding the long-term impacts of the interventions.

3.3.3. Single treated unit. Another key setting is that with a single treated unit, treated in multiple periods.

$$\underset{\text{(single unit)}}{\mathbf{W}^{\text{single}}} = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & 1 & \dots & 1 \end{pmatrix}.$$

This setting is prominent in the original applications of the synthetic control literature: Abadie and Gardeazabal (2003), Abadie et al. (2010), and Abadie (2021). This literature has exploded in terms of applications and theoretical work in the last twenty years. Here the number of time periods is typically too small to rely credibly on large T asymptotics, creating challenges for inference that are not entirely resolved. Large N asymptotics creates its own, different, challenges, stemming from the lack of multiple treated units.

3.3.4. Single treated unit and single treated period. An extreme case is that with only a single unit ever treated, and this unit only treated in a single period, typically the last period:

$$\underset{\text{(single unit/period)}}{\mathbf{W}^{\text{one}}} = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

This is a challenging setting for inference: we cannot rely on large sample approximations for the average outcomes for the treated unit/periods because there is only a single treated unit/period pair. Instead of focusing on population parameters it is natural here to focus on the effect for the single treated/time-period pair and construct prediction intervals. Because it is a special case of both the single treated period and the single treated unit case it is conceptually important for comparing estimation methods popular for those settings.

3.3.5. Block assignment. Another important case in practice is that with block assignment, where a subset of units is treated every period after a common starting date:

$$\underset{\text{(block)}}{\mathbf{W}^{\text{block}}} = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 1 & \dots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & \dots & 1 & 1 \end{pmatrix}.$$

This assignment matrix is the basis of the simulations reported in Bertrand et al. (2004) and Arkhangelsky et al. (2021). In this case there is typically a sufficient number of treated unit/time-period pairs to allow for reasonable approximations to be based on that number being large.

Here the presence of dynamic effects changes the interpretation of the average effect for the treated. The average effect for the treated now becomes an average over short and medium term effects during different periods. There is limited ability to separate out heterogeneity across calendar time and dynamic effects because, in any given time period, there are only treated units with an equal number of treated periods in their past.

3.3.6. Staggered adoption. The recent DID/TWFE literature has focused on the staggered adoption case where units remain in the treatment group once they adopt the treatment, but they vary in the time at which they adopt the treatment. Some may adopt early, while others adopt later:

$$\underset{\text{(staggered adoption)}}{\mathbf{W}^{\text{stag}} =} \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & 0 & 1 & \dots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & 1 & 1 & \dots & 1 & 1 \end{pmatrix}.$$

This case is also referred to as the absorbing treatment setting. Clearly, this setting leads to much richer information about the possible presence of dynamic effects, with the ability, under some assumptions, to separate dynamic effects from heterogeneity across calendar time.

A second issue is whether, for units adopting in period t, the best controls are units adopting in period $t + 1$, or later, or possibly the units never adopting the treatment (Callaway and Sant'Anna, 2021).

3.3.7. Event-study designs. A closely related design is the event-study design, where units are exposed to the treatment in at most one period.

$$\underset{\text{(event study)}}{\mathbf{W}^{\text{event}} =} \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & \dots & 0 & 0 \end{pmatrix}.$$

In this setting there are often dynamic effects of the treatment past the time of initial treatment. If these effects are identical to the initial effect, the analysis can end up being very similar to that of staggered adoption designs. Canonical applications include some in finance, e.g., Fama et al. (1969).

3.3.8. Clustered assignment. Finally, in many applications, units are grouped together in clusters, with units within the same clusters always assigned to the same treatment. The example below has C clusters, with a subset of the clusters assigned to the treatment from a common period onwards in a block assignment structure.

$$\underset{\text{(cluster/block)}}{\mathbf{W}^{\text{cluster}}} = \left(\begin{array}{cccccccc} & & & & & & & \text{cluster} \\ & & & & & & & \downarrow \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & \dots & 1 & 1 & 3 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & \dots & 1 & 1 & C \\ 0 & 0 & 0 & 1 & \dots & 1 & 1 & C \end{array}\right).$$

The clustering creates particular complications for inference, whether it is in the block assignment case, or other settings, in particular because often there are relatively few clusters. It also creates challenges for estimation if there are cluster components to the outcomes.

4. POTENTIAL OUTCOMES, GENERAL ASSUMPTIONS, AND ESTIMANDS

In this section we collect in a single section the notation that allows us to cover various parts of the literature. We focus on the proper panel data case with N units and T periods. We use the potential outcome notation (see Rubin, 1974; Imbens and Rubin, 2015). We also discuss basic estimands that have been the focus of this literature and some of the maintained assumptions.

Let $\underline{\mathbf{w}}$ denote the full T-component column vector of treatment assignments,

$$\underline{\mathbf{w}} \equiv (w_1, \dots, w_T)^\top,$$

and $\underline{\mathbf{W}}_i$ the vector of treatment values for unit i. Let $\underline{\mathbf{w}}^t$ the t-component column vector of treatment assignments up to time t:

$$\underline{\mathbf{w}}^t \equiv (w_1, \dots, w_t)^\top,$$

so that $\underline{\mathbf{w}}^T = \underline{\mathbf{w}}$, and similar for $\underline{\mathbf{W}}_i^t$. In general we can index the potential outcomes for unit i in period t by the full T-component vector of assignments $\underline{\mathbf{w}}$:

$$Y_{it}(\underline{\mathbf{w}}).$$

Even this notation already makes a key assumption, the stable unit treatment value assumption, or SUTVA (see Rubin, 1978; Imbens and Rubin, 2015). SUTVA requires that there is no interference or spillovers between units. This is a strong assumption, and in many applications it may be violated. There has been little attention paid to models allowing for such interference in the recent causal panel data literature to date, although there is extensive literature on interference in cross-section settings, e.g., in clustering settings (Manski, 1993; Hudgens and Halloran, 2008), in network settings (Auerbach, 2022; Leung, 2023), and in the general case (Aronow and Samii, 2017).

In applications where the spillover effects are only present within certain groups, e.g., clusters or economic markets, and the treatment is assigned at the same level, one can justify SUTVA by aggregating the individual data to the cluster or group level. In this case, the potential outcome introduced above would correspond to the aggregated data. This is directly connected to our discussion of grouped repeated cross-section (GRCS) data in Section 3.1.2. Of course, the aggregation changes the interpretation of the causal effects, which would now incorporate both direct and spillover effects.

Without further restrictions, our setup describes for each unit and each time period 2^T potential outcomes, as a function of multi-valued treatment $\underline{\mathbf{w}}$. As a result we can define for every period t unit-level treatment effects for every pair of assignment vectors $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}'$:

$$\tau_{it}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'} \equiv Y_{it}(\underline{\mathbf{w}}') - Y_{it}(\underline{\mathbf{w}}),$$

with the corresponding population average effect defined as

$$\tau_{t}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'} \equiv \mathbb{E}\left[Y_{it}(\underline{\mathbf{w}}') - Y_{it}(\underline{\mathbf{w}})\right].$$

These unit-level and average causal effects are the basic building blocks of many of the estimands considered in the literature. Note that we implicitly assume there is a large population over which we can take the expectation. Part of the literature has focused on finite sample issues using a design perspective. See Section 9 for more discussion on this.

If we are only interested in average causal effects of the form $\tau_{t}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'}$, then we have, in essence, a problem similar to the cross-sectional version of the problem of estimating average causal effects. One approach would be to analyse such problems using standard methods for multi-valued treatments under unconfoundedness, e.g., Imbens (2000). Here this would require comparing outcomes in period t for units with treatment vectors $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}'$.

If we have completely random assignment, all average causal efects of the type $\tau_{t}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'}$ are identified, given sufficient variation in the treatment paths. That also means that we can identify in this setting dynamic treatment effects. For example, in the two-period case

$$\tau_2^{(1,1),(0,1)},$$

is the average effect in the second period of being exposed to the sequence (1,1) rather than the sequence (0,1), so it measures the dynamic effect of a sequence of treatments on period 2 outcomes, being exposed to the treatment in both periods versus being exposed only in the second period.

A key challenge is that there are many, $2^{T-1} \times (2^T - 1)$ to be precise, distinct average effects of the form $\tau_{t}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'}$. Even with $T = 2$ there are already six different average causal effects, and with T larger, this number quickly increases. This means that in practice we need to limit or focus on summary measures of all these causal effects, e.g., averages over effects at different times. Typically we also put additional structure on these causal effects in the form of cross-temporal restrictions on the potential outcomes $Y_{it}(\underline{\mathbf{w}})$. That enables us to give comparisons of outcomes from different time periods a causal interpretation. See Chamberlain (1984) for a discussion of this in the case of linear models. Note that without additional restrictions, all the average treatment effects $\tau_{t}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'}$ are just-identified, so any additional assumptions will typically imply testable restrictions.

The first restriction that we consider is the commonly made no-anticipation assumption, see, e.g., Sun and Abraham (2020), Callaway and Sant'Anna (2021), and Athey and Imbens (2022). This requires that potential outcomes do not depend on future treatments.

ASSUMPTION 4.1. (NO ANTICIPATION) *The potential outcomes satisfy*

$$Y_{it}(\underline{\mathbf{w}}) = Y_{it}(\underline{\mathbf{w}}'),$$

for all i, and for all combinations of t, $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}'$ such that $\underline{\mathbf{w}}^t = \underline{\mathbf{w}}'^t$.

With experimental data, and sufficient variation in treatment paths, this assumption is testable. To do so we can compare outcomes in period t for units with the same treatment path up to and including t, but whose treatment paths diverge in the future, that is, after period t. The average difference between such average outcomes should be zero in expectation under the no-anticipation assumption.

This substantive assumption is appealing in situations where units are not active decision-makers, but rather passive recipients of the treatment. In such cases, the no-anticipation assumption can, in principle, be guaranteed by design. If random units are assigned treatment each period, or, in the staggered adoption case, if the adoption date is randomly assigned, potential outcomes cannot vary with the future random assignment. Of course, in observational studies, the assumption need not hold. In many applications, treatments are state-level regulations that are known to be coming prior to the time they formally take effect. One remedy for this problem is to allow for limited anticipation, assuming the treatment can be anticipated for a fixed number of periods, as in Callaway and Sant'Anna (2021). Algorithmically, this amounts to redefining $\underline{\mathbf{w}}$ by shifting it by the fixed number of periods.

At the same time, there are numerous economic applications where units are involved in an active decision-making process. Units can make decisions about variables for which $\underline{\mathbf{w}}$ is an important input, and beliefs about future treatment paths would affect those decisions. For example, current taxes and beliefs about future tax changes can be important determinants of current consumption. In this case, researchers first need to address a key conceptual issue. The premise of the potential outcome framework is that it describes the exhaustive set of counterfactual outcomes that can be realised in an experiment where the researcher controls the assignment of $\underline{\mathbf{w}}$. However, if the units are making decisions in environments with uncertainty, then they can change their behaviour in response to different distributions of the future treatment paths, in line with Lucas's (1976) critique. As a result, one cannot express potential outcomes as functions of $\underline{\mathbf{w}}$ only, but also needs to view them as functions of the experimental design itself, i.e., the known or anticipated distribution of $\underline{\mathbf{w}}$.

One solution to this problem is to define potential outcomes for a given randomised experimental design. Assumption 4.1 then becomes innocuous because the beliefs about the future treatment paths are incorporated in the definition of the potential outcomes, and the actual values are by construction unknown. This does, however, change the interpretation of the causal effects. This issue is well understood in macroeconomic literature, which emphasises the distinction between the effect of a surprise deviation from a given policy rule versus the effect of a permanent change in the policy rule itself. While the former quantity can be learned using various quasi-experimental strategies (see Nakamura and Steinsson, 2018, for a discussion), the identification of the latter typically relies on an economic model (see, however, McKay and Wolf, 2023). Causal panel data literature could benefit from explicitly incorporating these ideas. See Abbring and Heckman (2007) for a related discussion and additional references.

The situation becomes considerably more complicated in observational studies, where one cannot directly control the information about the future treatment paths available to the units. Nevertheless, in some applications, the researchers directly observe the arrival of such information. In this case, to make Assumption 4.1 plausible, one needs to guarantee that different units face the same informational environment. Failure to do so is akin to comparing outcomes across units participating in experiments with different designs. Abbring and Van den Berg (2003) and Abbring and Heckman (2007) show that many economic applications have data that would allow researchers to measure the information inflow and discuss how to adjust for the differences across units in this inflow to ensure that Assumption 4.1 holds.

The no-anticipation assumption reduces the total number of potential treatment effects from $2^{T-1} \times (2^T - 1)$ to $(\sum_{t=1}^{T} 2^{t-1})(\sum_{t=1}^{T} 2^t - 1)$. The basic building blocks, unit period-specific treatment effects, are now of the type

$$\tau_{it}^{\underline{\mathbf{w}}^t, \underline{\mathbf{w}}^{t'}} \equiv Y_{it}(\underline{\mathbf{w}}^{t'}) - Y_{it}(\underline{\mathbf{w}}^t),$$

with the potential outcomes for period t indexed by treatments up to period t only.

This current structure still allows us to distinguish between static treatment effects, i.e., $\tau_{it}^{(\underline{\mathbf{w}}^{t-1},0),(\underline{\mathbf{w}}^{t-1},1)}$, which measures the response of current outcome to the current treatment, holding the past ones fixed, and dynamic ones, i.e., $\tau_{it}^{(\underline{\mathbf{w}}^{t-1},w^t),(\underline{\mathbf{w}}'^{t-1},w^t)}$, which does the opposite. In the earlier panel data literature, the dynamic effects were explicitly modelled by putting a particular structure on them, but, in principle, one can identify them without imposing additional restrictions on the potential outcomes given assumptions on the assignment mechanism, such as random assignment, e.g., Bojinov et al. (2021). There is also a large literature in biostatistics on dynamic models that is relevant for these problems, e.g., Robins et al. (2000) and Murphy (2003).

A stronger assumption is that the potential outcomes only depend on the contemporaneous assignment, ruling out dynamic effects of any type.

ASSUMPTION 4.2. (NO-DYNAMIC/CARRY-OVER EFFECTS) *The potential outcomes satisfy*

$$Y_{it}(\underline{\mathbf{w}}) = Y_{it}(\underline{\mathbf{w}}'),$$

for all i and for all combinations of t, $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}'$ such that $w_{it} = w'_{it}$.

This is *not* a design assumption that can be guaranteed by randomisation in a suitably designed experiment. It restricts the treatment effects and, thus, the potential outcomes for the post-treatment periods. Like the no-anticipation assumption it has testable restrictions given the random assignment of the treatment and sufficient variation in the treatment paths. Note that it does *not* restrict the time path of the potential outcomes in the absence of any treatment, $Y_{it}(\mathbf{0})$, where $\mathbf{0}$ is the vector with all elements equal to zero. In fact, these outcomes can exhibit arbitrary correlations in the sequence of potential outcomes $Y_{it}(\underline{w})$ for any given $\underline{w}$.

If we are willing to make the no-dynamic effects assumption, we can write the potential outcomes, with some abuse of notation, as $Y_{it}(0)$ and $Y_{it}(1)$ with a scalar argument. In this case, the total number of treatment effects for each unit is greatly reduced to T (one per period), and we can simplify them to

$$\tau_{it} \equiv Y_{it}(1) - Y_{it}(0),$$

where τ_{it} has no superscripts because there are only two possible arguments of the potential outcomes, $w \in \{0, 1\}$.

So far, we have discussed assumptions on potential outcomes themselves. A conceptually different assumption is that of absorbing treatments, that is where the assignment mechanism corresponds to staggered adoption.

ASSUMPTION 4.3. (STAGGERED ADOPTION)

$$W_{it} \geq W_{it-1} \quad \forall t = 2, \ldots, T.$$

Defining the adoption date A_i as the date of the first treatment, $A_i \equiv T + 1 - \sum_{t=1}^{T} W_{it}$ for units that are treated in the sample, and $A_i \equiv \infty$ for never treated ones. In the staggered adoption case, we can write the potential outcomes, again with some abuse of notation, in terms of the adoption date, $Y_{it}(a)$, for $a = 1, \dots, T, \infty$, and the realised outcome as $Y_{it} = Y_{it}(A_i)$.

There are two cases that are sometimes viewed as staggered adoption designs, but that are different in substance although not always in terms of analyses. First, there may be interventions that are adopted and remain in place. States or other administrative units adopt new regulations at different times. For example, states adopted speed limits or minimum drinking ages at different times (Ashenfelter and Greenstone, 2004), and counties adopted enhanced 911 policies at different times (Athey and Stern, 2002). These staggered adoption designs were introduced in Section 3.3.6. Second, there may be one-time interventions that have a long-term or even permanent impact. We refer to such settings, introduced in Section 3.3.7 as event studies. In that case, the post-intervention period effects would be dynamic effects.

Given staggered adoption, but absent the no-anticipation and no-dynamics assumptions, we can write the building blocks as

$$\tau_{it}^{a,a'} \equiv Y_{it}(a') - Y_{it}(a),$$

with the corresponding population average

$$\tau_{t}^{a,a'} \equiv \mathbb{E}\left[Y_{it}(a') - Y_{it}(a)\right].$$

We also introduce notation for the average for subpopulations defined by the adoption date:

$$\tau_{t|a''}^{a,a'} \equiv \mathbb{E}\left[Y_{it}(a') - Y_{it}(a)|A_i = a''\right].$$

Compared to previously defined estimands, this one explicitly depends on the details of the assignment process, which determines which units adopt the treatment and when they do so. This estimand is conceptually similar to the average effect on the treated in cross-sectional settings, with the important difference that selection now operates over two dimensions: units and periods. As in the cross-sectional setting, this matters for interpretation in observational studies in which the researcher does not control the assignment process.

In the two-period case where all units are exposed to the control treatment in the initial treatment, the estimand $\tau_{t|1}^{0,1}$, the average effect of the treatment in the second period for those that adopt in the second period is very much like the average effect for the treated. In settings with more variation in the adoption date there are many such average effects, depending on when the units adopted, and which period we are measuring the effect in.

5. TWO-WAY FIXED EFFECT AND DIFFERENCE-IN-DIFFERENCES ESTIMATORS

In this section we give a brief introduction to conventional difference-in-differences (DID) or two-way fixed effect (TWFE) estimations. The discussion is framed in terms of the potential outcomes framework from the modern causal inference literature, but otherwise it is largely standard and following textbook discussions. For other recent surveys of this literature, see Chiu et al. (2023), De Chaisemartin and d'Haultfoeuille (2023), and Roth et al. (2023).

5.1. The Two-Way Fixed Effect Characterisation

We start with the two-way fixed effect (TWFE) specification in a proper panel setting with no anticipation and no dynamics and parallel trends or constant treatment effects. Traditionally this specification often motivates the DID estimator.

ASSUMPTION 5.1. (THE TWO-WAY FIXED EFFECT MODEL) *The control outcome $Y_{it}(0)$ satisfies a two-way fixed effect structure:*

$$Y_{it}(0) = \alpha_i + \beta_t + \varepsilon_{it}. \tag{5.1}$$

The unobserved component ε_{it} is (mean-)independent of the treatment assignment W_{it}.

ASSUMPTION 5.2. (PARALLEL TRENDS ASSUMPTION) *The potential outcomes satisfy*

$$Y_{it}(1) = Y_{it}(0) + \tau \quad \forall(i, t).$$

The combination of these two assumptions leads to a model for the realised outcome, defined as $Y_{it} \equiv W_{it}Y_{it}(1) + (1 - W_{it})Y_{it}(0)$,

$$Y_{it} = \alpha_i + \beta_t + \tau W_{it} + \varepsilon_{it}. \tag{5.2}$$

We can estimate the parameters of this model by least squares:

$$(\hat{\tau}^{\text{TWFE}}, \hat{\alpha}, \hat{\beta}) = \arg\min_{\tau,\alpha,\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \alpha_i - \beta_t - \tau W_{it})^2. \tag{5.3}$$

Here we need to impose one restriction on the α_i or β_t (e.g., fixing one of the α_i or one of the β_t equal to zero) to avoid perfect collinearity of the regressors, but this normalisation does not affect the value for the estimator of the parameter of interest, τ.

Under a block assignment structure we have $W_{it} = 1$ only for a subset of the units (the 'treatment group' with $i \in \mathcal{I}$, where the cardinality for the set $\mathcal{I}$ is N^{tr}, and $N^{\text{co}} \equiv N - N^{\text{tr}}$), and those units are treated only during periods t with $t > T_0$ ('post-treatment'). Defining the averages in the four groups as

$$\overline{Y}^{\text{tr,post}} \equiv \frac{\sum_{i\in\mathcal{I}} \sum_{t>T_0} Y_{it}}{N^{\text{tr}}(T - T_0)}, \quad \overline{Y}^{\text{tr,pre}} \equiv \frac{\sum_{i\in\mathcal{I}} \sum_{t\leq T_0} Y_{it}}{N^{\text{tr}} T_0},$$
$$\overline{Y}^{\text{co,post}} \equiv \frac{\sum_{i\notin\mathcal{I}} \sum_{t>T_0} Y_{it}}{N^{\text{co}}(T - T_0)}, \quad \text{and} \quad \overline{Y}^{\text{co,pre}} \equiv \frac{\sum_{i\notin\mathcal{I}} \sum_{t\leq T_0} Y_{it}}{N^{\text{co}} T_0},$$

we can write the estimator for the treatment effect as

$$\hat{\tau}^{\text{TWFE}} = \hat{\tau}^{\text{DID}} = \left(\overline{Y}^{\text{tr,post}} - \overline{Y}^{\text{tr,pre}}\right) - \left(\overline{Y}^{\text{co,post}} - \overline{Y}^{\text{co,pre}}\right),$$

in the familiar double difference form that motivated the DID terminology.

It is convenient to use the TWFE characterisation based on least squares estimation of the regression function in (5.2) because this characterisation also applies in settings where the estimator does not have the double difference form, including the staggered adoption setting and even more general assignment processes. For this reason we also generally use the TWFE rather than the DID terminology in the remainder of this discussion.

5.2. The Difference-In-Differences Estimator in the Grouped Repeated Cross-Section Setting

Here we study the grouped repeated cross-section case where we observe each physical unit only once, obviously implying that we observe different units at different points in time. We continue to focus on the case with the blocked assignment. To reflect this, our notation now only has a single index for the unit, $i = 1, \dots, N$. Let $G_i \in \mathcal{G} = \{1, \dots, G\}$ denote the cluster or group unit i belongs to, and $T_i \in \{1, \dots, T\}$ the time period unit i is observed in. The set of clusters $\mathcal{G}$ is partitioned into two groups, a control group $\mathcal{G}_C$ and a treatment group $\mathcal{G}_T$, with cardinality G_C and G_T respectively.

Units belonging to a group G_i with $G_i \in \mathcal{G}_C$ are not exposed to the treatment, irrespective of the time the units are observed. Units with $G_i \in \mathcal{G}_T$ are exposed to the treatment if and only if they are observed after the treatment date T_0, so that the treatment indicator is $W_i = \mathbf{1}_{G_i \in \mathcal{G}_T, T_i > T_0}$. Let $D_i = \mathbf{1}_{G_i \in \mathcal{G}_T}$ be the treatment group indicator that indicates whether unit i is in one of the treated groups, irrespective of whether this unit is observed in the post-treatment period, so that $W_i = D_i \mathbf{1}_{T_i > T_0}$.

To define the DID estimator we first average outcomes and treatments for all units within a cluster/time period and construct $\overline{Y}_{gt}$ and $\overline{W}_{gt}$. By assumption that the treatment within group and time-period pairs is constant, the cluster/time-period average treatment $\overline{W}_{gt}$ is binary if the original treatment is. The DID estimator is then the double difference

$$\hat{\tau}^{\text{DID}} = \frac{1}{G_T(T-T_0)} \sum_{g \in \mathcal{G}_T, t > T_0} \overline{Y}_{gt} - \frac{1}{G_C(T-T_0)} \sum_{g \in \mathcal{G}_C, t > T_0} \overline{Y}_{gt}$$
$$- \frac{1}{G_T T_0} \sum_{g \in \mathcal{G}_T, t \le T_0} \overline{Y}_{gt} + \frac{1}{G_C T_0} \sum_{g \in \mathcal{G}_C, t \le T_0} \overline{Y}_{gt}.$$

Alternatively, we can use the TWFE specification at the group level (it cannot be used at the unit level because we do not observe any unit multiple times). At the group level we do have a proper panel setup:

$$\overline{Y}_{gt}(0) = \alpha_g + \beta_t + \varepsilon_{gt}, \qquad \overline{Y}_{gt}(1) = \overline{Y}_{gt}(0) + \tau, \tag{5.4}$$

similar to that in (5.1). The potential outcomes $\overline{Y}_{gt}(0)$ and $\overline{Y}_{gt}(1)$ should here be interpreted as the average of the potential outcomes if all units in a group/time-period pair are exposed to the control (active) treatment. The group-level TWFE estimator is identical to the DID estimator.

5.3. Inference

To conduct inference about $\hat{\tau}^{\text{DID}}$ or $\hat{\tau}^{\text{TWFE}}$ we need to be explicit about the sampling and assignment schemes. In situations where the assignment process is known, such as in randomised experiments, we can do design-based or randomisation-based inference. We discuss this approach in detail in Section 9. Outside of such situations, researchers typically rely on sampling-based inference, which we outline below.

In the proper panel setting, one often assumes that all units are randomly sampled from a large population and thus exchangeable. In this case, inference about $\hat{\tau}^{\text{TWFE}}$ reduces to joint inference about four means with independent and identically distributed (i.i.d.) observations. This approach was used by Card and Krueger (1994) to quantify the uncertainty about the estimated effect of the minimum wage.

With GRCS data and the cardinality of the control and treatment groups $\mathcal{G}_C$ and $\mathcal{G}_T$ larger than one, the situation is different. Now in addition to accounting for variation within a group at the unit level, one can allow for nonvanishing errors at the group level, the ε_{gt} in the model in (5.4). This cannot be done in the two-group/two-period case as in Card and Krueger (1994) because one cannot estimate the between-group variation in the presence of group fixed effects, and mechanically the estimated residuals $\hat{\varepsilon}_{gt}$ are all equal to zero. The clustering approach allowing for nonvanishing ε_{gt} was advocated in Liang and Zeger (1986), Arellano (1987), Bertrand et al. (2004), Donald and Lang (2007), and Ibragimov and Müller (2016), and is routinely used in situations where the number of groups or periods exceeds two. See Abadie et al. (2023) for a recent discussion in a design setting.

In addition to accounting for the presence of nonvanishing ε_{gt}, since Bertrand et al. (2004) inference for TWFE estimators has typically taken into account the correlation in outcomes over time within units in applications with more than two periods. This implies that if one estimates the average treatment effect as in (5.3), it is not appropriate to use the robust, Eicker–Huber–White standard errors. Instead, one can use clustered standard errors (Liang and Zeger, 1986; Arellano, 1987), based on clustering observations by units. The appropriate standard errors can also be approximated by bootstrapping all observations for each unit. Hansen (2007) discusses a more general hierarchical setup.

5.4. The Parallel Trend Assumption

The fundamental justification for the TWFE estimator, in one form or another, is based on a parallel trend assumption. This states that, in one form or another, the units that are treated would have followed, in the absence of the treatment, a path that is parallel to the path followed by the control units, in an average sense. The substantive content and the exact form of the assumption depend on the specific setup, the proper panel case versus the grouped repeated cross-section case, whether one takes a model-based or design-based perspective, the number of groups, and the averaging that is performed.

Let us first consider the proper panel case, with block assignment and D_i the indicator for the event that unit i will be exposed to the treatment in the post-treatment period (after T_0). Then, the assumption is that the expected difference in control outcomes in any period for units that later are exposed to the treatment and units that are always in the control group is constant:

ASSUMPTION 5.3. *For all t, t',*

$$\mathbb{E}[Y_{it}(0)|D_i = 1] - \mathbb{E}[Y_{it}(0)|D_i = 0] = \mathbb{E}[Y_{it'}(0)|D_i = 1] - \mathbb{E}[Y_{it'}(0)|D_i = 0].$$

Equivalently, we can formulate this assumption in terms of changes over time. In that formulation, the assumption is that the expected change in control outcomes is the same for those that will eventually be exposed to the treatment and those that will not:

$$\mathbb{E}[Y_{it}(0) - Y_{it'}(0)|D_i = 1] = \mathbb{E}[Y_{it}(0) - Y_{it'}(0)|D_i = 0] \qquad \forall t, t'.$$

To motivate this assumption for the panel case an alternative is to postulate a TWFE model for the control outcomes, as in (5.2), with the additional assumption that the treatment assignment D_i is independent of the vector of residuals ε_{it}, $t = 1, \dots, T$ conditional on fixed effects:

$$D_i \perp\!\!\!\perp (\varepsilon_{i1}, \dots, \varepsilon_{iT})|\alpha_i,$$

as in, for example, Arellano (2003). From the point of view of the modern causal inference literature, the parallel trend assumption is somewhat nonstandard because it combines restrictions on the potential outcomes with restrictions on the assignment mechanism–see Ghanem et al. (2022) and Roth and Sant'Anna (2023) for additional discussions.

Consider next the grouped repeated cross-section (GRCS) case. Suppose in the population all groups are large (infinitely large) in each period, and we have random samples from these populations for each period. Then the expectations are well defined as population averages. In that case, the parallel trends assumption can be formulated as requiring that the difference in expected control outcomes between two groups remains constant over time:

Assumption 5.4. *For all pairs of groups g, g' and for all pairs of time periods t, t', the average difference between the groups remains the same over time, irrespective of their treatment status:*

$$\mathbb{E}\left[Y_{gt}(0)\middle|D_i = 1\right] - \mathbb{E}\left[Y_{g't}(0)\middle|D_i = 0\right] = \mathbb{E}\left[Y_{gt'}(0)\middle|D_i = 1\right] - \mathbb{E}\left[Y_{g't'}(0)\middle|D_i = 0\right].$$

Again, an alternative formulation is as the assumption that the expected change between periods t' and t is the same for all groups:

$$\mathbb{E}\left[Y_{gt}(0)\middle|D_i = 1\right] - \mathbb{E}\left[Y_{gt'}(0)\middle|D_i = 1\right] = \mathbb{E}\left[Y_{g't}(0)\middle|D_i = 0\right] - \mathbb{E}\left[Y_{g't'}(0)\middle|D_i = 0\right],$$

for all g, g', t, t'. If we were to observe $Y_{gt}(0)$ for all groups and time periods, then the presence of two groups and two time periods would be sufficient for this assumption to have testable implications. However, with at least one of the four cells defined by the group and time period exposed to the treatment, there are no testable restrictions implied by this assumption in the two-group/two-period case, as in, for example, in the New Jersey/Pennsylvania minimum wage study in Card and Krueger (1994).

Because we can view the panel case as a two-group setting, with the defined in terms of the indicator $D_i \in \{0, 1\}$, there are only testable restrictions from this assumption when we have more than two periods. With more than two groups, just as in the case with more than two periods, there are testable restrictions implied by the parallel trend assumption. In an early paper, Ashenfelter and Card (1985) argued against using the TWFE model for evaluation of training programmes based on the failure of parallel trends detected in the data. See Jakiela (2021), Rambachan and Roth (2023), and Roth et al. (2023) for a discussion and bounds based on limits on the deviations from parallel trends. Bridging some of the gap between design and sampling-based approaches, Roth and Sant'Anna (2023) show how parallel trends can be implied by random assignment of treatment. They also discuss the sensitivity to transformations of the parallel trend assumption. We return to this in Section 8, where we discuss nonlinear methods.

5.5. *Pre-treatment Variables*

Often researchers observe time-invariant characteristics of the units in addition to the time path of the outcome. Such characteristics cannot be incorporated simply by adding them to the TWFE specification in (5.2) because they would be perfectly colinear with the individual components α_i. Nevertheless, the pre-treatment variables can be important by facilitating a decomposition of these effects into explained components, as in Plümper and Troeger (2007), or by allowing the relaxation of some of the key assumptions. Specifically, one can assume that the parallel trend and constant treatment effect assumptions hold only within subpopulations defined by these characteristics. Two specific proposals have been made for the applications with time-invariant covariates.

5.5.1. Abadie (2005). In an early paper, Abadie (2005) proposes flexible ways of adjusting for time-invariant covariates while continuing with a conditional version of the parallel trends assumption. His solution was based on re-weighting the differences in outcomes by the propensity score to ensure balance.

5.5.2. Sant'Anna and Zhao (2020). Sant'Anna and Zhao (2020) use recent advances in the cross-section causal inference literature to adjust for time-invariant covariates in a doubly robust way by combining inverse propensity score weighting with outcome modelling. In cross-sectional settings, such doubly robust methods have been found to be more attractive than either outcome modelling or inverse propensity score weighting on their own. They do maintain the parallel trends assumption conditional on covariates.

With finite T, strictly exogenous time-varying covariates X_{it} can be converted to time-invariant $X_i \equiv (X_{i1}, \ldots, X_{iT})$. Applied researchers rarely follow this practice and instead rely on linear specifications with contemporaneous covariates. For a discussion of the problems with the conventional specifications and a potential solution, see Caetano et al. (2022).

5.6. *Unconfoundedness*

One key distinction between the repeated cross-section and proper panel case (and also the grouped, repeated cross-section case after aggregation) is that in the case with proper panel data there is a natural alternative to the TWFE estimator. This is most easily illustrated in the case with blocked assignment, where the treatment group is only exposed to the treatment in the last period. Viewing the pre-treatment outcomes as covariates, one could assume unconfoundedness:

$$D_i \perp\!\!\!\perp \Big(Y_{iT}(0), Y_{iT}(1)\Big) \Big| \; Y_{i1}, \ldots, Y_{iT-1}.$$

If one is willing to make this assumption, the larger literature on the estimation of treatment effects under unconfoundedness applies. See Imbens (2004) for a survey. Modern methods include doubly robust methods that combine modelling outcomes with propensity score weighting. See, for example, Bang and Robins (2005), Chernozhukov et al. (2017), and Athey et al. (2018).

Consider the case with two periods, $T = 2$. Because unconfoundedness is equivalent to assuming

$$D_i \perp\!\!\!\perp \Big(Y_{i2}(0) - Y_{i1}, Y_{i2}(1) - Y_{i1}\Big) \Big| \; Y_{i1},$$

it follows that the issue in the choice between TWFE and unconfoundedness is really whether one should adjust for differences between treated and control units in the lagged outcome, Y_{i1}. The TWFE approach implies one should not and that doing so may introduce biases that are otherwise absent, and the unconfoundedness approach implies one should adjust for these differences.

The unconfoundedness assumption and TWFE model validate different non-nested comparisons and applied researchers often do not carefully and explicitly motivate their choices in this regard. The key difference between the two models is the underlying selection mechanism. The TWFE model assumes that the treated units differ from the control ones in unobserved characteristics that are potentially correlated with a persistent component of the outcomes—the fixed effect α_i. The unconfoundedness assumption, however, is satisfied when the selection is based solely on past rather than future outcomes (and potentially other observed pre-treatment variables).

The methodological literature does not provide a lot of guidance on the choice between these two strategies, with exceptions in Angrist and Pischke (2008) and Xu (2023). It is somewhat segmented, with some subliteratures focusing solely on fixed effect strategies and some solely focusing on unconfoundedness approaches. For example, there is a large literature reanalysing the data originally studied in LaLonde (1986)—see also Dehejia and Wahba (1999) and Imbens and Xu (2024)—where the researcher has observations on two lagged outcomes. Although LaLonde reports estimates from various TWFE models in addition to estimates that adjust for initial period outcomes, in the subsequent literature the focus is almost entirely on methods assuming unconfoundedness. In contrast, most of the literature reanalysing the data originally studied in Card and Krueger (1994) where the researcher observes outcomes for a single pre-treatment period has focused on TWFE and related methods with relatively little attention paid to unconfoundedness approaches. It is not clear what motivates the differences in emphasis in these two applications. In an earlier study, Ashenfelter and Card (1985) carefully point out the limitations of the TWFE model and, in particular, its inability to capture temporary declines in earnings prior to enrolment in labour market programmes, the so-called Ashenfelter dip. In our view the unconfoundedness approach is perhaps under-utilised in the empirical panel data literature with the case for the fixed effect specification overstated.

There are two important cases where the unconfoundedness and TWFE approaches lead to similar results. Again, this is most easily seen in the two-period case. The results from the two approaches are similar if the averages of the initial period outcomes are similar for the two groups *or* if the average in the control group did not change much over time. One way to think about this case is to view it as one where there are multiple potential control groups. One can use the contemporaneous control group, or one can use the treatment group in the first period. If either the control group does not change over time or if the treatment group and the control group do not differ in the first period, then the two potential control groups deliver the same results. See for more on this multiple control group perspective (Rosenbaum, 2002).

When the control group changes over time, and in addition the control group and treatment group differ in the initial period, then the TWFE and unconfoundedness approaches give different results. However, the differences can be bounded, albeit under additional assumptions. Suppose unconfoundedness holds and the distribution of the pre-treatment outcomes in the treatment group stochastically dominates that in the control group. Then, the TWFE estimator will underestimate the true effect. However, if the TWFE model holds, then assuming unconfoundedness and adjusting for the lagged outcome will overestimate the true effect. See Angrist and Pischke (2008, ch. 5.4) for a derivation in the linear case, and Ding and Li (2019) for a nonparametric generalisation that allows for heterogeneity in treatment effects, i.e., failure of Assumption 5.2. Imai et al. (2023) take a middle ground between unconfoundedness assumptions and TWFE/DID methods by conditioning on lagged outcomes other than the most recent one, which is differenced out in a TWFE approach.

5.7. Distributional Effects

Our discussion in this section has focused on an average effect τ. This is without essential loss of generality as long as Assumption 5.2 holds, but, in practice, researchers do not expect this assumption to hold exactly. As we discuss in the next section, heterogeneity in treatment effects does not create problems for the DID estimator, which continues to estimate an interpretable average treatment effect. At the same time, when treatment effects are heterogeneous, researchers can be interested in estimands that capture the distributional effects of the treatment, and panel data provides the opportunity to estimate such effects.

In particular, in Bonhomme and Sauder (2011), the authors show how to use deconvolution techniques to identify the full distribution of the treatment effects for the treated units, as long as the treatment effects and treatment assignment W_{it} are statistically independent of the idiosyncratic errors ε_{it} introduced in Assumption 5.1. Note that this restriction does not put any structure on the correlation between the assignment, unit fixed effects α_i, and treatment effects, thus allowing for rich selection patterns, e.g., selection on the treatment effects themselves.

Callaway et al. (2018) and Callaway and Li (2019) use a different strategy and show that as long as the difference in outcomes is not correlated with the treatment assignment, researchers can identify quantile treatment effects under additional stability restrictions on the joint distribution of the potential outcomes. Their approach is connected to the changes-in-changes setup in Athey and Imbens (2006) which we discuss in more detail in Section 8.

6. THE STAGGERED ADOPTION CASE

Although much of the panel literature starts with the TWFE model for control outcomes (Assumption 5.1) with constant treatments effects (Assumption 5.2), the constant treatment effect assumption is not important in the setting with block assignment. Maintaining the TWFE for control outcomes, but allowing unrestricted heterogeneity in the treatment effects $Y_{it}(1) - Y_{it}(0)$, the TWFE estimator (which then has the double difference form) continues to estimate a well-defined average causal effect, namely the average treatment effect for the treated in the periods in which they received the treatment,

$$\frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\Big(Y_{it}(1) - Y_{it}(0)\Big).$$

The interpretation is more complex in settings with dynamic treatment effects, but the underlying estimand is still well defined. This robustness to treatment effect heterogeneity does not extend to settings outside of block assignment.

Part of the new causal panel literature builds on traditional TWFE methods in the staggered adoption setting, allowing for general heterogeneity in the treatment effects. The twin goals are understanding what is estimated by TWFE estimators in this setting that is common in empirical work (see De Chaisemartin and d'Haultfoeuille, 2023, for evidence on how common this setting is), and proposing modifications that ensure that a meaningful average causal effect is estimated. We review that literature here. Recall the staggered adoption setting,

$$\begin{array}{c}\mathbf{W}^{\text{stag}} = \\ \text{(staggered adoption)}\end{array} \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & 0 & 1 & \dots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & 1 & 1 & \dots & 1 & 1 \end{pmatrix} \quad \begin{array}{c} \text{(never adopter)} \\ \text{(very late adopter)} \\ \text{(late adopter)} \\ \text{(medium adopter)} \\ \vdots \\ \text{(early adopter).} \end{array}$$

Let $A_i \equiv T + 1 - \sum_{t=1}^{T} W_{it}$ be the adoption date (the first time unit i is treated if a unit is always treated), with the convention that $A_i \equiv \infty$ for units that never adopt the treatment, and recall that N_a is the number of units with adoption date $A_i = a$. Define also the average treatment effect by time and adoption date,

$$\tau_{t|a} \equiv \mathbb{E}\left[\, Y_{it}(1) - Y_{it}(0) \middle|\, A_i = a\right].$$

The key is that these average treatment effects can vary both by time and by adoption date. Such heterogeneity was rarely allowed for in the earlier literature, with an early exception in Chamberlain (1992) and more recently in Arellano and Bonhomme (2011a), Graham and Powell (2012), and Chernozhukov et al. (2013). We discuss the connection between this literature and the modern one at the end of this section. Note that in this setting, we cannot separate the presence of dynamic effects from heterogeneity in the treatment effects over time and by adoption date.

6.1. Decompositions of the TWFE Estimator

Here we discuss the interpretation of the TWFE estimator $\hat{\tau}$ based on the least squares regression

$$\min_{\alpha,\beta,\tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \alpha_i - \beta_t - \tau W_{it}\right)^2,$$

in the staggered adoption case. This decomposition is based on the discussion in Goodman-Bacon (2021). We maintain Assumption 5.1, which implies the TWFE structure for the control outcomes and the mean-independence between the residuals and the treatment indicator.

Define for all time periods t and all adoption dates a the average outcome in period t for units with adoption date a:

$$\overline{Y}_{t|a} \equiv \frac{1}{N_a} \sum_{i:A_i=a} Y_{i,t}.$$

Then, for all pairs of time periods $t > t'$ and pairs of adoption dates a, a' such that $t' < a \leq t$ (units with adoption date a change treatment between t and t') and either $a' \leq t'$ or $t < a'$ (units with adoption date a' do not change treatment status between t and t', they are either already treated before period t', or only adopt the treatment after period t), define the following double difference that is the building block for the TWFE estimator:

$$\hat{\tau}_{t,t'}^{a,a'} \equiv \left(\overline{Y}_{t|a} - \overline{Y}_{t'|a}\right) - \left(\overline{Y}_{t|a'} - \overline{Y}_{t'|a'}\right). \tag{6.1}$$

The interpretation of this double difference plays a key role in the interpretation of the TWFE estimator $\hat{\tau}$. The group with adoption date a changes treatment status between periods t' and t, so the difference $\overline{Y}_{t|a} - \overline{Y}_{t'|a}$ reflects a change in treatment, but this treatment effect is contaminated by the time trend in the control outcome under the TWFE structure:

$$\mathbb{E}\left[\overline{Y}_{t|a} - \overline{Y}_{t'|a'}\right] = \beta_t - \beta_{t'} + \tau_{t|a}.$$

For the group with an adoption date a', the difference $\overline{Y}_{t|a'} - \overline{Y}_{t'|a'}$ does not capture a change in treatment status. If $t < a'$, it is a difference in average control outcomes, and $\hat{\tau}_{t,t'}^{a,a'}$ is a standard DID estimand, which under the TWFE model for the control outcomes has an interpretation as an average treatment effect. Roth et al. (2023) refer to this as a ‘clean’ comparison.

However, if $a' < t'$, the difference $\overline{Y}_{t|a'} - \overline{Y}_{t'|a'}$ is a difference in average outcomes under the treatment. In the presence of treatment effect heterogeneity, and in the absence of a TWFE model for the outcomes under treatment, its expectation can be written as

$$\mathbb{E}\left[\overline{Y}_{t|a'} - \overline{Y}_{t'|a'}\right] = \beta_t - \beta_{t'} + \left(\tau_{t|a'} - \tau_{t'|a'}\right).$$

Hence, in the case with $a' < t'$, the basic building block in (6.1) has expectation

$$\mathbb{E}\left[\hat{\tau}^{a,a'}_{t,t'}\right] = \tau_{t|a} - \left(\tau_{t|a'} - \tau_{t'|a'}\right).$$

This is a weighted average of treatment effects, with the weights adding up to one, but with some of the weights negative. This is sometimes referred to as a ‘forbidden’ comparison (Roth et al., 2023). If the treatment effects are all identical, this does not, in fact, create a concern. However, if there is reason to believe there is substantial heterogeneity, as is likely in practice, researchers may be reluctant to report weighted averages with negative weights. Note that the concern with the comparisons $\hat{\tau}^{a,a'}_{t,t'}$ when $a' < t'$, but not when $a' > t$ fundamentally treats the treated state and the control state asymmetrically: the parallel trends assumption is maintained for the control outcomes, but not for the treated outcomes.

The TWFE estimator $\hat{\tau}^{\mathrm{TWFE}}$ can be characterised as a linear combination of the building blocks $\hat{\tau}^{a,a'}_{t,t'}$, including those where the non-changing group has an early adoption date $a' < t'$. The coefficients in that linear combination depend on various aspects of the data, including the number of units N_a in each of the corresponding adoption groups, as discussed in detail in Baker et al. (2021) and Goodman-Bacon (2021) for the staggered adoption case and in Imai and Kim (2021) for the general case. As a result, the TWFE estimator has two distinct problems. First, without further assumptions, the estimator does not have an interpretation as an estimate of an average treatment effect with non-negative weights in general. Second, the combination of weights on the building blocks chosen by the TWFE regression depends on the data, in particular on the distribution of units across the adoption groups. As a result, two identical populations in terms of potential outcome distributions (and thus identical treatment effect distributions) that have different adoption patterns would lead to different estimated quantities.

We emphasise that the expectations above are computed with respect to the errors ε_{it} holding the adoption dates fixed. This is in line with the fixed effects tradition in the panel data literature, which does not restrict the conditional distribution of unit-specific parameters, such as τ_{it}, given the covariates of interest, which in our case corresponds to A_i. In some situations, e.g., in randomised experiments, the adoption date is unrelated to the τ_{it} and thus the conditional distribution of the τ_{it} is equal to its marginal distribution, and the negative weights issue does not necessarily arise, e.g., Arkhangelsky et al. (2021). We return to this point and its connection to the random effects tradition in the panel data literature in Section 9.

We also note that in other settings, including linear regression, researchers often report estimates that in the presence of treatment effect heterogeneity represent weighted averages of treatment effects with some of the weights negative. While that is not necessarily ideal, there are in the current setup trade-offs with other assumptions, including the parallel trend assumptions, that may force the researcher to make some assumptions that are, at best, approximations. Similar trade-offs motivate the use of higher-order kernels in nonparametric regression, which also lead to estimators with negative weights. We, therefore, do not view the negative weights of some estimators as necessarily disqualifying. We also find the terminology ‘clean’ and ‘forbidden’ not doing justice to the potential benefits from such methods.

6.2. Alternative DID-Type Estimators for the Staggered Adoption Setting

To deal with the negative weights, researchers have recently, more or less contemporaneously, proposed a number of different modifications to the TWFE estimator. Here we discuss four of these modifications that have attracted considerable attention. It should be noted that all maintain the TWFE assumption for the control outcomes, and all four avoid the additional assumption on treatment effect heterogeneity.

6.2.1. Callaway and Sant'Anna (2021). Callaway and Sant'Anna (2021) propose two ways of dealing with the negative weights. Their first approach takes a group with adoption date a, and compares average outcomes in any post-adoption period $t \geq a$ ($\overline{Y}_{t|a}$ for $t \geq a$) to average outcomes for the same group (the group with adoption date a) immediately prior to the adoption ($\overline{Y}_{a-1|a}$). It then subtracts the difference in outcomes for the same two time periods for the single group that never adopts the treatment ($a = \infty$). Formally, consider, for $t \geq a$, the double difference

$$\hat{\tau}^{a,\infty}_{t,a-1} = \left(\overline{Y}_{t|a} - \overline{Y}_{a-1|a}\right) - \left(\overline{Y}_{t|\infty} - \overline{Y}_{a-1|\infty}\right). \tag{6.2}$$

A concern is that this particular control group, those who never adopt the treatment, may not be particularly attractive. One might worry that the very fact that this group never adopts the treatment is an indication that they are fundamentally different from the other groups and thus less suitable as a comparison for the trends in the absence of the treatment. In addition, very few of these never adopters may exist, especially in long panels, so the precision of the estimators based on such comparisons may make them unattractive.

Recognising this concern, Callaway and Sant'Anna (2021) suggest using as an alternative control group the average of the groups that do adopt the treatment, but restricting this to those who adopt after period t:

$$\hat{\tau}^{a,>t}_{t,a-1} \equiv \left(\overline{Y}_{t|a} - \overline{Y}_{a-1|a}\right) - \frac{1}{T-t}\sum_{a'=t+1}^{T}\left(\overline{Y}_{t|a'} - \overline{Y}_{a-1|a'}\right).$$

Given these two estimators, Callaway and Sant'Anna (2021) suggest reporting averages over periods t and adoption dates a, using a variety of possible weight functions $\omega(a, t)$ that depend on the adoption date and the time period. One of their preferred weight functions is

$$\omega_e(a, t) = \mathbf{1}_{a+e=t} \cdot \mathrm{pr}(A_i = a | A_i \leq T - e),$$

which leads to an average of treatment effects, over different adoption dates, at exactly e periods after adoption, for their two control groups,

$$\hat{\tau}^{\mathrm{CS,I}}(e) = \sum_{a=2}^{T-e} \omega_e(a, t) \cdot \hat{\tau}^{a,\infty}_{t,a-1}, \quad \text{or} \quad \hat{\tau}^{\mathrm{CS,II}}(e) = \sum_{a=2}^{T-e} \omega_e(a, t) \cdot \hat{\tau}^{a,>t}_{t,a-1}.$$

We should note that Callaway and Sant'Anna (2021) also allow for the possibility that the treatment is anticipated, and so that up to some known number of periods prior to the treatment, the outcome may already be affected by this.

6.2.2. Sun and Abraham (2020). Sun and Abraham (2020) start with one of the same building blocks as Callaway and Sant'Anna (2021), $\hat{\tau}^{a,\infty}_{t,a-1}$ in (6.2). Given double differences of this type they suggest reporting the average of this:

$$\hat{\tau}^{\mathrm{SA}} = \sum_{t=2}^{T} \sum_{a=2}^{t} \hat{\tau}_{t,a-1}^{a,\infty} \cdot \frac{\mathrm{pr}(A_i = a|2 \leq A_i \leq t) \cdot \mathbf{1}_{2 \leq a \leq t \leq T}}{T-1}.$$

This is a simple unweighted average over the periods t after the first period, with the weights within a period equal to the fraction of units with an adoption date prior to that, excluding first-period adopters.

An additional issue emphasised by Sun and Abraham (2020) is related to the validation of the two-way model. In applications, this validation is done by testing for parallel trends using pre-treatment data. Sun and Abraham (2020) show that common implementation of such tests using two-way specifications with leads of treatments also include comparisons with negative weights. As a result, they caution against such procedures.

6.2.3. De Chaisemartin and d'Haultfœuille (2020). De Chaisemartin and d'Haultfœuille (2020) deal with the negative weights by focusing on one period ahead of double differences, with control groups that adopt later ($a > t$):

$$\hat{\tau}_{t,t-1}^{t,a} = \left(\overline{Y}_{t|t} - \overline{Y}_{t-1|t}\right) - \left(\overline{Y}_{t|a} - \overline{Y}_{t-1|a}\right).$$

They aggregate these by averaging over all groups that adopt later:

$$\hat{\tau}_{+,t} = \frac{1}{T-(a-1)} \sum_{a>t} \hat{\tau}_{t,t-1}^{t,a}.$$

Then they average over the time periods, weighted by the fraction of adopters in each period:

$$\hat{\tau}^{\mathrm{CH}} = \sum_{t=2}^{T} \hat{\tau}_{+,t} \cdot \mathrm{pr}(A_i = a|A_i \geq 2).$$

One challenge with the De Chaisemartin and d'Haultfœuille (2020) approach is that by limiting the comparisons to those that are separated by a single period, the standard errors may be large relative to those for estimators based on more comparisons. Although the additivity assumption may be more likely to hold over such short horizons, there is also increased sensitivity to the presence of dynamic effects.

6.2.4. Borusyak et al. (2021). Borusyak et al. (2021) focus on a model for the baseline outcomes that is richer than the TWFE model:

$$Y_{it}(0) = A_{it}^{\top}\lambda_i + X_{it}^{\top}\delta + \epsilon_{it},$$

where A_{it} and X_{it} are observed covariates, leading to a factor-type structure. This setup reduced to the TWFE for $A_{it} \equiv 1$ and $X_{it} \equiv (\mathbf{1}_{t=1}, \ldots, \mathbf{1}_{t=T})$. They propose estimating λ_i and δ by least squares using only observations for control units only, and later construct unit time-specific imputations for the unobserved control outcomes for the treated units, leading to unit/period-specific treatment effect estimates:

$$\hat{\tau}_{it} = Y_{it} - A_{it}^{\top}\hat{\lambda}_i + X_{it}^{\top}\hat{\delta}.$$

These unit-specific estimators can then be aggregated into an estimator for the target of interest; let us call the estimator $\hat{\tau}^{\mathrm{BJS}}$. Notably, despite each unit time-specific treatment effect estimator $\hat{\tau}_{it}$ being inconsistent, after these objects are averaged, the estimator is well behaved. Moreover, Borusyak et al. (2021) show that the resulting estimator is efficient as long as ϵ_{it} is i.i.d. over i and t, which relies on a version of the Gauss-Markov theorem for their setup.

6.2.5. Discussion. If one is concerned with the negative weights in the TWFE estimator in a setting with staggered adoption, how should one choose between these four alternatives, $\hat{\tau}^{\mathrm{CS,I}}$ (or $\hat{\tau}^{\mathrm{CS,II}}$), $\hat{\tau}^{\mathrm{SA}}$, $\hat{\tau}^{\mathrm{CH}}$, and $\hat{\tau}^{\mathrm{BJS}}$? The first key issue is the choice of the estimand. In staggered designs there are many average effects one can estimate, and the choice of which one to report should be addressed carefully depending on the underlying research question. Once this choice is made, there are some substantive arguments that matter for the choice of the estimator: (i) the never adopter group may well be substantively different from groups that eventually adopt, (ii) for long differences (where we compare outcomes for time periods far apart) the assumption that differences between units are additive and stable over time becomes increasingly less plausible, (iii) one-period differences may be quite different from differences based on comparisons separated by multiple periods if there are dynamic effects, and (iv) efficiency considerations. These concerns do not lead to one proposal clearly dominating the others, and, in practice, looking for a single estimator may be the wrong goal.

What should one do instead? One option is to report all of the proposed estimators, as, for example, Braghieri et al. (2022), who report estimates based on all four approaches in addition to the standard TWFE estimator. However, that does not do justice to the fact that the estimators rely on fundamentally different assumptions, in particular about treatment effect heterogeneity, and focus on different estimands. Moreover, some of these comparisons may have little power in terms of uncovering heterogeneity of particular forms. Finally, other than Borusyak et al. (2021), the methods all rely on some version of parallel trend assumptions. Ultimately, instead of reporting all estimators, we therefore recommend exploring directly the presence of systematic variation in the $\hat{\tau}_{t,t'}^{a,a'}$, by adoption date, a, by the length of the period between before and after, $t-t'$, and the time since adoption, $t-a$.

6.2.6. Relation to earlier literature. Here we relate the discussion in this section to some earlier results in econometric panel data literature. In Chamberlain (1992), Arellano and Bonhomme (2011a), and Graham and Powell (2012), the authors analyse a class of panel data models that incorporates the two-way model with heterogeneous treatment effects as a special case. For example, Arellano and Bonhomme (2011a) postulate the following model:

$$\mathbf{Y_i} = \mathbf{Z}_i\delta + \mathbf{X}_i\gamma_i + \boldsymbol{\varepsilon}_i, \quad \mathbb{E}[\boldsymbol{\varepsilon}_i|\gamma_i, \mathbf{Z}_i, \mathbf{X}_i] = 0, \tag{6.3}$$

where $\mathbf{Y_i}$ is a T-dimensional vector of outcomes, and $\mathbf{Z}_i$ and $\mathbf{X}_i$ are matrices of regressors for unit i, and $\boldsymbol{\varepsilon}_i$ is a T-dimensional vector of errors. To see that this model includes those discussed in this section, first define $Z_i = \mathcal{I}_T$, a $T \times T$ identity matrix, and $\delta = (\beta_1, \ldots, \beta_T)^\top$. Next, define

$$\mathbf{X}_i = \begin{pmatrix} 1 & W_{i,1} & 0 & \ldots & 0 \\ 1 & 0 & W_{i,2} & \ldots & 0 \\ \ldots & & & & \\ 1 & 0 & 0 & \ldots & W_{i,T} \end{pmatrix}$$

and $\gamma_i = (\alpha_i, \tau_{i1}, \ldots, \tau_{iT})^\top$. Equation (6.3) then reduces to the two-way model with heterogeneous treatment effects:

$$Y_{it} = \alpha_i + \beta_t + \tau_{it}W_{it} + \varepsilon_{it}.$$

A similar relation applies to the setup described in Graham and Powell (2012). Both of these models are a particular instance of the setup described in Section 4 of Chamberlain (1992).

The earlier econometric literature did not focus on the properties of the fixed effects estimator for a misspecified version of (6.3), but instead was concerned with directly estimating distributional characteristics of γ_i, such as $\mathbb{E}[\gamma_i]$ in Chamberlain (1992) and Graham and Powell (2012), or $\mathbb{V}[\gamma_i]$ and higher-order moments in Arellano and Bonhomme (2011a). Because of this, the key assumption in these papers is the presence of units for which the matrix $\mathbf{X}_i^\top \mathbf{X}_i$ has a full rank. This assumption is needed to impute the value of γ_i:

$$\hat{\gamma}_i = \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^{-1} \mathbf{X}_i^\top (\mathbf{Y}_i - \mathbf{Z}_i \hat{\delta}),$$

where $\hat{\delta}$ is a consistent estimator for the common parameter δ, which is typically available.

Because of the dimension of $\mathbf{X}_i$, this approach is infeasible in the two-way model with heterogeneous treatment effects, and it is impossible to identify any distributional characteristics of γ_i for any subpopulation of units without additional assumptions. At the same time, often we are not interested in the distributional characteristics of the entire vector γ_i, but instead focus on components thereof, such as the average treatment effect in the periods when units are treated. For such estimands the results are more positive as illustrated by the causal panel data literature. From this perspective, one can view the strategies discussed above, in particular the imputation approach of Borusyak et al. (2021), as an extension of Arellano and Bonhomme (2011a) to settings where only some components of γ_i can be estimated.

Chernozhukov et al. (2013) is another example from the econometric panel data literature that emphasises the heterogeneity in treatment effects. For cases with binary regressors, their model has the following structure:[3]

$$Y_{it} = \beta_t + \lambda_t\,(\alpha_i + \tau_i W_{it}) + \varepsilon_{it}, \quad \mathbb{E}[\varepsilon_{it}|\alpha_i, \tau_i, W_{i1}, \ldots, W_{iT}] = 0.$$

This model allows for more flexible baseline potential outcomes generalising Assumption 5.1. This aspect connects it to factor models that we discuss in Section 7. At the same time, the heterogeneity in treatment effects is limited and connected to heterogeneity in the baseline outcomes, unlike in other models discussed in this section. The authors show that certain average treatment effects can be estimated in this model as long as there is enough variation in the treatment W_{it}. These results would typically apply to staggered designs as long as there are at least two different adoption periods. The authors also discuss the estimation of quantile effects, which relies on additional distributional restrictions on ε_{it}.

6.3. *Modelling Heterogeneity*

In Section 4, we discussed that the distinctive feature of the panel data analysis arises from restricting the potential outcomes. Assumptions 5.1 and 5.2 do exactly that by imposing a very special structure on the underlying potential outcomes. Both of these assumptions are restrictive, and the new causal panel data literature discussed in this section focused on fully relaxing Assumption 5.2 while maintaining the two-way model for a particular set of potential outcomes that are effectively used as a control group.

In practice, the choice of the control group depends on application details and, to a certain degree, is arbitrary. From this perspective, the analysis that fully relaxes Assumption 5.2 while

[3] For binary regressors, this form is equivalent to the nonseparable model described in the paper. The assumptions that the authors make imply additional distributional restrictions on ε_{it}, which are not needed for the estimation of average effects.

keeping Assumption 5.1 is somewhat internally inconsistent. If we are willing to assume that the baseline potential outcomes follow a simple model, then it is not clear why we would not be willing to make a similar assumption for the treatment effects. After all, the differences in the baseline potential outcomes partly arise due to other unobserved treatments, and if their effect is fully heterogeneous, then the two-way model is unlikely to hold.

One interpretation of the results in this section is that we do not need to take a stand on the degree of treatment effect heterogeneity because there exist methods that are fully robust to it as long as the two-way model holds, allowing us to relax one assumption at a time. This conclusion, however, is unlikely to hold in more complicated settings, e.g., see Arellano and Honoré (2001) for a related impossibility result in sequentially exogenous models. One can still attempt to relax Assumption 5.2, but simultaneously relax at least part of Assumption 5.1. Which one of those approaches deserves more attention is fundamentally an empirical question. More broadly, we recommend letting the data determine the degree of underlying heterogeneity in potential outcomes. In the next section, we discuss a class of methods that does exactly that.

7. MOVING AWAY FROM THE TWO-WAY FIXED EFFECT STRUCTURE

A key strand of the recent causal panel literature starts with the introduction of the synthetic control (SC) method by Alberto Abadie and co-authors, initially in Abadie and Gardeazabal (2003), with more detailed methodological discussions in Abadie et al. (2010; 2015). This brought a substantially different perspective to the questions studied in the TWFE literature. Initially, the SC literature remained completely separate from the TWFE discussions. The SC literature focused on imputing missing potential outcomes by creating synthetic versions of the treated units constructed as convex combinations of control units. This more algorithmic, as opposed to model-based, approach has inspired much new research, ranging from factor-model approaches that motivate synthetic control type algorithms to hybrid approaches that link synthetic control methods to the earlier TWFE methods and highlight their connections.

In this section we first discuss the basic synthetic control method in Section 7.1. Next, in Section 7.2, we discuss the direct estimation of factor models. In Section 7.3 we discuss some hybrid methods that combine synthetic control and TWFE components.

7.1. Synthetic Control Methods

In the original paper, Abadie and Gardeazabal (2003) were interested in estimating the causal effect on terrorism on the Basque region economy. They constructed a comparison for the Basque region based on a convex combination of other regions in Spain. The weights were chosen to ensure that this synthetic Basque region matched the actual Basque region closely in the years pre-treatment (prior to the terrorism) years.

In a short period of time, this synthetic control method has become a widely used approach, popular in empirical work in social sciences as well as in the popular press (including *The Economist* and *The Guardian*), with many theoretical advances in econometrics, statistics, and computer science. The key papers by Abadie, Diamond, and Hainmueller that discuss the details of the original synthetic control proposals are Abadie et al. (2010; 2015). For recent reviews see Samartsidis et al. (2019) and Abadie (2021).

7.1.1. Estimation. Here we use a characterisation of the SC method as a least squares estimator, as discussed in Doudchenko and Imbens (2016), that is slightly different from that in Abadie et al. (2010). We focus on the case without covariates. Suppose unit N is the sole treated unit, and is treated in period T only. Define the weights $\hat{\omega}$ as the regression estimates subject to restrictions:

$$\hat{\omega} \equiv \arg \min_{\omega | \omega \geq 0, \sum_j \omega_j = 1} \sum_{t=1}^{T-1} \left(Y_{Nt} - \sum_{j=1}^{N-1} \omega_j Y_{jt} \right)^2, \tag{7.1}$$

and then impute the missing potential outcome as

$$\hat{Y}_{NT}(0) = \sum_{j=1}^{N-1} \hat{\omega}_j Y_{jT}.$$

The non-negative weights $\hat{\omega}_j$ define the 'synthetic' control that gave the methods its name. One remarkable finding in the initial papers by Abadie and co-authors is that this solution is typically sparse, with positive weights $\hat{\omega}_j > 0$ only for a small subset of the control units. Although this is not always important substantively, it greatly facilitates the interpretation of the results. For example, in the German reunification application in Abadie et al. (2015) where the full set of potential controls consists of sixteen OECD countries, only five countries, Austria, Japan, the Netherlands, Switzerland, and the United States, have positive weights.

The characterisation of the SC estimator in (7.1) allows for an interesting comparison with methods based on the unconfoundedness assumption discussed in Section 5.6. With a linear model specification, unconfoundedness would suggest an estimator

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^{N-1} \left(Y_{iT} - \beta_0 - \sum_{s=1}^{T-1} \beta_s Y_{is} \right)^2, \tag{7.2}$$

followed by the imputation of the missing potential outcome as

$$\hat{Y}_{NT}(0) = \hat{\beta}_0 + \sum_{s=1}^{T-1} \hat{\beta}_s Y_{Ns}.$$

The difference is that, in using the terminology of Athey et al. (2021), SC in the regression in (7.1) relies on *vertical regression* with $T - 1$ observations and $N - 1$ predictors, with some restrictions on the parameters (with the units of observations corresponding to the columns of $\mathbf{Y}$), and (7.2) relies on *horizontal regression* with $N - 1$ observations and T regressors after including an intercept (with the units of observations corresponding to the columns of $\mathbf{Y}$). If the minimisers in these least squares regressions are not unique, we take the solution to be the one that minimises the L_2 norm (Spiess et al., 2023). See Shen et al. (2022) for more insights into the comparison between the horizontal and vertical regressions in this setting. In particular, they demonstrate the interesting insight that point estimates for the counterfactuals are identical for the vertical and horizontal regressions in the absence of the non-negativity and adding up restrictions.

One interesting aspect of the synthetic control approach is that it is more algorithmic than many other methods used in these settings. Consider the estimator based on unconfoundedness in (7.2). Such an approach is typically motivated by a linear model

$$Y_{iT} = \gamma_0 + \sum_{s=1}^{T-1} \gamma_s Y_{is} + \varepsilon_i,$$

with assumptions on the ε_i given the lagged outcomes. The corresponding model for the SC estimator would be

$$Y_{Nt} = \sum_{j=1}^{N-1} \omega_j Y_{jt} + \eta_t,$$

with assumptions on η_t given the contemporaneous outcomes for other units. However, such assumptions are rarely postulated, and for good reason. It would postulate a relationship between the cross-section units, e.g., states, that is oddly asymmetric. If, as in the application in Abadie et al. (2010), California is the treated state, this model would postulate a relationship between California and the other states of a form that cannot also hold for all other states. Attempts to specify models that justify the synthetic control estimator had limited success. Abadie et al. (2010) discuss factor models as a data-generating process, but that begs the question of why one would not directly estimate the factor model. Researchers have done so, as discussed in Section 7.2 below, but interestingly, such attempts have not always outperformed the Abadie–Diamond–Haimueller synthetic control methods, suggesting the latter have attractive properties that are not fully understood yet. See the review in Abadie (2021) for more discussion on conditions under which synthetic control methods are appropriate.

In Arkhangelsky and Samkov (2024), the authors show that in environments where the contribution of idiosyncratic errors to the overall variation in the outcomes is small, i.e., most of the variation is explained by the two-way fixed effects and the factor model, a version of the SDID estimator discussed in Section 7.3 is asymptotically equivalent to a particular methods-of-moments estimator for the underlying factor model. These results are relevant for applications where researchers rely on aggregated data, such as the GRCS data discussed in Section 3.1.2, and bridge the gap between methods that directly estimate factors models and those based on the SC ideas. Further theoretical research is needed to better understand this connection in other settings.

7.1.2. Modifications. A number of modifications have been suggested to the basic version of the SC estimator. Hsiao et al. (2012), Doudchenko and Imbens (2016), and Ferman and Pinto (2021) suggest making the estimator more flexible by allowing for an intercept in the regression (or, equivalently, applying the method to outcomes in deviations from time averages). Hsiao et al. (2012), Doudchenko and Imbens (2016), and Gardeazabal and Vega-Bayo (2017) also discuss allowing the weights to be outside the unit interval. This improves the in-sample fit, but has the potential of making the out-of-sample predictions less accurate.

Li (2023) proposes an alternative to TWFE estimation that relies on selecting a set of controls. One can think of this as a special case of SC where the weights for the control units are either zero, or $1/N_C$, where N_C is the number of control units selected. The proposal includes a greedy algorithm for selecting the set of controls with an objective function that closely mimics the SC criterion for the case with an intercept.

Typically, in the synthetic control method, only the control units are weighted. In principle, however, one could also weight the treated units to make it easier to find a set of (weighted) control units that are similar to these weighted treated units during the pre-treatment period, as suggested in Kuosmanen et al. (2021).

Kellogg et al. (2021) suggest combining matching and synthetic control methods. Whereas synthetic control methods avoid extrapolation at any cost, combining it with matching allows researchers to lower the bias from either method.

7.1.3. Regularisation. In settings where the number of control units is large relative to the number of pre-treatment periods, this requires some form of regularisation. Hsiao et al. (2012) use statistical information criteria. Doudchenko and Imbens (2016) suggest regularising the weights by imposing an elastic net penalty on the weights ω_i, with the penalty chosen by cross-validation. Spiess et al. (2023) avoid the choice of a penalty term by choosing the minimum L_2 norm value for the weights within the set of weight combinations that lead to the optimal in-sample fit, in the spirit of the recent double descent literature (Belkin et al., 2019). Abadie and L'hour (2021) recognise that weights of a convex combination of control units that are all far away from the treated unit are not as attractive as a convex combination of control units that are all close to the target treated unit. They suggest choosing the weights by minimising the sum of the original synthetic control criterium and a term that penalises the distance between any of the control units and the target unit

$$\hat{\omega} = \arg \min_{\omega|\omega\geq 0, \sum_j \omega_j = 1} \sum_{t=1}^{T-1} \left(Y_{Nt} - \sum_{j=1}^{N-1} \omega_j Y_{jt} \right)^2 + \lambda \sum_{j=1}^{N-1} \omega_j \sum_{t=1}^{T-1} (Y_{Nt} - Y_{jt})^2,$$

with the tuning parameter λ chosen through cross-validation, for example, on the control units.

7.1.4. Inference. Inference has been a major challenge in synthetic control settings, and there is, as of yet, no consensus regarding the best way to estimate variances or construct confidence intervals. One particular challenge is that the methods are often used in settings with just a single treated unit/period, or relatively few treated unit/period pairs, making it difficult to rely on central limit theorems for the distribution of estimators. In applications where the number of units and periods is large, the situation is different; see the results in Arkhangelsky et al. (2021) and Ferman (2021).

One approach has been to use placebo methods to test sharp null hypotheses, typically for the null hypothesis of no effect of the intervention. Abadie et al. (2010) propose such a method. Suppose there is a single treated unit, say unit N. Abadie et al. (2010) construct a distribution of estimates based on each control unit being analysed as the treated unit and then calculate the p-value for unit N as the quantile in that distribution of placebo estimates. See also Firpo and Possebom (2018) for an extension and additional analysis of this method.

Doudchenko and Imbens (2016) suggest that the same placebo approach can be based on changing the time period that was treated. Essentially here the idea is to think of the time of the treatment as random, generating a randomisation distribution of estimates. In a related approach Chernozhukov et al. (2021), Lei and Candès (2021), and Viviano and Bradic (2023) develop conformal inference procedures that rely on the exchangeability of the residuals from some model over time. Cattaneo et al. (2021) propose the construction of prediction intervals for the counterfactual outcome.

7.2. Matrix Completion Methods and Factor Models

A second set of methods that relaxes the TWFE assumptions focuses directly on factor models, where the outcome is assumed to have the form

$$Y_{it}(0) = \sum_{r=1}^{R} \alpha_{ir}\beta_{tr} + \varepsilon_{it}. \tag{7.3}$$

First, note that this generalises the TWFE specification: if we fix the rank at $R = 2$, and set $\alpha_{i2} = 1$ for all i and $\beta_{t1} = 1$ for all t, this is identical to the TWFE specification, but the factor model obviously allows for more general dependence structures in the data. Although such factor models have a long tradition in panel data, e.g., Chamberlain and Rothschild (1983), Anderson (1984), Stock and Watson (1998), Bai and Ng (2002), and Bai (2009), the recent causal literature has used them in different ways.

7.2.1. Matrix completion with nuclear norm regularisation. Athey et al. (2021) take an approach that models the entire matrix of potential control outcomes as

$$Y_{it}(0) = L_{it} + \alpha_i + \beta_t + \varepsilon_{it},$$

where the ε_{it} is random noise, uncorrelated with the other components. The matrix $\mathbf{L}$ with typical element L_{it} is a low-rank matrix. As mentioned above the unit and time components α_i and β_t could be subsumed in the low-rank component as they on their own form a rank-two matrix, but in practice it improves the performance of the estimator substantially to keep these fixed effect components in the specification separately from the low-rank component $\mathbf{L}$. The reason is that we regularise the low-rank component $\mathbf{L}$, but not the individual and time components. Building on the matrix completion literature (Candès and Recht, 2009; Candès and Plan, 2010), Athey et al. (2021) propose the nuclear-norma-matrix-completion (NNMC) estimator based on minimising

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(1 - W_{it})\left(Y_{it} - L_{it} - \alpha_i - \beta_t\right)^2 + \lambda\|\mathbf{L}\|_*,$$

over $\mathbf{L}$, α, and β. The missing $Y_{it}(0)$ values are then imputed using the estimated parameters. Here the nuclear norm $\|\mathbf{L}\|_*$ is the sum of the singular values $\sigma_l(\mathbf{L})$ of the matrix $\mathbf{L}$, based on the singular value decomposition $\mathbf{L} = \mathbf{S}\Sigma\mathbf{R}$, where $\mathbf{S}$ is $N \times N$, Σ is the $N \times T$ diagonal matrix with the singular values and $\mathbf{R}$ is $T \times T$. The penalty parameter λ is chosen through out-of-sample cross-validation. The nuclear norm regularisation shrinks towards a low-rank estimator for $\mathbf{L}$, similar to the way LASSO shrinks towards a sparse solution in linear regression.

7.2.2. Robust synthetic control. Amjad et al. (2018) focus on the case with a single treated unit. They start with a factor model $\mathbf{Y} = \mathbf{L} + \varepsilon$. They would like to use a synthetic control estimator with denoised matrix $\mathbf{L}$ as the control outcomes, rather than the actual outcomes $\mathbf{Y}$. They implement this through a two step procedure. In the first step the matrix $\mathbf{L}$ is estimated by taking the singular value decomposition, and setting all singular values below a threshold μ equal to zero. This leads to a low-rank estimate $\hat{\mathbf{L}}$, which is then scaled by one over p, where p is the maximum of the fraction of observed outcomes and $1/((N-1)T)$.

In the second step Amjad et al. (2018) use the part of this rescaled matrix corresponding to the control units, in combination with the pre-treatment period values for a treated unit, in a standard synthetic control approach. The idea is that using de-noised outcomes $\hat{\mathbf{L}}$ instead of the actual outcomes $\mathbf{Y}$ leads to better predictors by removing an estimate of the noise component ε. In this second synthetic control step Amjad et al. (2018) do not impose the convexity restrictions on the weights, but do add a regularisation penalty.

7.2.3. Interactive fixed effect or factor models. Building on the factor model literature in econometrics (Chamberlain and Rothschild, 1983; Holtz-Eakin et al., 1988; Chamberlain, 1992; Pesaran, 2006; Bai, 2009; Moon and Weidner, 2015; 2018; Freyberger, 2018), Xu (2017) study direct estimation of factor models as an alternative to synthetic control methods. The basic setup models the control potential outcome as in (7.3). The number of factor is then estimated or pre-specified and the model is directly estimated after some normalisation. Based on this model one can impute the missing potential outcomes for the treated unit/time-period pairs and use that to estimate the average effect for the treated. See Gobillon and Magnac (2016) for an application.

7.2.4. Grouped panel data. Bonhomme and Manresa (2015) and Bonhomme et al. (2022) consider a factor model, but impose a group structure. In our causal setting, their setup would correspond to

$$Y_{it}(0) = \theta_{G_i,t} + \varepsilon_{it},$$

with the group membership unknown. They focus on the case with the number of groups G known. In that case one can write the model as a factor model with G factors λ_{rt} and the loadings equal to indicators, $\alpha_{ir} = \mathbf{1}_{G_i=r}$, so that

$$Y_{it}(0) = \theta_{G_i,t} + \varepsilon_{it} = \sum_{r=1}^{G} \alpha_{ir}\lambda_{rt} + \varepsilon_{it}.$$

Computationally this grouped structure creates substantial challenges. Chetverikov and Manresa (2022); Mugnier (2022) suggest alternative estimation methods that are computationally more attractive.

7.2.5. Tuning. One disadvantage that the methods discussed in this section share is the need to specify the tuning parameters. This sets them apart from the conventional TWFE methods that we discussed before and makes them harder to adopt in practice. In the case of the matrix completion estimator proposed by Athey et al. (2021), this tuning parameter is the regularisation parameter λ that quantifies the importance of the nuclear norm penalty. In the context of the standard interactive fixed effects estimators, one needs to specify the rank of the underlying factor model. The same applies to the estimator based on finitely many groups. In principle, one can use traditional techniques from the machine learning literature, such as cross-validation, to find appropriate values of these parameters. The panel dimension, however, creates an additional challenge on how exactly to implement the cross-validation. It is thus attractive to have methods that generalise the two-way methodology and do not require explicit tuning. One such proposal is Moon and Weidner (2018), where the authors analyse the limiting version of the estimator from Athey et al. (2021) with λ approaching zero. They show that the resulting estimator is consistent under relatively weak assumptions, albeit can converge at a slower rate.

7.3. *Hybrid Methods*

Two recent methods combine some of the benefits from the synthetic control approach with either TWFE ideas or with unconfoundedness methods. These methods are particularly attractive because of the nest TWFE being able to accommodate more flexible outcome models. There are in essence two approaches. One can directly generalise the outcome model, or one can use a local version of the TWFE model. This is somewhat similar to the way one can generalise a linear regression model by making the regression function more flexible through the inclusion of additional function of the regressors, or by estimating it locally through kernel methods.

7.3.1. Synthetic difference in differences. For expositional reasons let us consider the case with a single treated unit and time period, say unit N in period T, although the insights readily extend to the block assignment case. Once the researcher has calculated the SC weights, the SC estimator for the treatment effect can be characterised as a weighted least squares regression,

$$\min_{\beta,\tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\omega}_i \left(Y_{it} - \beta_t - \tau W_{it}\right)^2. \tag{7.4}$$

It is useful to contrast this with the TWFE estimator, which is based on a slightly different least squares regression:

$$\min_{\beta,\alpha,\tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \alpha_i - \beta_t - \tau W_{it}\right)^2. \tag{7.5}$$

The two differences are that (i), the SC regression in (7.4) uses weights $\hat{\omega}_i$, and (ii) the TWFE regression in (7.5) has unit-specific fixed effects α_i.

In light of this comparison, and more generally in the context of the larger panel data literature, the omission of the unit fixed effects from the synthetic control regression may seem surprising. Arkhangelsky et al. (2021) exploit this by proposing what they call the synthetic difference-in-difference (SDID) estimator that includes both the unit fixed effects α_i and the SC weights $\hat{\omega}_i$, as well as analogous time weights $\hat{\lambda}_t$, leading to

$$\min_{\beta,\alpha,\tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\omega}_i \hat{\lambda}_t \left(Y_{it} - \alpha_i - \beta_t - \tau W_{it}\right)^2.$$

The time weights $\hat{\lambda}_t$ are calculated in a way similar to the unit weights,

$$\min_{\lambda} \sum_{i=1}^{N-1} \left(Y_{iT} - \sum_{s=1}^{T-1} \lambda_s Y_{is}\right)^2,$$

subject to the restriction that $\lambda_s \geq 0$, and $\sum_{s=1}^{T-1} \lambda_s = 1$. The weights for treated units and periods are equal to 1.

7.3.2. Augmented synthetic control. Ben-Michael et al. (2021) augment the SC estimator by regressing the outcomes in the treatment period on the lagged outcomes using data for the control units. Suppose that, following Ben-Michael et al. (2021), one uses ridge regression for this first step, again in the setting with unit N and period T the only treated unit/time-period pair:

$$\hat{\eta} = \arg\min_{\eta} \sum_{i=1}^{N-1} \left(Y_{iT} - \eta_0 - \sum_{s=1}^{T-1} \eta_s Y_{is} \right)^2 + \lambda \sum_{s=1}^{T-1} \eta_s^2,$$

with ridge parameter λ chosen through cross-validation. A standard unconfoundedness approach would predict the potential control outcome for the treated unit/time-period pair as

$$\hat{Y}_{NT} = \hat{\eta}_0 + \sum_{s=1}^{T-1} \hat{\eta}_s Y_{Ns}.$$

The augmented SC estimator modifies this by combining it with SC weights in a way that can be seen either as a bias-adjustment to the unconfoundedness estimator, or a bias-adjustment to the SC estimator:

$$\begin{aligned}\hat{Y}_{NT} &= \hat{\eta}_0 + \sum_{s=1}^{T-1} \hat{\eta}_s Y_{Ns} + \sum_{i=1}^{N-1} \omega_i \left(Y_{iT} - \hat{\eta}_0 - \sum_{s=1}^{T-1} \hat{\eta}_s Y_{is} \right) \\ &= \sum_{i=1}^{N-1} \omega_i Y_{iT} + \sum_{s=1}^{T-1} \hat{\eta}_s \left(Y_{Ns} - \sum_{j=1}^{N-1} \omega_j Y_{js} \right).\end{aligned}$$

Ben-Michael et al. (2022) extend this approach to the case with staggered adoption.

7.3.3. The connection between unconfoundedness, difference in differences, synthetic control, and matrix completion. Although methods based on unconfoundedness and synthetic control estimators, difference-in-differences, and matrix completion estimators appear to be quite different, they are in fact closely related. We want to highlight two insights regarding these connections.

We focus on the case with a single treated unit/period pair, say unit N in period T. The observed control outcomes are $\mathbf{Y}$, an $N \times T$ matrix with the (N, T) entry missing. We partition this matrix as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_0 & \mathbf{y}_1 \\ \mathbf{y}_2^\top & ? \end{pmatrix},$$

where $\mathbf{Y}_0$ is a $(N-1) \times (T-1)$ matrix, and $\mathbf{y}_1$ and $\mathbf{y}_2$ are $(N-1)$ and $(T-1)$ component vectors, respectively.

First, Shen et al. (2022) discuss an interesting connection between SC estimators and estimators based on unconfoundedness in combination with linearity. In that case we first estimate a linear regression

$$Y_{iT} = \gamma_0 + \sum_{s=1}^{T-1} \gamma_s Y_{is} + \varepsilon_i,$$

and then impute the missing outcome as $\hat{Y}_{NT} = \hat{\gamma}_0 + \sum_{s=1}^{T-1} \hat{\gamma}_s Y_{Ns}$. Shen et al. (2022) show that if we drop the intercept from this regression, $\gamma_0 = 0$, then the unconfoundedness imputation is identical to the SC imputation (where we also restrict the weights to be non-negative and sum to one).

In an alternative connection, Athey et al. (2021) show that in some cases linear versions of all four estimators can all be characterised as solutions to the same optimisation problem, subject to different restrictions on parameters of that optimisation problem. To see this, define for a given positive integer R, an $N \times R$ matrix $\mathbf{U}$, an $T \times R$ matrix $\mathbf{V}$, an N-vector α and a T-vector β, and a scalar λ the objective function

$$Q(R, \mathbf{U}, \mathbf{V}, \alpha, \beta, \lambda) \equiv \sum_{i=1}^{N} \sum_{t=1}^{T} (1 - W_{it}) \left(Y_{it} - \sum_{r=1}^{R} U_{ir} V_{tr} - \alpha_i - \beta_t \right)^2 + \lambda \left(\sum_{i=1}^{N} \sum_{r=1}^{R} U_{ir}^2 + \sum_{t=1}^{T} \sum_{r=1}^{R} V_{tr}^2 \right). \tag{7.6}$$

When $R = 0$, we take the product $\mathbf{U}\mathbf{V}^\top$ to be the $N \times T$ matrix with all elements equal to zero. Given $\mathbf{U}$, $\mathbf{V}$, α and β the imputed value for Y_{NT} is $\hat{Y}_{NT} = \sum_{r=1}^{R} U_{Nr} V_{Tr} - \alpha_i - \beta_t$.

First, note that minimising the objective function (7.6) over the rank R, the matrices $\mathbf{U}$, $\mathbf{V}$ and the vectors α and β given $\lambda = 0$, does not lead to a unique solution. By choosing the rank R to the minimum of N and T, we can find for any pair α and β a solution for $\mathbf{U}$ and $\mathbf{V}$ such that $(1 - W_{it})(Y_{it} - \sum_{r=1}^{R} U_{ir} V_{tr} - \gamma_i - \delta_t) = 0$ for all (i, t), with different imputed values for Y_{NT}. The implication is that we need to add some structure to the optimisation problem. The next result shows that unconfoundedness regression, the SC estimator, the DID estimator, and the matrix completion (MC) estimator can all be expressed as minimising the objective function under different restrictions on, or with different approaches to, regularisation of $(R, \mathbf{U}, \mathbf{V}, \alpha, \beta)$.

Nuclear norm matrix completion: the nuclear norm matrix completion estimator chooses λ through cross-validation.

Unconfoundedness: the unconfoundedness regression is based on regressing $\mathbf{Y}_1$ on $\mathbf{y}_2^\top$ and an intercept. It can also be characterised as the solution to minimising (7.6) with the restrictions

$$R = T - 1, \qquad \mathbf{U} = \begin{pmatrix} \mathbf{Y}_0 \\ \mathbf{y}_2^\top \end{pmatrix}, \quad \alpha = 0, \quad \beta_1 = \beta_2 = \ldots = \beta_{T-1} = 0, \qquad \lambda = 0.$$

Synthetic control (SC): the SC estimator imposes the restrictions subject to

$$R = N - 1, \quad \mathbf{V} = \begin{pmatrix} \mathbf{Y}_0^\top \\ \mathbf{y}_1^\top \end{pmatrix}, \quad \alpha = 0, \quad \beta = 0, \quad \forall\, i, U_{iT} \geq 0, \ \sum_{i=1}^{N-1} U_{iT} = 1, \quad \lambda = 0.$$

Difference in differences: The DID estimator fixes $R = 0$.

7.3.4. The selection mechanism. Another set of insights concerning the differences between the various estimators emerges from a focus on the selection mechanism. First, Ghanem et al. (2022) show that outside of a few special cases to justify the conventional parallel trends assumption, one needs to assume that the treatment is unrelated to time-varying components of the outcomes. The restrictiveness of this assumption presents a challenge for the DID estimator, which relies on this parallel trends assumption. These concerns are less relevant for other estimators; see Arkhangelsky and Hirshberg (2023) and Imbens and Viviano (2023) for two recent discussions.

We can see this in the setting with the block design, where there are $T_0 + 1$ periods, and some units are treated in the last period, with D_i being the treatment indicator, $D_i = \mathbf{1}_{W_{iT}=1}$. Suppose the underlying potential outcomes follow a static two-way model of Section 5:

$$Y_{it}(0) = \alpha_i + \beta_t + \varepsilon_{it}, \quad \varepsilon_{it} \perp\!\!\!\perp \alpha_i, \quad \tau = Y_{it}(1) - Y_{it}(0).$$

The key feature that determines the performance of different algorithms in this environment is the relationship between D_i and the vector of errors $(\varepsilon_{i1}, \ldots, \varepsilon_{iT_0+1})$. As long as D_i is mean-independent from $(\varepsilon_{i1}, \ldots, \varepsilon_{iT_0+1})$, then the discussed estimators will have good statistical properties. This should not be surprising for the DID (which does not rely on large T_0) or matrix completion estimator because their statistical properties are established under this assumption. The fact that the SC estimator would work well in this situation follows from the results in Arkhangelsky et al. (2021) and Arkhangelsky and Hirshberg (2023).

This conclusion changes dramatically if we allow $(\varepsilon_{i1}, \ldots, \varepsilon_{iT_0+1})$ to be correlated with D_i. If this correlation is completely unrestricted, then any observed differences in outcomes in the two groups can be attributed to differences in errors, and it is impossible to identify the effect using any method. Suppose, however, that we make a natural selection assumption

$$D_i \perp\!\!\!\perp Y_{iT_0+1}(0) | \alpha_i, Y_{i1}, \ldots, Y_{iT_0}(0),$$

which restricts the correlation of D_i with ε_{i,T_0+1}. Note that this restriction combines both the selection on fixed effect assumption discussed in Section 5.4 and the unconfoundedness assumption discussed in Section 5.6.

As long as ε_{it} are autocorrelated, the DID estimator is inconsistent, even when T_0 goes to infinity. The reason for this failure is that $\varepsilon_{iT_0+1} - \sum_{t \leq T_0} \varepsilon_{it}/T_0$ remains correlated with D_i which introduces bias. The performance of the SC estimator is different, and the results in Arkhangelsky and Hirshberg (2023) show that the SC estimator is consistent and asymptotically unbiased as long as T_0 goes to infinity. The consistency properties of the matrix completion estimator and the unconfoundedness regression are not established for this setting.

Imbens and Viviano (2023) focus on a factor model with block assignment where D_i can be correlated with the factor loadings and the time of initial exposure can be correlated with the factors. They present conditions under which the SC estimator is consistent.

This discussion illustrates that to analyse the behaviour of algorithmically related estimators one needs to take a stand on the underlying selection mechanism. Most of the recent results in the causal panel data literature are established under strict exogeneity, which does not allow D_i to be correlated with ε_{it}. Understanding the performance of different estimators in environments where such correlation is present is an attractive area of future research that can benefit from the econometric panel data literature.

7.3.5. Simulation comparisons. There have been a number of studies comparing various DID and SC estimators in simulations. These have not always been in realistic settings, limiting their usefulness for practitioners. In fact, it is not as easy in longitudinal setting to come up with realistic simulation settings that capture both the degree of cross-section and time-series dependence that is present in a given data set. In pure cross-section settings, Athey et al. (2024) suggest using generative adversarial networks to generate realistic data-generating processes, but that does not immediately extend to longitudinal settings. Arkhangelsky et al. (2021) compare DID, SC, SDID, and NNMC estimators in settings motivated by state and country panel data sets. They first estimate factor models with four factors and use that to construct a data-generating process with additive fixed effects, four additional factors, and an autoregressive error process. They find that there is substantial variation in the performance of the methods, with SDID typically outperforming the other methods.

8. NONLINEAR MODELS

In this section, we discuss some nonlinear panel data models. By nonlinear models, we mean here models where the conditional mean function is not linear in parameters. Part of this literature is motivated by the concern that the standard fixed effect models maintain additivity and linearity in a way that does not do justice to the type of data that are often analysed. With binary outcomes, it is particularly difficult to justify the standard TWFE model. At the same time, estimating the unit and time fixed effects inside a logistic or probit model does not lead to consistent estimators for the effects of interest in typical settings.

8.1. Changes in Changes

Athey and Imbens (2006) focus on the repeated cross-section case with two periods and two groups, one treated in the second period and one never treated. They are concerned with the functional form dependence of the standard TWFE specification in levels. If the model,

$$Y_i(0) = \mu + \alpha \mathbf{1}_{C_i=1} + \beta \mathbf{1}_{T_i=1} + \varepsilon_i,$$

holds in levels, then obviously it cannot hold in general in logarithms. In fact, in some cases, one can test that the model cannot hold in levels. Suppose the outcome is binary, and suppose that the potential control outcome averages by group and time period are $\overline{Y}_{11}(0) = 0.2$ (for the first-period control group), $\overline{Y}_{12} = 0.8$ (for the second-period control group), and $\overline{Y}_{21}(0) = 0.7$ (for the first-period treatment group). Then the additive TWFE model implies that the second-period treatment group, in the absence of the treatment, would have had average outcome $0.7 + (0.8 - 0.2) = 1.3$, which of course, is not feasible with binary outcomes.

To address this concern Athey and Imbens (2006) propose a scale-free changes-in-changes (CIC) model for the potential control outcomes,

$$Y_i = g(U_i, T_i),$$

where the U_i is an unobserved component that has a different distribution in the treatment group and the control group, but a distribution that does not change over time. The standard TWFE model can be viewed as the special case where $g(u, t)$ is additively separable in u and t:

$$g(u, t) = \beta_0 + u + \beta_1 t,$$

implying that the expected control outcomes can be written in the TWFE form as

$$\mathbb{E}[Y_i(0)|T_i = t, G_i = g] = \beta_0 + \beta_1 t + \alpha \mathbb{E}[U_i|G_i = g] - \mathbb{E}[U_i|G_i = 1].$$

Athey and Imbens (2006) show that if U_i is a scalar, and $g(u, t)$ is strictly monotone in u, one can infer the second-period distribution of the control potential outcome in the treatment group as

$$F_{Y_i(0)|T_i=2,G_i=2}(y) = F_{Y_i(0)|T_i=1,G_i=2}\left(F^{-1}_{Y_i(0)|T_i=1,G_i=1}\left(F_{Y_i(0)|T_i=2,G_i=1}(y)\right)\right).$$

This, in turn, can be used to estimate the average effect of the intervention on the second-period outcomes for the treatment group.

The expression for the counterfactual distribution of the control outcome for the second-period treatment group has an analogue in the literature on wage decompositions, see Altonji and Blank (1999). Arkhangelsky (2019) discusses a similar approach to the CIC estimator in Athey and Imbens (2006), where the role of the groups and time periods are reversed and also considers an extension for multiple outcomes. Wooldridge (2022) also studies nonlinear versions of DID/TWFE approaches. In the two-period two-group setting, his starting point assumes there is a known function $g : \mathbb{R} \mapsto \mathbb{R}$ such that $\mathbb{E}[Y_{it}(0)|D_i] = g(\mu + \alpha D + \gamma_t)$, so that there is a parallel trend inside the known transformation $g(\cdot)$. The transformation $g(\cdot)$ could be the exponential function, $g(a) = \exp(a)$ in case of non-negative outcomes, or the logistic function $g(a) = \exp(a)/(1 + \exp(a))$ in case of binary outcomes.

8.2. Distributional Synthetic Controls

Gunsilius (2023) develops a model that has similarities to both the CIC and SC control approaches. He focuses on a setting with repeated cross-sections, where we have a relatively large number of units observed in a modest number of groups, with a modest number of time periods. As in the canonical synthetic control case there is a single treated group. Whereas the synthetic control method chooses weights on the control units so that the weighted controls match the treated outcomes in the pre-treatment periods, the Gunsilius (2023) approach chooses weights on the control groups so that the marginal distribution for the weighted controls matches that for the treated group. The metric is based on the quantile function $F^{-1}_{Y_{gt}}(v)$, for group g and period t. First, weights $\hat{\omega}_{tg}$ are calculated separately for each pre-treatment period t based on the following objective:

$$\hat{\omega}_{tg} = \arg \min_{\omega:\omega\geq 0,\sum_{g=1}^{G-1}\omega_{gt}=1} \frac{1}{M}\sum_{m=1}^{M}\left(\sum_{g=1}^{G-1}\omega_{gt}\hat{F}^{-1}_{Y_{gt}}(V_m) - \hat{F}^{-1}_{Y_{Gt}}(V_m)\right)^2,$$

where the quantile functions are evaluated at M randomly choosing values $v_1, \ldots, v_M$.

In the next step the weights are averaged over time,

$$\hat{\omega}_g = \frac{1}{T-1}\sum_{t=1}^{T-1}\hat{\omega}_{gt}.$$

Finally, the quantile function for the treated group in the absence of the treatment is estimated as the synthetic control average of the control quantile functions:

$$\hat{F}^{-1}_{GT}(v) = \sum_{g=1}^{G-1}\hat{\omega}_g\hat{F}^{-1}_{gT}(v).$$

Note that in the case with $G = 2$, so there is just a single control group, the quantile function for the treated group in the last period in the absence of the treatment is identical to the quantile function for the control group in the last period, and the pre-treatment distributions are immaterial.

8.3. Balancing Statistics to Control for Unit Differences

Arkhangelsky and Imbens (2022) focus on settings where the treatment can switch on and off, as in the assignment matrix in (3.1), unlike the staggered adoption case where the treatment can only switch on. They also assume there are no-dynamic effects. Their focus is on flexibly adjusting for differences beyond additive effects. Allowing for completely unrestricted differences between units would require relying solely on within-unit comparisons. Often the number of time periods is not sufficient to rely on such comparisons and still obtain precise estimates. Arkhangelsky and Imbens (2022) balance these two concerns, the restrictiveness of the TWFE model and the lack of precision when focusing purely on within-unit comparisons, by making assumptions that allow the between-unit differences to be captured by a low-dimensional vector, which then can be adjusted for in a flexible, nonlinear way using some of the insights from the cross-section causal inference literature.

To see the insights most clearly it is useful to start with a simpler setting. Specifically, let us first consider a clustered sampling setting with cross-section data studied in Arkhangelsky and Imbens (2023). In that case a common approach is based on a fixed effect specification

$$Y_i = \alpha_{C_i} + \tau W_i + \beta X_i + \varepsilon_i,$$

where C_i is the cluster indicator for unit i. Estimating τ by least squares is the same as estimating the following regression function by least squares,

$$Y_i = \mu + \tau W_i + \gamma \overline{W}_{C_i} + \beta X_i + \delta \overline{X}_{C_i} + \eta_i,$$

$\overline{W}_c$ is the cluster average of the treatment for cluster c, and similar for $\overline{X}_c$. This equivalence has been known since Mundlak (1978).

Arkhangelsky and Imbens (2023) build on the Mundlak insight, still in the clustered setting, by making the unconfoundedness assumption that

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \,\Big|\, X_i, \overline{X}_{C_i}, \overline{W}_{C_i}.$$

Implicitly, this uses the two averages $\overline{X}_{C_i}$ and $\overline{W}_{C_i}$ as proxies for the differences between the clusters. This idea is related to Altonji and Matzkin (2005), who also use exchangeability to control for unobserved heterogeneity. Given this uconfoundedness assumption, one can then adjust for differences in $(X_i, \overline{X}_{C_i}, \overline{W}_{C_i})$ in a flexible way, through nonparametric adjustment methods, possibly in combination with inverse propensity score weighting. Arkhangelsky and Imbens (2023) then generalise this by assuming that

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \,\Big|\, X_i, S_{C_i},$$

where the sufficient statistic S_c captures the relevant features of the cluster, possibly including distributional features such as the average of W_i in the cluster, but also other averages such as the average of the product of X_i and W_i in the cluster.

Arkhangelsky and Imbens (2022) extend these ideas from the clustered cross-section case to the panel data case. They focus on the no-dynamics case where the potential outcomes are indexed only by the binary contemporaneous treatment. In panel data settings, an alternative to two-way fixed effect regressions is the least squares regression

$$Y_{it} = \tau \ddot{W}_{it} + \varepsilon_{it},$$

where $\ddot{W}_{it}$ is the double difference

$$\ddot{W}_{it} = W_{it} - \overline{W}_{i\cdot} - \overline{W}_{\cdot t} + \overline{\overline{W}},$$

with

$$\overline{W}_{i\cdot} = \frac{1}{T}\sum_{t=1}^{T} W_{it}, \quad \overline{W}_{\cdot t} = \frac{1}{N}\sum_{i=1}^{N} W_{it}, \quad \text{and} \quad \overline{\overline{W}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}.$$

See, for example, Vogelsang (2012). Wooldridge (2021) shows the same estimator can be obtained through what he calls the Mundlak regression,

$$Y_{it} = \tau W_{it} + \gamma \overline{W}_{i\cdot} + \delta \overline{W}_{\cdot t} + \varepsilon_{it}.$$

Arkhangelsky and Imbens (2022) postulate the existence of a known function $S_i(W_{i1}, \ldots, W_{iT})$ that captures all the relevant components of the assignment vector $\underline{W}_i = (W_{i1}, \ldots, W_{iT})$ (and possibly other covariates, time-varying or time-invariant). Given this balancing statistic, they assume that the potential outcomes are independent of the treatment assignment vector given this balancing statistic:

$$\underline{W}_i \perp\!\!\!\perp Y_{it}(w) \Big| S_i. \tag{8.1}$$

Consider the case where the balancing statistic is fraction treated periods, $S_i = \overline{W}_i$. The unconfoundedness assumption in (8.1) implies that one can compare treated and control units in the same period, as long as they have the same fraction of treated periods over the entire sample. More generally S_i could capture both the fraction of treated periods, as well as the number of transitions between treatment and control groups.

The estimator proposed by Arkhangelsky and Imbens (2022) has a built-in robustness property: it remains consistent if the two-way model is correctly specified or the unconfoundedness given S_i holds. As a result, it does not require researchers to commit to a single identification strategy. This approach creates a link between the TWFE literature and the design-based analysis we discuss in Section 9.

8.4. *Negative Controls, Proxies, and Deconvolution*

The results in Arkhangelsky and Imbens (2022) show how to use panel data to construct a variable that eliminates the unobserved confounding. A related, but different strategy, is to use a panel to construct a set of proxy measures for the unobservables. If these proxy measures do not directly affect either outcomes or treatments, then this restriction can be used for identification. In biostatistics, such proxy variables are called negative control variables. To emphasise the connections between this literature and economic applications, we use these two terms, proxy variables and negative controls, interchangeably. In biostatistics a recent literature focuses on nonparametric identification results for average treatment effects that are based on negative controls (Sofer et al., 2016; Shi et al., 2020). See Ying et al. (2021) for an introductory article. This literature is closely connected to econometric literature on nonparametric identification with measurement error (Hu and Schennach, 2008) and the CIC model (Athey and Imbens, 2006). In a DID setting one can view the pre-treatment outcomes as proxies or negative controls in the sense of this literature. Recently, these arguments have been extended to prove identification results for a class of panel data models in Deaner (2021b).

Proxy variables have a long history in economics. In early applications, Chamberlain (1977) and Griliches (1977) use data on several test scores to estimate returns to schooling accounting for unobserved ability (see also Deaner, 2021a). Using modern terminology, these test scores serve as negative controls. Versions of these strategies have also been successfully used in the traditional panel data literature. For example, Holtz-Eakin et al. (1988) use data on past outcomes to estimate a dynamic linear panel data model with interactive fixed effects with a finite number of periods. They achieve this by eliminating the interactive fixed effects via a quasi-differencing scheme, which is called a bridge function in the negative control literature (see Imbens et al., 2021; Ying et al., 2021).

A similar idea is used in Freyaldenhoven et al. (2019), where the authors consider a setting with an unobserved confounder that can vary arbitrarily over i and t. To eliminate this confounder, the authors assume the presence of a proxy variable that is affected by the same confounder, but is not related to the treatment. As a result, one can eliminate the unobservables by subtracting a scaled proxy variable from the outcome of interest. The appropriate scaling is estimated using the pre-treatment data. In essence, this strategy is analogous to quasi-differencing and is another example of using bridge functions.

An important aspect of the negative control literature, which it shares with most of the methods discussed in this survey, is that it aims to isolate and eliminate the unobserved confounders rather than identify causal effects conditional on unobservables. Alternatively, one can obtain identification under different distributional assumptions that connect the unobservables to outcomes and treatments using general deconvolution techniques. This approach has been successfully employed to answer causal questions in linear panel data models (Arellano and Bonhomme, 2011a; Bonhomme and Sauder, 2011) and nonlinear quantile panel data models (Arellano and Bonhomme, 2016), but so far has not been widely adopted by a broader causal community.

8.5. *Combining Experimental and Observational Data*

Another direction this literature has explored is the combination of experimental and observational data. Athey et al. (2020) study the case with an experimental data set that has observations on short-term outcomes, and an observational sample that has information on the short-term outcome and the primary outcome. A key assumption is that the observational sample has an unobserved confounder that leads to biases in the comparison of the short-term outcome by treatment group. The experimental data allows one to remove the bias and isolate the unobserved confounder, which then can be used to eliminate biases in the primary outcome comparisons essentially as a proxy variable as discussed in the previous section. See also Kallus and Mao (2020), Imbens et al. (2021), and Ghassami et al. (2022).

9. DESIGN-BASED APPROACHES TO ESTIMATION AND INFERENCE

An issue that features prominently in the recent panel data literature, but is largely absent in the earlier one, is a re-interpretation of the uncertainty in the estimates as coming from the stochastic nature of the causal variables. In most empirical analyses in economics and in most of the methodological literature in econometrics, uncertainty is assumed to be arising from sampling variation. This is a natural perspective if, say, we have data on individuals that can be at least approximately viewed as a random sample from a well-defined population. Had we sampled a different set of individuals, our estimates would have been different, and the standard errors reflect the variation that would have been seen if we repeatedly obtained different random samples from that population. This sampling-based perspective is still a natural one in panel data settings when the units can be viewed as a sample from a larger population, e.g., individuals in the Panel Study of Income Dynamics or the National Longitudinal Survey of Youth.

The sampling-based perspective is less natural in cases where the sample is the same as the population of interest. This is quite common in panel data settings, for example, when we analyse state-level data from the United States, or country-level data from regions of the world, or all firms in a particular class. It is not clear why viewing such a sample as a random sample from a population is appropriate. Researchers have struggled with interpreting the uncertainty of their estimates in that case. Manski and Pepper write in their analysis of the impact of gun regulations with data from the fifty US states: ‘measurement of statistical precision requires specification of a sampling process that generates the data. Yet we are unsure what type of sampling process would be reasonable to assume in this application. One would have to view the existing United States as the sampling realisation of a random process defined on a superpopulation of alternative nations’ (Manski and Pepper, 2018, p. 234).

An alternative approach to formalising uncertainty focuses on the random assignment of causes, taking the potential outcomes as fixed. This approach has a long history in the analysis of randomised experiments (e.g., Fisher, 1937; Rubin, 1990), where the justification for viewing the causes as random is immediate. For modern discussions, see Imbens and Rubin (2015), Rosenbaum (2023), and Ritzwoller et al. (2024). Recently these ideas have been used to capture uncertainty in observational studies, see Abadie et al. (2020; 2023). The justification in panel data settings is not always quite as clear. Consider one of the canonical applications of synthetic control methods to estimate the causal effect of German reunification in 1989 on West German gross domestic product (GDP). A design-based approach would require the researcher to contemplate an alternative world where either other countries would have joined with East Germany or an alternative world where the reunification with West Germany would have happened in a different year. Both are difficult to contemplate. However, a sampling-based approach would require the researcher to consider a world with additional countries that could experience a unification event, which again is not an easy task.

Design-based perspective has interesting connections with the econometric panel data literature. One such connection comes from the part of the panel data analysis that treats fixed effects as parameters rather than realisations of unobserved random variables, thus tying the inference to a particular set of observed units. The uncertainty in these models comes from errors, which we can interpret as realisations of *unobserved* treatments. In contrast, the design-based uncertainty comes from realisations of the observed treatment. A different connection is with panel data literature on random effects, which assumes that unit-level heterogeneity is uncorrelated with the regressors. This assumption is unlikely to hold in observational studies, but it holds by design in experiments. Again, the difference in analysis comes from fixing different quantities. The traditional analysis fixes the covariates and focuses on the distribution of the unit-level heterogeneity, whereas the design-based analysis does the opposite.

9.1. The TWFE Estimator in the Staggered Adoption Case with Random Adoption Dates

As an example of a design-based approach Athey and Imbens (2022) analyse the properties of the TWFE estimator under assumptions on the assignment process in the staggered adoption setting, keeping the potential outcomes fixed. In that case the assignment process is fully determined by the distribution of the adoption date. Athey and Imbens (2022) derive the randomisation-based distribution of the TWFE estimator under the random assignment assumption alone and present an interpretation for the estimand corresponding to that estimator. They show that as long as the adoption date is randomly assigned, the estimand can be written as a linear combination of average causal effects on the outcome in period t if assigned adoption date a' relative to being assigned adoption date a:

$$\tau_t^{a,a'} = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_{it}(a') - Y_{it}(a)\Big),$$

with the weights summing to one, but generally including negative weights.

Athey and Imbens (2022) show the implications for the estimand of the assumption that there is no anticipation of the treatment (so that the potential outcomes are invariant to the future date of adoption). They also show how the interpretation of the estimand changes further under the additional assumption that there are no-dynamic effects so that the potential outcomes only depend on whether the adoption has taken place or not, but not on the actual adoption date. Rambachan and Roth (2020) discuss the implications of variation in the assignment probabilities and the biases this can create.

9.2. *Switchback Designs*

One design that has recently received considerable attention after a long history is what Cochran (1939) called the *rotation experiment*, and what more recently has been referred to as a *switchback experiment* in Bojinov et al. (2022) or *crossover experiment* in Brown (1980). In such experiments units are assigned to treatment or control in each of a number of periods, with individual units potentially switching between treatment and control groups. Such experiments were originally used in agricultural settings, where, for example, cattle were assigned to different types of feed for some period of time. Using each unit as its own control can substantially improve the precision of estimators compared to assigning each unit to the treatment or control group for the entire study period. Such designs have become popular in tech company settings to deal with spillovers. For example, Lyft and Uber often randomise markets to treatment and control groups, with the assignment changing over time.

9.3. *Experimental Design with Staggered Adoption*

This subsection focuses on the design of experiments where the adoption date, rather than the treatment in each period, is randomly assigned. Early studies, including Hussey and Hughes (2007), Hemming et al. (2015), and Barker et al. (2016), focused on simple designs such as those where a constant fraction of units adopted the treatment in each period after the initial period. Sometimes these designs suggested analyses that allowed for spillovers so outcomes for one or two periods after the adoption would be discarded from the analyses if the focus was on the average treatment effect.

Xiong et al. (2019) focused on the question of optimally choosing the fraction of adopters in each period and showed that instead of it being constant, it was initially small and then larger for some periods, after which it declined again. Bajari et al. (2023) discuss randomisation-based inference for some of these settings and present exact variances for some estimators.

9.4. *Design with Dynamic Effects*

Bojinov et al. (2021) propose unbiased estimators and derive their properties under the randomisation distribution. They allow for dynamics in the treatment effects and essentially unrestricted heterogeneity. They also discuss the biases of the conventional TWFE specifications in their setting. Bojinov et al. (2022) discuss optimal design from a minimax perspective, allowing for carry-over effects where the treatment status in recent periods may affect current outcomes.

9.5. Robust Methods

The design-based approach to estimation and inference is natural in the context of randomised experiments. However, in practice, applied researchers continue using conventional panel data methods, such as TWFE, even with experimental data (e.g., Broda and Parker, 2014; Colonnelli and Prem, 2022). There are multiple practical reasons for this: one can believe that the experiment's description is inconsistent with how it was implemented or think that the TWFE estimator is more precise. These concerns are even more salient in quasi-experimental environments where the data does not come from an experiment, but it is appealing to treat it as such (e.g., Borusyak and Hull, 2022).

To address these issues Arkhangelsky et al. (2021) propose a version of the TWFE estimator that incorporates the design information. The key property of this method is that it delivers a consistent estimator even if the design assumptions do not hold as long as the TWFE model is correctly specified. Algorithmically, it amounts to estimating a weighted version of the standard TWFE model:

$$(\hat{\tau}^{rob}, \hat{\alpha}, \hat{\beta}) = \arg\min_{\tau,\alpha,\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \alpha_i - \beta_t - \tau W_{it})^2 \, \omega_i,$$

where the weights $\{\omega_i\}_{i=1}^{n}$ are constructed using the information about the design. In particular, in environments with staggered design, this amounts to estimating a duration model for the treatment adoption time. See Shaikh and Toulis (2021) for a related approach of using duration models for inference in staggered adoption designs.

10. OPEN QUESTIONS FOR FURTHER RESEARCH

Here we discuss some open questions in the current causal panel data literature.

10.1. Modelling Dynamics in Potential Outcomes

The recent panel data literature has only paid limited attention to dynamic treatment effects, compared to the earlier literature (Anderson and Hsiao, 1981; Heckman and Navarro, 2007; see also Abbring and Heckman, 2007, for an overview), as well as compared to its importance in practice. For example, a curious feature of many of the current methods, including factor models and synthetic control methods, is that they pay essentially no attention to the time ordering of the observations. If the time labels were switched, the estimated causal effects would not change. This seems implausible. Suppose one has data available for T_0 pre-treatment periods. For many of the methods, the researcher would be indifferent between having available the first $T_0/2$ pre-treatment period versus the second $T_0/2$ pre-treatment observations, whereas in practice, one would think that the more recent data would be more valuable.

It seems likely the current literature will take the dynamics more seriously. One direction may be to follow Robins (1986) and a number of follow-up studies that developed a sequential unconfoundedness approach. Viviano and Bradic (2021) discuss this approach in economic contexts and propose an implementation that combines traditional linear models with modern balancing approaches. This analysis, however, ignores unobserved heterogeneity, which is central to the current empirical practice. See also Brodersen et al. (2015), Masini and Medeiros (2021; 2022), and Ben-Michael et al. (2023) for studies that take the time-series dimension of these settings more seriously. Other recent work includes Han (2021) and Brown and Butts (2023).

10.2. Validation

LaLonde (1986) has become an influential paper in the causal inference literature because it provided an experimental data set that could be used to validate new methods for estimating average causal effects under unconfoundedness. There are few longer panel data sets that can deliver the comparisons for validating the various new methods. However, there are methods that can be used to assess the performance of proposed estimators with purely observational data. An early paper with suggested tests is Heckman and Hotz (1989). Currently, many approaches in panel data rely on placebo tests where the researcher pretends the treatment occurred some periods prior to when it actually did; the researcher then estimates the treatment effect for these periods where, in the absence of anticipation effects, the treatment effect is known to be zero. Finding estimates close to zero, both substantially and statistically, is then taken as evidence in favour of the proposed methods. See for examples Imbens et al. (2001) and Abadie et al. (2015). This strategy, however, relies on strict exogeneity and can backfire in models where the selection into treatment is based on shocks to past outcomes, as discussed in Section 7.3.4. See Arkhangelsky and Hirshberg (2023) for a particular illustration of this point and a discussion of alternative validation strategies.

10.3. Connections with Macroeconomics

Nakamura and Steinsson (2018) discuss the impact that ideas from the causal inference literature had on empirical research in macroeconomics. At the same time, the causal inference literature itself can benefit from incorporating macroeconomic ideas, which are particularly relevant in applications with panel data. For example, in Section 4, we discuss that Lucas's critique can be relevant for the interpretation of causal quantities in applications in microeconomics. More broadly, panel data sets allow us to connect micro-level and aggregate time-series variation, providing identification strategies for aggregate effects (e.g., Gabaix and Koijen, 2020) as well as local-level effects (e.g., Arkhangelsky and Korovkin, 2019). Wolf (2023) shows how to combine credible micro and macro evidence to analyse policy-relevant counterfactual in macroeconomic models. We view this as an attractive area of future research.

10.4. Bridging Unconfoundedness and the TWFE Approach

Much of the discussion on unconfoundedness and the TWFE model has been framed in terms of a choice. It is difficult to imagine that a clear consensus will emerge, and finding practical methods that build on both approaches would be useful.

10.5. Continuous Treatments

Much of the recent literature has emphasised the binary treatment case. This has led to valuable new insights, but it is clear that many applications go beyond the binary treatment case. There is a small literature studying these cases, including Callaway et al. (2021) and De Chaisemartin and d'Haultfoeuille (2023), and earlier work in the cross-section setting, e.g., Imbens (2000), but more work is needed. Note that the earlier econometric panel data literature did not distinguish between settings where the variables of interest were binary or continuous.

11. RECOMMENDATIONS FOR EMPIRICAL PRACTICE

The recent literature has greatly expanded the set of methods available to empirical researchers in social sciences in settings that are important in practice. This survey is an attempt to put these methods in context and show the close relationship between various approaches, including two-way fixed effect and synthetic control methods, to provide practitioners with additional guidance on when to use the various methods.

11.1. The Blocked Assignment Setting

Although the standard TWFE estimator (simplifying to the double difference or DID estimator in the special case with the blocked assignment) continues to be widely used, there are now methods available that have superior properties in settings with both cross-section and time dimensions at least modestly large. (In cases with few units and few time periods, there may not be enough information in the data to go beyond the simpler methods.) These methods relax the parallel trends assumption that is unattractive both from a conceptual perspective (because it is tied to a particular functional form) and from a practical perspective (because it is unlikely to hold over long periods of time). Some of the new methods allow for factor structures that generalise the two-way fixed effect setup. Others use synthetic control approaches, sometimes in combination with fixed effects. While none of these methods is likely to dominate uniformly, preliminary simulation evidence in the blocked assignment case (e.g., Arkhangelsky et al., 2021) suggests that many of them dominate TWFE in realistic settings. Recent results in Arkhangelsky and Hirshberg (2023) also suggest that some of these methods, in particular those based on synthetic control, dominate TWFE in settings with more complicated selection mechanisms.

11.2. The Staggered Adoption Case

The staggered adoption case, common in empirical work, opens up new opportunities for estimation strategies (exploiting the variation in adoption times), but also forecloses some options (the standard synthetic control estimator). Some of the recent proposals modify the TWFE estimator and relax the parallel trends assumptions by limiting the comparisons between treated and control outcomes to a subset of the set of possible comparisons. This subset may avoid comparisons distant in time, avoid the use of units that are to be treated at a future date as controls, or, in contrast, avoid the use of units that are never treated. In all cases there is an asymmetry in the way treated outcomes and control outcomes are used that does not appear to do justice to the *ex ante* arbitrariness in the treatment versus control labels. There have not been systematic simulation studies that are informative about realistic settings. Nevertheless, we expect that methods which model both treated and control potential outcomes, implying models for both control outcomes and treatment effects, taking account of dynamic effects as the earlier panel literature did more carefully, will be the most effective.

ACKNOWLEDGEMENTS

This paper originated in the Sargan Lecture presented online at the 2021 Royal Economic Society Meetings. We are grateful for the comments by Manuel Arellano, Apoorva Lal, Ganghua Mei, the editor, Jaap Abbring, and two reviewers. We thank the Office of Naval Research for support under grant numbers N00014-17-1-2131 and N00014-19-1-2468 and Amazon for a gift.

REFERENCES

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., A. Agarwal, R. Dwivedi and A. Shah (2024). Doubly robust inference in causal latent factor models. *arXiv: Econometrics* 2402.11652.

Abadie, A., S. Athey, G. W. Imbens and J. M. Wooldridge (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica 88*(1), 265–96.

Abadie, A., S. Athey, G. W. Imbens and J. M. Wooldridge (2023). When should you adjust standard errors for clustering? *Quarterly Journal of Economics 138*(1), 1–35.

Abadie, A., A. Diamond and J. Hainmueller (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., A. Diamond and J. Hainmueller (2015). Comparative politics and the synthetic control method. *American Journal of Political Science 59*, 495–510.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review 93*, 113–32.

Abadie, A. and J. L'hour (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association 116*(536), 1817–34.

Abbring, J. H. and J. J. Heckman (2007). Econometric evaluation of social programs, Part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, vol. 6B, 5145–303. Amsterdam: Elsevier.

Abbring, J. H. and G. J. van den Berg (2003). The nonparametric identification of treatment effects in duration models. *Econometrica 71*(5), 1491–517.

Abowd, J. M. and D. Card (1989). On the covariance structure of earnings and hours changes. *Econometrica 57*, 411–45.

Abowd, J. M., F. Kramarz and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*(2), 251–333.

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis 11*(4), 581–98.

Altonji, J. G. and R. M. Blank (1999). Race and gender in the labor market. In O. C. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, vol. 3C, 3143–259. Amsterdam: Elsevier.

Altonji, J. G. and R. L. Matzkin (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica 73*(4), 1053–102.

Alvarez, J. and M. Arellano (2003). The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica 71*(4), 1121–59.

Amjad, M., D. Shah and D. Shen (2018). Robust synthetic control. *Journal of Machine Learning Research 19*(1), 802–52.

Anderson, T. W. (1984). Estimating linear statistical relationships. *Annals of Statistics 12*(1), 1–45.

Anderson, T. W. and C. Hsiao (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association 76*(375), 598–606.

Angrist, J. D. and A. Krueger (1991). Does compulsory schooling affect schooling and earnings. *Quarterly Journal of Economics 106*(4), 979–1014.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricists' Companion*. Princeton, NJ: Princeton University Press.

Arellano, M. (1987). Computing robust standard errors for within group estimators. *Oxford Bulletin of Economics and Statistics 49*(4), 431–4.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford: Oxford University Press.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies 58*(2), 277–97.

Arellano, M. and S. Bonhomme (2011a). Identifying distributional characteristics in random coefficients panel data models. *Review of Economic Studies 79*(3), 987–1020.

Arellano, M. and S. Bonhomme (2011b). Nonlinear panel data analysis. *Annual Review of Economics 3*, 395–424.

Arellano, M. and S. Bonhomme (2016). Nonlinear panel data estimation via quantile regressions. *The Econometrics Journal 3*(19), C61–94.

Arellano, M. and J. Hahn (2007). Understanding bias in nonlinear panel models: Some recent developments. In R. Blundell, W. Newey and T. Persson (Eds.), *Advances in Economics and Econometrics*, vol. 3, 381–409. Cambridge: Cambridge Unversity Press.

Arellano, M. and B. Honoré (2001). Panel data models: some recent developments. In J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, vol. 5, 3229–96. Amsterdam: Elsevier.

Arkhangelsky, D. (2019). Dealing with a technological bias: The difference-in-difference approach, Working Paper wp2019_1903, CEMFI, Spain.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens and S. Wager (2021). Synthetic difference-in-differences. *American Economic Review 111*(12), 4088–118.

Arkhangelsky, D. and D. Hirshberg (2023). Large-sample properties of the synthetic control method under selection on unobservables. *arXiv: Econometrics* 2311.13575.

Arkhangelsky, D. and G. W. Imbens (2022). Doubly robust identification for causal panel data models. *Econometrics Journal 25*(3), 649–74.

Arkhangelsky, D. and G. W. Imbens (2023). Fixed effects and the generalized Mundlak estimator. *Review of Economic Studies*, rdad089. Published ahead of print, 7 September.

Arkhangelsky, D., G. W. Imbens, L. Lei and X. Luo (2021). Double robust two-way fixed effect regression for panel data. *arXiv: Econometrics* 1909.09412.

Arkhangelsky, D. and V. Korovkin (2019). On policy evaluation with aggregate time-series shocks. *arXiv: Econometrics* 1905.13660.

Arkhangelsky, D. and A. Samkov (2024). Sequential synthetic difference in differences. *arXiv: Econometrics* 2404.00164.

Aronow, P. M. and C. Samii (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics 11*(4), 1912–47.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *Review of Economics and Statistics 60*, 47–57.

Ashenfelter, O. and D. Card (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics 67*(4), 648–60.

Ashenfelter, O. and M. Greenstone (2004). Using mandated speed limits to measure the value of a statistical life. *Journal of Political Economy 112*(S1), S226–267.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens and K. Khosravi (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association 116*(536), 1716–30.

Athey, S., R. Chetty and G. Imbens (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv: Statistics, Methodology* 2006.09676.

Athey, S. and G. W. Imbens (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica 74*(2), 431–97.

Athey, S. and G. W. Imbens (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics 226*, 62–79.

Athey, S., G. W. Imbens, J. Metzger and E. Munro (2024). Using Wasserstein generative adversarial networks for the design of Monte Carlo simulations. *Journal of Econometrics 240*, 105076.

Athey, S., G. W. Imbens and S. Wager (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(4), 597–623.

Athey, S. and S. Stern (2002). The impact of information technology on emergency health care outcomes. *RAND Journal of Economics 33*(3), 399–432.

Auerbach, E. (2022). Identification and estimation of a partially linear regression model using network data. *Econometrica 90*(1), 347–65.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–79.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bajari, P., B. Burdick, G. W. Imbens, L. Masoero, J. McQueen, T. S. Richardson and I. M. Rosen (2023). Experimental design in marketplaces. *Statistical Science 1*(1), 1–19.

Baker, A., D. F. Larcker and C. C. Wang (2021). How much should we trust staggered difference-in-differences estimates? Working Paper 3794018, SSRN.

Baltagi, B. (2008). *Econometric Analysis of Panel Data*. Chichester: Wiley.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–73.

Barker, D., P. McElduff, C. D'Este and M. Campbell (2016). Stepped wedge cluster randomised trials: A review of the statistical methodology used and available. *BMC Medical Research Methodology 16*(1), 1–19.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica 62*, 657–81.

Belkin, M., D. Hsu, S. Ma and S. Mandal (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences 116*(32), 15849–54.

Bell, A. and K. Jones (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods 3*(1), 133–53.

Ben-Michael, E., D. Arbour, A. Feller, A. Franks and S. Raphael (2023). Estimating the effects of a California gun control program with multitask Gaussian processes. *Annals of Applied Statistics 17*(2), 985–1016.

Ben-Michael, E., A. Feller and J. Rothstein (2021). The augmented synthetic control method. *Journal of the American Statistical Association 116*(536), 1789–803.

Ben-Michael, E., A. Feller and J. Rothstein (2022). Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(2), 351–81.

Bertrand, M., E. Duflo and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics 119*(1), 249–75.

Blundell, R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics 87*(1), 115–43.

Bojinov, I., A. Rambachan and N. Shephard (2021). Panel experiments and dynamic causal effects: A finite population perspective. *Quantitative Economics 12*(4), 1171–96.

Bojinov, I., D. Simchi-Levi and J. Zhao (2022). Design and analysis of switchback experiments. *Management Science 69*, 3759–77.

Bonhomme, S. (2012). Functional differencing. *Econometrica 80*(4), 1337–85.

Bonhomme, S. (2020). Econometric analysis of bipartite networks. In B. Graham and A. de Paula (Eds.), *The Econometric Analysis of Network Data*, 83–121. Amsterdam: Elsevier.

Bonhomme, S., T. Lamadon and E. Manresa (2022). Discretizing unobserved heterogeneity. *Econometrica 90*(2), 625–43.

Bonhomme, S. and E. Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica 83*(3), 1147–84.

Bonhomme, S. and U. Sauder (2011). Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *Review of Economics and Statistics 93*(2), 479–94.

Borusyak, K. and P. Hull (2022). Non-random exposure to exogenous shocks. Working Paper 27845, NBER.

Borusyak, K., X. Jaravel and J. Spiess (2021). Revisiting event study designs: Robust and efficient estimation. *arXiv: Econometrics* 2108.12419.

Braghieri, L., R. Levy and A. Makarin (2022). Social media and mental health. *American Economic Review 112*(11), 3660–93.

Broda, C. and J. A. Parker (2014). The economic stimulus payments of 2008 and the aggregate demand for consumption. *Journal of Monetary Economics 68*, S20–36.

Brodersen, K. H., F. Gallusser, J. Koehler, N. Remy and S. L. Scott (2015). Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics 9*, 247–74.

Brown, B. W., Jr. (1980). The crossover experiment for clinical trials. *Biometrics 36*, 69–79.

Brown, N. and K. Butts (2023). Dynamic treatment effect estimation with interactive fixed effects and short panels. Working paper, Queens's University, Kingston, Canada.

Caetano, C., B. Callaway, S. Payne and H. S. Rodrigues (2022). Difference in differences with time-varying covariates. *arXiv: Econometrics* 2202.02903.

Callaway, B., A. Goodman-Bacon and P. H. Sant'Anna (2021). Difference-in-differences with a continuous treatment. *arXiv: Econometrics* 2107.02637.

Callaway, B. and T. Li (2019). Quantile treatment effects in difference in differences models with panel data. *Quantitative Economics 10*(4), 1579–618.

Callaway, B., T. Li and T. Oka (2018). Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. *Journal of Econometrics 206*, 395–413.

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*, 200–30.

Candès, E. J. and Y. Plan (2010). Matrix completion with noise. *Proceedings of the IEEE 98*(6), 925–36.

Candès, E. J. and B. Recht (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics 9*, 717–72.

Card, D. and A. Krueger (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review 84*, 772–93.

Card, D., J. Rothstein and M. Yi (2022). Industry wage differentials: A firm-based approach. Ms., University of California, Berkeley.

Cattaneo, M. D., Y. Feng and R. Titiunik (2021). Prediction intervals for synthetic control methods. *Journal of the American Statistical Association 116*(536), 1865–80.

Chamberlain, G. (1977). Education, income, and ability revisited. *Journal of Econometrics 5*(2), 241–57.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies 47*(1), 225–38.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics 18*(1), 5–46.

Chamberlain, G. (1984). Panel data. In Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics*, vol. 2, 1247–318. Amsterdam: North-Holland.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica 60*, 567–96.

Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica 78*(1), 159–68.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*, 1281–304.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen and W. Newey (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review 107*(5), 261–5.

Chernozhukov, V., I. Fernández-Val, J. Hahn and W. Newey (2013). Average and quantile effects in nonseparable panel models. *Econometrica 81*(2), 535–80.

Chernozhukov, V., K. Wüthrich and Y. Zhu (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association 116*(536), 1849–64.

Chetverikov, D. and E. Manresa (2022). Spectral and post-spectral estimators for grouped panel data models. *arXiv: Econometrics* 2212.13324.

Chiu, A., X. Lan, Z. Liu and Y. Xu (2023). What to do (and not to do) with causal panel analysis under parallel trends: Lessons from a large reanalysis study. Working Paper 4490035, SSRN.

Cochran, W. (1939). Long-term agricultural experiments. *Supplement to the Journal of the Royal Statistical Society 6*(2), 104–40.

Colonnelli, E. and M. Prem (2022). Corruption and firms. *Review of Economic Studies 89*(2), 695–732.

Deaner, B. (2021a). Many proxy controls. *arXiv: Econometrics* 2110.03973.

Deaner, B. (2021b). Proxy controls and panel data. *arXiv: Econometrics* 1810.00283.

Deaton, A. (1985). Panel data from time series of cross-sections. *Journal of Econometrics 30*(1–2), 109–26.

De Chaisemartin, C. and X. d'Haultfœuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–96.

De Chaisemartin, C. and X. d'Haultfoeuille (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *Econometrics Journal 26*(3), C1–30.

Dehejia, R. H. and S. Wahba (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association 94*(448), 1053–62.

Dehejia, R. H. and S. Wahba (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics 84*(1), 151–61.

Ding, P. and F. Li (2019). A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis 27*(4), 605–15.

Donald, S. G. and K. Lang (2007). Inference with difference-in-differences and other panel data. *Review of Economics and Statistics 89*(2), 221–33.

Doudchenko, N. and G. W. Imbens (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, NBER.

Eissa, N. and J. B. Liebman (1996). Labor supply response to the earned income tax credit. *Quarterly Journal of Economics 111*(2), 605–37.

Engle, R. F., D. F. Hendry and J.-F. Richard (1983). Exogeneity. *Econometrica 51*, 277–304.

Fama, E. F., L. Fisher, M. C. Jensen and R. Roll (1969). The adjustment of stock prices to new information. *International Economic Review 10*(1), 1–21.

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), 1764–72.

Ferman, B. and C. Pinto (2021). Synthetic controls with imperfect pretreatment fit. *Quantitative Economics 12*(4), 1197–221.

Fernández-Val, I. and M. Weidner (2016). Individual and time effects in nonlinear panel models with large *N*, *T*. *Journal of Econometrics 192*(1), 291–312.

Fernández-Val, I. and M. Weidner (2018). Fixed effects estimation of large-*T* panel data models. *Annual Review of Economics 10*, 109–38.

Firpo, S. and V. Possebom (2018). Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference 6*(2), 20160026.

Fisher, R. A. (1937). *The Design of Experiments*, London: Oliver And Boyd.

Freyaldenhoven, S., C. Hansen and J. M. Shapiro (2019). Pre-event trends in the panel event-study design. *American Economic Review 109*(9), 3307–38.

Freyberger, J. (2018). Non-parametric panel data models with interactive fixed effects. *Review of Economic Studies 85*(3), 1824–51.

Gabaix, X. and R. S. Koijen (2020). Granular instrumental variables. Technical report, NBER.

Gardeazabal, J. and A. Vega-Bayo (2017). An empirical comparison between the synthetic control method and Hsiao et al.'s panel data approach to program evaluation. *Journal of Applied Econometrics 32*(5), 983–1002.

Ghanem, D., P. H. Sant'Anna and K. Wüthrich (2022). Selection and parallel trends. *arXiv: Econometrics* 2203.09001.

Ghassami, A., A. Yang, D. Richardson, I. Shpitser and E. T. Tchetgen (2022). Combining experimental and observational data for identification and estimation of long-term causal effects. *arXiv: Methodology* 2201.10743.

Gobillon, L. and T. Magnac (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics 98*(3), 535–51.

Goldberger, A. S. (1991). *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Goldsmith-Pinkham, P. (2024). Tracking the credibility revolution across fields. Ms., Yale University and NBER.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–77.

Graham, B. S. and J. L. Powell (2012). Identification and estimation of average partial effects in 'irregular' correlated random coefficient panel data models. *Econometrica 80*(5), 2105–52.

Griliches, Z. (1977). Estimating the returns to schooling: Some econometric problems. *Econometrica 45*, 1–22.

Gunsilius, F. F. (2023). Distributional synthetic controls. *Econometrica 91*(3), 1105–17.

Hahn, J. and G. Kuersteiner (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both *n* and *T* are large. *Econometrica 70*(4), 1639–57.

Han, S. (2021). Identification in nonparametric models for dynamic treatment effects. *Journal of Econometrics 225*, 132–47.

Hansen, C. B. (2007). Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *Journal of Econometrics 140*(2), 670–94.

Heckman, J. J. (1981). Statistical models for discrete panel data. In C. F. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, 114–78. Cambridge, MA: MIT Press.

Heckman, J. J. and V. J. Hotz (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association 84*(408), 862–74.

Heckman, J. J. and S. Navarro (2007). Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics 136*(2), 341–96.

Hemming, K., T. P. Haines, P. J. Chilton, A. J. Girling and R. J. Lilford (2015). The stepped wedge cluster randomised trial: Rationale, design, analysis, and reporting. *BMJ 350*, h391.

Holtz-Eakin, D., W. Newey and H. S. Rosen (1988). Estimating vector autoregressions with panel data. *Econometrica 56*, 1371–95.

Honoré, B. E. (1992). Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica 60*, 533–65.

Honoré, B. E. and E. Tamer (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica 74*(3), 611–29.

Hsiao, C. (2022). *Analysis of Panel Data*. Cambridge: Cambridge University Press.

Hsiao, C., H. S. Ching and S. Ki Wan (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), 705–40.

Hu, Y. and S. M. Schennach (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica 76*(1), 195–216.

Hudgens, M. and E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association 103*(482), 832–42.

Hussey, M. A. and J. P. Hughes (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials 28*(2), 182–91.

Ibragimov, R. and U. K. Müller (2016). Inference with few heterogeneous clusters. *Review of Economics and Statistics 98*(1), 83–96.

Imai, K. and I. S. Kim (2021). On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis 29*(3), 405–15.

Imai, K., I. S. Kim and E. H. Wang (2023). Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science 67*, 587–605.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika 87*, 706–10.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics 86*, 4–29.

Imbens, G. W., N. Kallus and X. Mao (2021). Controlling for unmeasured confounding in panel data using minimal bridge functions: From two-way fixed effects to factor models. *arXiv: Statistics: Methodology* 2108.03849.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

Imbens, G. W., D. B. Rubin and B. I. Sacerdote (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American Economic Review 91*, 778–94.

Imbens, G. W. and D. Viviano (2023). Identification and inference for synthetic controls with confounding. *arXiv: Econometrics* 2312.00955.

Imbens, G. W. and Y. Xu (2024). LaLonde (1986) after nearly four decades: Lessons learned. *arXiv: Econometrics* 2406.00827.

Jakiela, P. (2021). Simple diagnostics for two-way fixed effects. *arXiv: General Economics* 2103.13229.

Kallus, N. and X. Mao (2020). On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv: Machine Learning* 2003.12408.

Kellogg, M., M. Mogstad, G. A. Pouliot and A. Torgovitsky (2021). Combining matching and synthetic control to tradeoff biases from extrapolation and interpolation. *Journal of the American Statistical Association 116*(536), 1804–16.

Kuosmanen, T., X. Zhou, J. Eskelinen and P. Malo (2021). Design flaw of the synthetic control method. MPRA paper, University Library of Munich, Germany.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review 76*, 604–20.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics 95*(2), 391–413.

Lei, L. and E. J. Candès (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 83*(5), 911–38.

Leung, M. P. (2023). Network cluster-robust inference. *Econometrica 91*(2), 641–67.

Li, K. T. (2023). Frontiers: A simple forward difference-in-differences method. *Marketing Science 43*, 239–468.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Liu, L., Y. Wang and Y. Xu (2024). A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *American Journal of Political Science 68*, 160–76.

Lucas, R. E., Jr. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy 1*, 19–46.

Lynch, J. (1984). Canonical row-column-exchangeable arrays. *Journal of Multivariate Analysis 15*(1), 135–40.

Magnac, T. (2004). Panel binary variables and sufficiency: generalizing conditional logit. *Econometrica 72*(6), 1859–76.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies 60*(3), 531–42.

Manski, C. F. and J. V. Pepper (2018). How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics 100*, 232–44.

Masini, R. and M. C. Medeiros (2021). Counterfactual analysis with artificial controls: Inference, high dimensions, and nonstationarity. *Journal of the American Statistical Association 116*(536), 1773–88.

Masini, R. and M. C. Medeiros (2022). Counterfactual analysis and inference with nonstationary data. *Journal of Business and Economic Statistics 40*, 227–39.

McKay, A. and C. K. Wolf (2023). What can time-series regressions tell us about policy counterfactuals? *Econometrica 91*, 1695–725.

Meyer, B. D., W. K. Viscusi and D. L. Durbin (1995). Workers' compensation and injury duration: Evidence from a natural experiment. *American Economic Review 85*, 322–40.

Moon, H. R. and M. Weidner (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica 83*(4), 1543–79.

Moon, H. R. and M. Weidner (2018). Nuclear norm regularized estimation of panel regression models. *arXiv: Econometrics* 1810.10987.

Mugnier, M. (2022). Make the difference! computationally trivial estimators for grouped fixed effects models. *arXiv: Econometrics* 2203.08879.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica 46*, 69–85.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(2), 331–55.

Nakamura, E. and J. Steinsson (2018). Identification in macroeconomics. *Journal of Economic Perspectives 32*(3), 59–86.

Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*, 1–32.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica 49*, 1417–26.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*(4), 967–1012.

Plümper, T. and V. E. Troeger (2007). Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects. *Political Analysis 15*(2), 124–39.

Rambachan, A. and J. Roth (2020). Design-based uncertainty for quasi-experiments. *arXiv: Econometrics* 2008.00602.

Rambachan, A. and J. Roth (2023). A more credible approach to parallel trends. *Review of Economic Studies 90*, 2555–91.

Ritzwoller, D. M., J. P. Romano and A. M. Shaikh (2024). Randomization inference: Theory and applications. *arXiv: Econometrics* 2406.09521.

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), 1393–512.

Robins, J., M. A. Hernan and B. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology 11*, 550–60.

Rosenbaum, P. R. (2002). Multiple control groups. In *Observational Studies*, 253–75. New York: Springer.

Rosenbaum, P. R. (2023). *Causal Inference*. Cambridge, MA: MIT Press.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Roth, J. and P. H. Sant'Anna (2023). When is parallel trends sensitive to functional form? *Econometrica 91*, 737–47.

Roth, J., P. H. Sant'Anna, A. Bilinski and J. Poe (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics 235*, 2218–44.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics 6*, 34–58.

Rubin, D. B. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science 5*(4), 465–72.

Samartsidis, P., S. R. Seaman, A. M. Presanis, M. Hickman and D. de Angelis (2019). Assessing the causal effect of binary interventions from observational panel data with few treated units. *Statistical Science 34*(3), 486–503.

Sant'Anna, P. H. and J. Zhao (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics 219*(1), 101–22.

Shaikh, A. M. and P. Toulis (2021). Randomization tests in observational studies with staggered adoption of treatment. *Journal of the American Statistical Association 116*(536), 1835–48.

Shen, D., P. Ding, J. Sekhon and B. Yu (2022). A tale of two panel data regressions. *arXiv: Econometrics* 2207.14481.

Shi, X., W. Miao, J. C. Nelson and E. J. T. Tchetgen (2020). Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of The Royal Statistical Society: Series B (Statistical Methodology) 82*, 521–40.

Sofer, T., D. B. Richardson, E. Colicino, J. Schwartz and E. J. T. Tchetgen (2016). On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical Science 31*, 348–61.

Spiess, J., G. Imbens and A. Venugopal (2023). Double and single descent in causal inference with an application to high-dimensional synthetic control. *arXiv: Econometrics* 2305.00700.

Stock, J. H. and M. W. Watson (1998). Diffusion indexes. Working Paper 6702, NBER.

Sun, L. and S. Abraham (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*, 175–99.

Viviano, D. and J. Bradic (2021). Dynamic covariate balancing: estimating treatment effects over time. *arXiv: Econometrics* 2103.01280.

Viviano, D. and J. Bradic (2023). Synthetic learner: Model-free inference on treatments over time. *Journal of Econometrics 234*, 691–713.

Vogelsang, T. J. (2012). Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects. *Journal of Econometrics 166*(2), 303–19.

Wolf, C. K. (2023). The missing intercept: A demand equivalence approach. *American Economic Review 113*(8), 2232–69.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Wooldridge, J. M. (2021). Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators. Working Paper 3906345, SSRN.

Wooldridge, J. M. (2022). Simple approaches to nonlinear difference-in-differences with panel data. Working Paper 4183726, SSRN.

Xiong, R., S. Athey, M. Bayati and G. Imbens (2019). Optimal experimental design for staggered rollouts. *arXiv: Econometrics* 1911.03764.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis 25*(1), 57–76.

Xu, Y. (2023). Causal inference with time-series cross-sectional data: A reflection. Working Paper 3979613, SSRN.

Ying, A., W. Miao, X. Shi and E. J. T. Tchetgen (2021). Proximal causal inference for complex longitudinal studies. *arXiv: Methodology* 2109.07030.

Managing editor Jaap Abbring handled this manuscript.