Causal Models for Longitudinal and Panel Data: A Survey

Dmitry Arkhangelsky & Guido Imbens

November 2023

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

Keywords: Panel Data, Causal Effects, Synthetic Control Methods, Two-Way-Fixed-Effects, Difference-in-Differences

---

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 data settings. This literature has taken some of the elements of the earlier panel data literature and brought in 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) methods 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. Part of the challenge in writing this survey is the organization. We have chosen one that is different from that in the rest of the literature (with the exception of [ABD+21a], but one we felt was helpful for organizing our own thinking. It is clear that one could organize the material differently.

Although we intend to make this survey of interest to empirical researchers, it is not a guide with recommendations for specific cases. Rather, we intend to lay out out views on this literature and present the key features of various methods 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 or not, but as approximations that may be accurate and useful in particular settings. For example, the two-way-fixed-effect set up has been critized 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. Similarly, assuming the absence of dynamic effects is unlikely to ever hold, and similarly treatment effects are almost always heterogeneous. Nevertheless, in many cases fixed effect models with constant treatment effects may be good working models that lead to robust estimates and inferences. Understanding when they do so, and when relaxing them is likely to improve the combination of bias-squared and variance, and in what way to generalize them without sacrificing too much precision is close to what we intend to do here.

After this 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 literatures. We also point out conceptual issues that have been raised in the earlier literature and that have received less attention, or even are in

danger of being entirely ignored, in the current literature.

Next, in Section 3 we discuss three ways of organizing the panel data literature, first by types of data available, *e.g.,* proper panel data, repeated cross-sections, or two-way data structures, second by shape of the data frame, many units or many periods, and third by the assignment for the causal variable of interest. We find this organization useful because it matters for the relevance of various methods that have been proposed. 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 realization that the structure of the assignment matrix 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 unconfoundedness literature has focused largely on the setting with panel data with large N and relatively small T, 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 N and T are comparable in size, and one or few units are treated from some period onwards. The recent DID literature has paid particular attention to the setting with staggered adoption patterns in the assignment. The singular focus of some of these literatures 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 and Two-Way-Fixed-Effect (TWFE) setup as a stepping stone to our following discussion of the recent developments in causal panel data literature in Sections 6 through 10. We see four main threads to the new causal panel literature. First, one strand of the literature has focused on a more general setup where different groups adopt treatments at different points in time. In this staggered adoption case recent research has highlighted some concerns with the standard TWFE. In particular, in cases with general treatment effect heterogeneity, the implicit weights on the building blocks of the TWFE estimator have been argued to be counter-intuitive, and alternatives have been proposed. We discuss this literature in Section 6.

Second, the literature has generalized the popular two-way additive fixed effect structure. One direction has allowed for factor models, whereas another strand of the literature generalized the models to allow for nonlinear structures. An important part of the first extension that allows for factor structures is the synthetic control approach developed in a series of papers by Alberto Abadie and coauthors [ADH10]. Although

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

this literature shares features with the TWFE and DID literature, it has partly developed separately, ignoring some of the gains arising from combining the insights from each of them. We discuss this literature in Section 7. The second generalization, allowing for non linear models, is discussed in Section 8.

Fourth, 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. We discuss this literature in Section 9.

Finally, in Section 10 we discuss open questions which we view as exciting avenues for future research.

There are some excellent recent surveys of the new DID and causal panel data literature. They differ in their focus, and in the perspectives of the authors, and complement ours in various ways. Some of these surveys [dCD22, RSBP23] focus primarily on the DID setting with heterogenous treatment effects, with less attention paid to 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, [Aba19] focuses primarily on synthetic control methods. In the current survey we stress deeper linkages between these ideas as well as potential benefits from combining them. Recent surveys in the political science literature, [Xu23, LWX22], like the current survey, also discusses the connections between synthetic control, unconfoundedness, and TWFE approaches.

In this discussion we use a number of acronyms. We list those in Table 1.

2.   The Econometrics Panel Data Literature

Although the new panel literature focuses on different estimands and settings, and emphasizes different concerns about internal and external validity, many of the methods

overlap with 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 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 this entire literature, partly because that is a vast literature in itself, but mainly because there are many excellent surveys and textbooks, including [Are03, Bal08, Hsi22, Woo10, AB11b].

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 [Cha82, Cha84, HSCKW12, Are03, AH01, Hsi22, Bal08, Woo10, AB11b]. This literature was initially motivated by the increased availability of various large public-use 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 is primarily a repeated cross-section data set with some short term panel features. 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 [Gol91]. Part of the literature analyzed fully parametric models, but more often semiparametric settings were considered. The parameters of interest could be causal in the modern sense, but the term 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, fixed effects that were conditioned on in the analyses and were modeled as unrestricted in their correlation with other variables, *versus* random effects that were treated as stochastic and often uncorrelated with observed covariates (though not always, see the correlated random effects discussion in [Cha84]). See for general dsicussions [Hsi22, BJ15]. This distinction was often used as an organizing 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 labeled the incidental parameter problem [NS48]. Especially when the fixed effects entered in in non-additive and non-linear ways in short (fixed length) panels, with limited dependent or discrete outcomes, this led to challenging identification

problems *e.g.,* [Cha80, Mag04, Hon92]. In cases where identification in fixed *T* settings was not feasible the literature introduced various methods for bias-correction (see [AH07b] for a survey), or developed bounds analyses (*e.g.*, [HT06]). More recently, these bias-reduction ideas were extended to nonlinear two-way models [FVW16, FVW18].

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 labor supply. Motivated by these questions this literature distinguished between state dependence and unobserved heterogeneity (*e.g.*, [Hec81, Cha84]) and various dynamic forms of exogeneity (*e.g.*, weak, strong and strict exogeneity, and predeterminedness, see [EHR83, AB91]), issues that 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 and the use of instrumental variable approaches [Nic81, AB91, BB98, HK02, AA03]. This literature had a huge impact on empirical work in social sciences.

In contrast, an important theme in the current literature is allowing for general heterogeneity in causal effects, both over time and across units, associated with observed as well as unobserved characteristics. Much of the earlier literature often restricted the heterogeneity in treatment effects to parametric forms or assumed its absence entirely. Exceptions include [Cha92, AB11a, GP12, CFVHN13], which emphasized heterogeneity in a way that is more in line with the current literature. The recognition of the importance of heterogeneity has led to findings that previously popular estimators are sensitive to the presence of heterogeneity and the development of more robust alternatives.

3.   Set Up and Data Configurations

In this section, we consider different types of data as well as two classifications of panel data configurations. One of these classifications, in terms of the relative size of the cross-section and time-series dimensions, is familiar from the earlier literature, but the second, 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 modeled as fixed effects or random effects, *e.g.,* [Cha84]. This distinction plays less of a role in the current literature.

5

Both these 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 in context.

To put the following discussions into context, it is also helpful to keep in mind that in most cases we are interested in the average causal effect of some intervention on the outcomes for the treated units, during the periods they were treated. 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 the interpretation of those estimands.

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 grouped repeated cross-section settings. 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. Most of the time 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 where we observe units for different lengths of time, or from missing data. We ignore both complications in the current discussion.

Classic examples of this proper panel setting include [Ash78] with information on earnings for over 90,000 individuals for 11 years, and [AC86] with information on wages for 1,448 individuals also for 11 years. Another classic example is [CK94] with data for two periods and 399 fast food restaurants.

3.1.2. Grouped Repeated Cross-section Data

In a Grouped Repeated Cross-section Data (GRCD) setting we have observations on N units. Unit i is observed only once, in period T_i, which is 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), *e.g.,* just two or three, with many units sharing the same value for T_i, although 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.[1] 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:

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

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 as pointed out in [Dea85] and [Woo10]. Many methods in this 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 the settings 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

---

[1]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 this repeated cross-section case.

groups and two periods, [MVD95], 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 [EL96] the units are different taxpayers in two different years, with the number of groups again equal to two. The case with multiple groups is studied in [BDM04], 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 panel or repeated cross-section data corresponds to what we refer to as row-column exchangeable data (RCED) [Ald81, Lyn84]. 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 panel data is that there is no ordering for the second index (time in the 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 rideshare company where we observe outcomes for trips involving customer i and driver j. Although this is not a particularly common data configuration, it is useful to discuss it explicitly to see that panel data differ in two aspects from cross-section data: the double indexing and the time ordering.

In this case it is more plausible to model the units $i = 1, \ldots, N$ as exchangeable, but also model the units $j = 1, \ldots, J$ as exchangeable, whereas with proper panel data the exchangeability of the time periods is typically implausible. Many, but not all, methods ostensibly developed for use with panel data are also applicable in this RCED setting, and this is an important reason to distinguish between the two types of data. For example, TFWE 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, these methods should be used with caution in proper panel data settings.

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 [AKM99, CRY22]. See [Bon20] for a recent survey of the relevant methods.

3.2. Shapes of Data Frames

One classification the panel data literature can be organized 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 depends on the magnitude of the cross-section and time-series correlations. Nevertheless, it is useful to reflect on the relative magnitude of the cross-section and time-series dimensions as it has implications for the relative merits 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. See [AA03] for a discussion of the properties of various panel data estimators under different asymptotic sequences.

3.2.1. Many Units, Few Time Periods: $N >> T$

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:

$$
\mathbf{Y}^{\text{thin}} = \begin{pmatrix} Y_{11} & Y_{12} & Y_{13} \\ Y_{23} & Y_{23} & Y_{23} \\ Y_{33} & Y_{33} & Y_{33} \\ Y_{43} & Y_{43} & Y_{43} \\ Y_{53} & Y_{53} & Y_{53} \\ Y_{63} & Y_{63} & Y_{63} \\ \vdots & \vdots & \vdots \\ Y_{N3} & Y_{N3} & Y_{N3} \end{pmatrix}
$$

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 PSID and 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 [NS48] are relevant (see [Lan00] 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.,* [Hon92, Cha10, Bon12]. However,

the fact that the time series dimension is small or modest does mean that random effect assumptions are very powerful because they place a stochastic structure on the individual components so that these individual components can be integrated out .

3.2.2. Few Units, Many Time Periods: $N << T$

The second setting is one where the number of time periods is large relative to the number of cross-section units:

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

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, and 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.,* [BGK$^+$15, BMAF$^+$23]. The work on inference using conformal methods is also in this spirit, [CWZ21].

3.2.3. 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.

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

A common example is that were 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, 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 more 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 depend also 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 [AK91] the authors 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 [Bek94] 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).

3.3. Assignment Mechanisms

The second 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.

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:

(1)
$$\mathbf{W}^{\mathrm{gen}} = \text{(general)} \begin{pmatrix} 1 & 1 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & 1 & \ldots & 0 \\ 1 & 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 1 & 0 & \ldots & 0 \end{pmatrix}$$

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 be important for credibly estimating causal effects. This setting is particularly relevant for the row-column exchangeable data configuration. It is not very 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 of dynamic treatment effects are crtical, and in settings where such assumptions are not plausible many of the commonly used methods may 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.

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

In settings where the number of time periods is relatively small, this case is often analyzed 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 [RR83]. A classic example in the economics literature is the Lalonde-Dehejia-Wabha data originally collected in [LaL86] with the data set now commonly used first analyzed in [DW99]. In that case there are three periods of outcome data (earnings), but only in the last period are some units treated. The original study [LaL86] reported results for a variety of models, including some two-way-fixed-effect regressions. Much of the subsequent literature since [DW99, DW02], however, has focused more narrowly on methods relying on unconfoundedness, sometimes in combination with functional form assumptions. Asymptotics are typically based on large N and fixed T.

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

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

This setting is prominent in the original applications of the synthetic control literature [AG03, ADH10, Aba19]. 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.

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:

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

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 after a common starting date:

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

This assignment matrix is the basis of the simulations reported in [BDM04, AAH⁺21]. In this case there is often 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 difference-in-differences 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, with some early adopters, and some later adopters:

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

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.

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.

$$\mathbf{W}^{\text{event}} = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & \dots & 0 & 0 \end{pmatrix}$$

(event study)

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.,* [FFJR69]

3.3.8. Clustered Assignment

Finally, in many applications units are grouped together in clusters, with units within the same the 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.

$$
\mathbf{W}^{\text{cluster}} =
\text{(cluster/block)}
\begin{pmatrix}
 & & & & & & & \text{cluster} \\
 & & & & & & & \downarrow \\
0 & 0 & 0 & 0 & \ldots & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & \ldots & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & \ldots & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & \ldots & 0 & 0 & 2 \\
0 & 0 & 0 & 0 & \ldots & 0 & 0 & 2 \\
0 & 0 & 0 & 1 & \ldots & 1 & 1 & 3 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 1 & \ldots & 1 & 1 & C \\
0 & 0 & 0 & 1 & \ldots & 1 & 1 & C
\end{pmatrix}
$$

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 panel data case with N units and T periods. We use the potential outcome notation [Rub74, IR15]. 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 vector of treatment assignments,

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

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

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

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 [Rub78, IR15]. 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 [HH08, AS17].

Without further restrictions, this setup describes for each unit 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 $\underline{\mathbf{w}}'$:

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

with the corresponding average effect (averaged over all units) defined as

$$\tau_{t}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'} \equiv \frac{1}{N} \sum_{i=1}^{N} \Big(Y_{it}(\underline{\mathbf{w}}') - Y_{it}(\underline{\mathbf{w}}) \Big).$$

These unit-level and average causal effects are the basic building blocks of many of the estimands considered in the literature.

If all we are interested in is average causal effects of the form $\tau_{t}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'}$, 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 analyze such problems using standard methods for multi-valued treatments under unconfoundedness, *e.g.*, [Imb02]. Here this implies we could compare outcomes in period t for units with treatment vectors $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}'$.

If we have completely random assignment, all $\tau_{t}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'}$ are identified given sufficient variation in the treatment paths. That also means 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 for of being exposed to the sequence $(1, 1)$ rather than the sequence $(0, 1)$, so that measures the dynamic effect of being exposed to

18

the treatment in period one on period 2 outcomes for units exposed to the treatment 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 $\tau_t^{\mathbf{w},\mathbf{w}'}$. Even with $T = 2$ this means there are 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 $Y_{it}(\mathbf{w})$. That enables us to give comparisons of outcomes from different time periods a causal interpretation. See [Cha84] for a discussion of this in the case of linear models. Note that without additional restrictions all the average treatment effects $\tau_t^{\mathbf{w},\mathbf{w}'}$ 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 [AI21, CS20, SA20].

ASSUMPTION 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 $\underline{\mathbf{w}}^t = \underline{\mathbf{w}}'$.

With experimental data, and sufficient variation in treatment paths, this assumption is testable. We can compare outcomes in period t for units with the same treatment path up to and including t, but whose treatments paths diverge in the future, that is, after period t.

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. In observational studies of course the assumption need not hold. In many applications treatments are state-level regulations that are anticipated some time prior to the time they formally take effect. One remedy for this problem is to impose limited anticipation, assuming the treatment can be anticipated for a fixed number of periods, as in [CS20]. Algorithmically, this amounts to redefining the variable $\underline{\mathbf{w}}$ by shifting it by the fixed number of periods.

At the same time, there are numerous economic applications where units are involved into active decision-making process. This process can be about $\underline{\mathbf{w}}$ itself, e.g., firms deciding on the adoption of a new technology, or students deciding on exiting schools. Alternatively, agents can decide on other variables for which $\underline{\mathbf{w}}$ is an important input, e.g., current and future tax changes can be important determinants of the current consumption. In such cases, no anticipation assumption cannot be guaranteed by design, because units can change their behavior in response to different experimental designs [LJ76]. See [AVdB03] for a related discussion of the no anticipation assumption, and [HN07] for a distinction between this assumption for structural and reduced-form models in the context of staggered adoption designs.

This 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 one 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 modeled by putting a particular structure on them, but in principle, one can estimate them without imposing additional restrictions on the potential outcomes given assumptions on the assignment mechanism, such as random assignment [BRS21]. There is also a large literature in biostatistics on dynamic models that is relevant for these problems, *e.g.,* [RHB00].

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

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

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

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

ASSUMPTION 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 $w_{it} = w'_{it}$.

Like the no-anticipation assumption this is not a design assumption that can be guaranteed by randomization. It restricts the treatment effects, and thus the potential

outcomes for the post-treatment periods. Again like the no-anticipation assumption it has testable restrictions given random assignment of the treatment, and given 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, which can exhibit arbitrary correlations.

If we are willing to make the no-dynamics 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

$$(4) \qquad\qquad \tau_{it} \equiv Y_{it}(1) - Y_{it}(0),$$

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 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 first treatment, $A_i \equiv T + 1 - \sum_{t=1}^{T} W_{it}$. 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 $a = 1, \ldots, T$, and the realized outcome as $Y_{it} = Y_{it}(A_i)$.

There are two broad classes of settings that are sometimes viewed as staggered adoption designs but that are substantially different. 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 [AG04], and counties adopted enhanced 911 policies at different times [AS02]. Second, there may be one-time interventions that have a long-term or even permanent impact. We refer to such settings as event studies. In that case the post intervention period effects would be dynamic effects.

Given staggered adoption, but without the no-anticipation and no-dynamics assump-

tions, we can write the building blocks as

$$(5) \qquad \tau_{it}^{a,a'} \equiv Y_{it}(a') - Y_{it}(a),$$

with the corresponding average

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

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

$$\tau_{t|a''}^{a,a'} \equiv \frac{1}{N_{a''}} \sum_{i:A_i=a''} \left(Y_{it}(a') - Y_{it}(a) \right),$$

where $N_a \equiv \sum_{i=1}^{N} \mathbf{1}_{A_i=a}$ is the number of units with adoption date a. 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.

In the two period case where all units are exposed to the control treatment the estimand $\tau_{t|1}^{0,1}$, the average effect of the treatment in the second period for those who 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 Estimators and Difference-In-Differences

In this section we give a brief introduction to conventional Difference-In-Differences (DID) estimation. 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 [CLLX23, dCD22, RSBP23].

5.1. The Two-Way-Fixed-Effect Characterization

We start with the Two-Way-Fixed-Effect (TWFE) specification in a proper panel setting with no anticipation and no dynamics and constant treatment effects. Traditionally this specification often motivates the DID estimator.

ASSUMPTION 4. (THE TWO-WAY-FIXED-EFFECT MODEL) *The control outcome $Y_{it}(0)$ satisfies*

$$Y_{it}(0) = \alpha_i + \beta_t + \varepsilon_{it}. \tag{6}$$

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

ASSUMPTION 5. (CONSTANT STATIC TREATMENT EFFECTS) *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 realized outcome, $Y_{it} = W_{it}Y_{it}(1) + (1 - W_{it})Y_{it}(0)$,

$$Y_{it} = \alpha_i + \beta_t + \tau W_{it} + \varepsilon_{it}. \tag{7}$$

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

$$(\hat{\tau}^{\text{TWFE}}, \hat{\alpha}, \hat{\beta}) = \arg \min_{\tau, \alpha, \beta} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \alpha_i - \beta_t - \tau W_{it}\right)^2. \tag{8}$$

Here we need to impose one restriction on the α_i or β_t to avoid perfect collinearity, but this normalization 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 $N^{\text{co}} = N - N^{\text{tr}}$), and for those units only for $t > T_0$ ("post-treatment"). Defining the averages in the four groups as

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

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

we can write the estimator for the treatment effect as

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

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

It is convenient to use the TWFE characterization based on least squares estimation

of the regression function in (7) because this characterization also applies in settings where the estimator does not have the double difference form.

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 blocked assignment. To reflect this our notation now only has a single index for the unit, $i = 1, \ldots, N$. Let $G_i \in \mathcal{G} = \{1, \ldots, G\}$ denote the cluster or group unit i belongs to, and $T_i \in \{1, \ldots, T\}$ the time period unit i is observed in. The set of clusters $\mathcal{G}$ is partitioned into two groups, $\mathcal{G}_C$ and a treatment group $\mathcal{G}_T$, with cardinality G_C and G_T.

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 $W_i = \mathbf{1}_{G_i \in \mathcal{G}_T, T_i > T_0}$. Let $D_i = \mathbf{1}_{G_i \in \mathcal{G}_T}$ be an indicator that unit i is in one of the treated groups, irrespective of whether the unit is observed in the post-treatment period.

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

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

$$-\frac{1}{G_T T_0} \sum_{g \in \mathcal{G}_T, t \leq T_0} \overline{Y}_{gt} + \frac{1}{G_C T_0} \sum_{g \in \mathcal{G}_C, t \leq T_0} \overline{Y}_{gt}$$

Alternatively we can use the TWFE specification at the group level where we do have a proper panel set up:

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

similar to that in (6). 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.

5.3. Inference

To conduct inference about $\hat{\tau}^{\mathrm{DID}}$ or $\hat{\tau}^{\mathrm{TWFE}}$ we need to be explicit about the sampling scheme. In the proper panel setting one often assumes that the units are randomly sampled from a large population, and thus independent across each other. In this case, inference about $\hat{\tau}^{\mathrm{TWFE}}$ reduces to joint inference about four means with i.i.d. observations. This approach was used by [Car94] to quantify the uncertainty about the estimated effect of the minimum wage.

With GRCS data the situation can be different, and instead of dealing with uncertainty at the unit-level, one can allow for non-vanishing errors at the group level, as in the model in Equation (9). This clustering approach was advocated in [LZ86, Are87, BDM04], and is routinely used in situations where the number of groups is large. See [AAIW23] for a recent discussion in a design setting. Note that clustering adjustments to standard errors is infeasible in the two group case, as in [Car94].

Regardless of the level of aggregation, since [BDM04] the inference for TWFE/DID estimators has typically taken into account correlation in outcomes over time within units. This implies that if one estimates the average treatment effect by least squares regression for (8), it is not appropriate to use the robust, Eicker-Huber-White standard errors. Instead one can use clustered standard errors [LZ86, Are87], based on clustering observations by units. The appropriate standard errors can also be approximated by bootstrapping all observations for each unit. [Han07] discusses a more general hierarchical set up.

5.4. The Parallel Trend Assumption

The fundamental motivation for the DID estimator, in one form or another, is based on a parallel trend assumption that, in one form or another, the units who 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 set up, 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 assumption is that the expected difference in control outcomes in any period for units who later are exposed to the treatment, and units who are always

in the control group is constant:

ASSUMPTION 6. *For all t, t',*

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

We can also 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 who will eventually be exposed to the treatment and those who 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 (7), with the additional assumption that the treatment assignment D_i is independent of the vector of residuals ε_{it}, $t = 1, \ldots, T$ conditional on fixed effects:

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

as in, for example, [Are03]. From the point of view of the causal inference literature, the parallel trend assumption is somewhat non-standard because it combines restrictions on the potential outcomes through the TWFE specification with restrictions on the assignment mechanism.

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 7. *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:*

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

Alternatively we could formulate this 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 group and time period exposed to the treatment, there are not testable restrictions implied by this assumption in the two group / two period case.

Because we can view the panel case as a two-group setting, 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 [AC85] argued against using the TWFE model for evaluation of training programs based on the observed failure of parallel trends. See [RSBP23, RR23, Jak21] 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 [RS23, GSW22] show how parallel trends can be implied by random assignment of treatment. They also discuss the sensitivity to transformations of the parallel trend assumption.

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 (7) because they would be perfectly collinear with the individual components α_i [PT07]. Nevertheless, they can be important by relaxing 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.

5.5.1. Abadie (2005)

In an early paper [Aba05] proposes flexible ways of adjusting for time-invariant covariates, while continuing with the conditional parallel trends assumption. His solution was based on re-weighting the differences in outcomes by the propensity score.

5.5.2. Sant'Anna and Zhao (2020)

[SZ20] use recent advances in the cross-section causal literature to adjust for time-invariant covariates in a more robust way, by combining inverse-propensity score weighting with outcome modeling. In cross-sectional settings, such doubly robust methods have been found to be more attractive than either outcome modeling or inverse-propensity-score weighting on their own. They do maintain the parallel trends assumption conditional on covariates.

5.6. Unconfoundedness

One key distinction between the repeated cross-section and proper panel case is that in the case with proper panel data there is a natural alternative to the TWFE model. 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:

$$(12) \qquad D_i \perp\!\!\!\perp \left(Y_{iT}(0), Y_{iT}(1) \right) \mid Y_{i1}, \ldots, Y_{iT-1}.$$

If one is willing to make this assumption the larger lierature on estimation of treatment effects under unconfoundedness applies. See [Imb04] for a survey. Modern methods include doubly robust methods that combine modelling outcomes with propensity score weighting. See, for example, [BR05, CCD$^+$17, AIW18].

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

$$D_i \perp\!\!\!\perp \left(Y_{i2}(0) - Y_{i1}, Y_{i2}(1) - Y_{i1} \right) \mid Y_{i1},$$

it follows that the issue in the choice between DID/TWFE and unconfoundedness is really whether one should adjust for differences between treated and control units in the lagged outcome, Y_{i1}. The DID/TWFE approach implies one should not, and that doing so would 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 motivate their choices clearly 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, on the other hand, is satisfied when the selection is based on past outcomes and potentially other observed pre-treatment variables.

Methodological literature does not provide a lot of guidance on the choice between these two strategies, with exceptions in [AP08, Xu23], and is somewhat segmented. For example, there is a large literature re-analyzing the data originally studied in [LaL86] (see also [DW99]) where the researcher has observations on two lagged outcomes. Although Laonde 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 reanalyzing the data originally studied in [CK94] 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 early study [AC85] point out the limitations of the TWFE model and in particular its inability to capture temporary declines in earnings prior to enrollment in labor market programs, the so-called Ashenfelter dip.

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, than the two potential control groups deliver the same results. See for more on this perspective [Ros02].

When the control group changes over time, and the control group and treatment group differ in the initial period, then the TWFE and unconfoundedness approach give different results. However, the differences can be bounded [AP08, DL19]. Suppose the true average effect is positive, and suppose unconfoundedness holds. Then the TWFE estimator will over-estimate the true effect. On the other hand, if the effect is positive, and the TWFE model holds, then assuming unconfoundedness and adjusting for the lagged outcome will under-estimate the true effect.

[IKW21] takes 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.

6. The Staggered Adoption Case

Although much of the DID literature starts with the TWFE model for control outcomes in (Assumption 4) with constant treatments effects (Assumption 5), 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/DID estimator continues to estimate a well-defined average causal effect, namely for the average treatment effect for the treated,

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

This robustness to treatment effect heterogeneity does not extend to settings outside of block assignment. Part of the new DID 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 TFWE estimators in this setting that is common in empirical work (see [dCD22]), and proposing modifications that ensure that a meaningful average causal effect is estimated. We review that literature here. Recall the staggered adoption setting,

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

Let $A_i = \arg\min\{t = 1, \dots, T | W_{it} = 1\} = T + 1 - \sum_{t=1}^{T} W_{it}$ be the adoption date (the first time unit i is treated), with $A_i = \infty$ for units who never adopt the treatment, and let N_a be the number of units with adoption date $A_i = a$. Define also the average treatment

effect by time and adoption date,

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

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 [Cha92] and more recently in [AB11a, GP12, CFVHN13], which allow for heterogeneity over units, but not over time.

6.1. Decompositions of the TWFE Estimator

Here we discuss the interpretation of the TWFE estimator 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, following the discussion in [GB21]. We maintain the TWFE structure for the control outcomes in (6), as well as the mean-independence of the residuals and the treatment group 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 $t' < a \leq t$ and either $a' \leq t'$ or $t < a'$, define the building block for the TWFE estimator:

(13)
$$\hat{\tau}_{t,t'}^{a,a'} \equiv \left(\overline{Y}_{t|a} - \overline{Y}_{t'|a} \right) - \left(\overline{Y}_{t|a'} - \overline{Y}_{t'|a'} \right)$$

The interpretation of this building block plays a key role. 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:

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

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. [RSBP23] 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 (13) has expectation

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

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. 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, researchers may be reluctant to report weighted averages with negative weights. Note that the concern with the comparisons $\hat{\tau}_{t,t'}^{a,a'}$ when $a' < t'$ but not when $a' > t$ fundamentally treats the treated state and the control state as different: the parallel trends assumption is assumed for the control outcomes, but not for the treated outcomes.

The TWFE estimator $\hat{\tau}^{\text{TWFE}}$ can be characterized as a linear combination of the building blocks $\hat{\tau}_{t,t'}^{a,a'}$, including those where the non-changing group has an early adoption date $a' < t'$. The coefficients in that linear combination depend various aspects of the data, including the number of units N_a in each of the corresponding adoption groups, as discussed in detail in [GB21, BLW21] for the staggered adoption case and in [IK21] 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 nonnegative 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, given two identical populations in terms of potential outcome distributions that, for some reason, have different adoption patterns, one would estimate different quantities.

We emphasize 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 randomized 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. We can also immediately see that this problem does not arise if $\tau_{it} \equiv \tau_i$, which was assumed in [Cha92] and the follow-up literature.

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 set up tradeoffs with other assumptions, including the parallel trend assumptions, that may force the researcher to make some assumption that are at best approximations. Similar tradeoffs motivate the use of higher order kenels which also lead to estimators with negative weights. We therefore do not view the negative weights of some estimators as necessarily disqualifying and 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 widespread attention. It should be noted that all maintain the TWFE assumption for the control outcomes, and all four avoid additional assumption on treatment effect heterogeneity.

6.2.1. Callaway and Sant'Anna (2020)

[CS20] propose two ways of dealing with the negative weights. The first compares outcomes for a group that adopts the treatment in any period after the adoption ($\overline{Y}_{t,a}$ for $t \geq a$) to average outcomes for the same group immediately prior to the adoption ($\overline{Y}_{a-1,a}$, and subtracts the difference in outcomes for the same two time periods for the group that never adopts the treatment. Formally, consider, for $t \geq a$, the double difference

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

The particular control group, those who never adopt the treatment, may not be particularly attractive. In some cases one would be concerned that the 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 there may be very few of these never-adopters, especially in long panels, so that the precision of the natural estimators for such estimands may make them unattractive.

Recognizing this concern [CS20] 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:

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

Given these two estimators, [CS20] suggest reporting averages 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 e periods after adoption, for their two control groups,

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

We should note that [CS20] 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. See [CS20] for details.

6.2.2. Sun and Abraham (2020)

[SA20] propose a different way of dealing with the negative weights by comparing the change for a group that adopted in period a between the period just before the adoption $a-1$ and an arbitrary later period t, $\overline{Y}_{t,a} - \overline{Y}_{a-1,a}$, relative to the change for the never-

adopters over the same period:

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

Given double differences of this type they suggest reporting the average of this:

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

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 important additional issue emphasized by [SA20] 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. [SA20] 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

[DCd20] deal with the negative weights by focusing on one-period ahead 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:

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

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

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

6.3. Borusyak, Jaravel, and Spiess, 2021

[BJS21] 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 $X_{it} \equiv (\{t = 1\}, \ldots, \{t = T\})$. 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 treated units:

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

These unit-specific estimators can then be aggregated into an estimator for the target of interest, let us call the estimator $\hat{\tau}^{\text{BJS}}$. Notably, despite each unit-time specific treatment effect estimator $\hat{\tau}_{it}$ being inconsist, after these objects are averaged the estimator behaves well. Moreover, [BJS21] show that the resulting estimator is efficient as long as ϵ_{it} are i.i.d. over i and t, which relies on a version of Gauss-Markov theorem for their setup.

6.3.1. 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, $\hat{\tau}^{\text{CS,I}}$ (or $\hat{\tau}^{\text{CS,II}}$), $\hat{\tau}^{\text{SA}}$, $\hat{\tau}^{\text{CH}}$, and $\hat{\tau}^{\text{BJS}}$? There are a couple of substantive arguments that matter for this choice: (*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 multiple periods apart if there are dynamic effects, and (*iv*) efficiency considerations. These concerns do not lead to one of the proposals clearly dominating the others, and in practice, looking for a single estimator may be the wrong goal.

What should one to instead? One option is to report all of the proposed estimators, as for example, [BLM22b] who reports 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 different assumptions, in particular about treatment effect

heterogeneity, and focus on different estimands. Moroever, some of these comparisons may have little power in terms of uncovering heterogeneity of particular forms. Instead of reporting all estimators, one may therefore wish to explore directly the presence of systematic variation in the $\hat{\tau}_{t,s}^{a,a'}$, by adoption date, a, by the length of the period between before and after, $t - s$, and the time since adoption, $t - a$.

7. Relaxing 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 coauthors, ininitially in [AG03], with more detailed methodological discussions in [ADH10] and [ADH15]. This brought a substantially different perspective to the questions studied in the DID/TWFE literature. Initially the SC literature remained very separate from the earlier DID/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 DID/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 direct estimation of factor models. In Section 7.3 we discuss some hybrid methods that combine synthetic control and DID/TWFE components.

7.1. Synthetic Control Methods

In the original paper, [AG03], Abadie and Gardeazabal 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 synthetic Basque region, with the weights 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 very 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 [ADH10, ADH15]. For a recent review see [Aba19, SSP$^+$19].

7.1.1. Estimation

Here we use a slightly different characterization of the SC method as a least squares estimator, as discussed in [DI16], focusing 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

$$(14) \qquad \hat{\omega} \equiv \arg \min_{\omega | \omega \geq 0, \sum_j \omega_j = 1} \sum_{t=1}^{T-1} \left(Y_{Nt} - \sum_{j=1}^{N-1} \omega_j Y_{jt} \right)^2 ,$$

and then impute the missing potential outcome as

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

The nonnegative weights ω_j define the "synthetic" control that gave the methods its name. One remarkable finding in the initial papers by Abadie and coauthors is that this solution is typically sparse, with positive weights $\omega_j > 0$ for only 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 re-unification application in [ADH15] where the full set of potential controls consists of sixteen OECD countries, only five countries, Austria, Japan, The Netherlands, Switzerland and the US, have positive weights.

The characterization of the SC estimator in (14) allows for an interesting comparison with methods based on unconfoundedness assumption discussed in Section 5.6. With a linear model specification unconfoundedness would suggest an estimator

$$(15) \qquad \hat{\beta} = \arg \min_{\beta} \sum_{i=1}^{N-1} \left(Y_{iT} - \beta_0 - \sum_{s=1}^{T-1} \beta_s Y_{is} \right)^2 ,$$

and then impute 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 [ABD⁺21b], SC in the regression in (14) relies on *vertical regression* with $T - 1$ observations and $N - 1$ predictors, with some restrictions on the parameters, and (15) relies on *horizontal regression* with $N - 1$ observations and T regressors (by including an intercept). If the minimizers in these least squares regressions are not unique, we take the solution to be the one that minimizes the L_2 norm [SIV23]. See [SDSY22] for more insights into the comparison between the horizontal and vertical regressions in this setting.

One interesting aspect of the synthetic control approach is that it is more algorithmic that many other methods used in these settings. Consider the estimator based on unconfoundedness in (15). 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 contemparenous outcomes for other units. However, such a model is never 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 [ADH10], 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 the other states. Attempts to specify models that motivate the synthetic control estimator have met with limited success. [ADH10] discusses factor models as data generating processses, but that begs the question why one would not directly estimate the factor model. Researchers have done so, as discussed in Secction 7.2 below, but such attempts have not always outperformed synthetic control methods. See the review in [Aba19] for more discussion on conditions under which synthetic control methods are appropriate.

7.1.2. Modifications

A number of modifications have been suggested to the basic version of the SC estimator. [DI16] and [FP21] 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).

[DI16] and [AL21] are concerned with settings where the number of potential control units, $N - 1$ is large relative to the number of time periods that is used to estimate the weights. [DI16] suggest regularizing the weights by imposing an elastic net penalty on the weights ω_i, with the penalty chosen by cross-validation. [SIV23] 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 within sample fit, in the spirit of the recent double descent literature [BHMM19]. [AL21] recognizes that weights that a convex combination of control units that are all far away from the treated unit is 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 minimizing the sum of the original synthetic control criterium and a term that penalizes the distance between any of the control units and the target unit

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

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

Typically in synthetic control methods 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 [KZEM21].

[KMPT21] 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. Inference

Inference has been a major challenge in synthetic control settings, and there is as of yet not 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.

One approach has been to use placebo methods to test sharp null hypotheses, typically for the null hypothesis of no effect of the intervention. [ADH10] proposes such a

method. Suppose there is a single treated unit, say unit N. [ADH10] construct a distribution of estimates based on each unit in term being analyzed as the treated unit, and then calculating the p-value for unit N as the quantile in that distribution of placebo estimates.

[DI16] suggests 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 randomization distribution of estimates. Closely related [CWZ21] builds on [LC20] to develop conformal inference procedures that rely on exchangeability of the residuals from some model over time. [CFT21] does not rely on large samples of treated units by proposing the construction of prediction intervals for the counteractual 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

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

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 the TWFE specification, but it obviously allows for more general dependence structures in the data. Although these factor models have a long tradition in panel data, *e.g.,* [And84, CR83, SW98, BN02, Bai09], the recent causal literature has used them in different ways.

7.2.1. Matrix Completion with Nuclear Norm Regularization

[ABD$^+$21b] 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 to keep these fixed effect components in the specification separately from the low-rank component $\mathbf{L}$. The reason is that we regularize the low rank component $\mathbf{L}$,

but not the individual and time components. The parameters $\mathbf{L}$, α and β are estimated by minimizing

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

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 crossvalidation. The nuclear norm regularization 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

[ASS18] focusing 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 [ASS18] use the part of this rescaled matrix corresponding to the control units, in combination with the pre-treatment-period values of for 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 [ASS18] do not impose the convexity restrictions on the weights, but do add a regularization penalty.

7.2.3. Interactive Fixed Effect or Factor Models

Building on the factor model literature in econometrics [CR83, Pes06, Bai09, MW15, MW18, Fre18] researchers have studied direct estimation of factor models as an alternative to synthetic control methods [Xu17]. The basic set up models the control potential outcome as

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

where the ε_{it} are independent of the factors and loadings. Based on this model on 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 [GM16] for an application.

7.2.4. Grouped Panel Data

[BM15, BLM22a] consider a factor model but impose a group structure. In our causal setting, their set up 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. [Mug22, CM22] suggest alternative estimation methods that are computationally more attractive.

7.2.5. Tuning

One disadvantage the methods discussed in this section share is the need to specify the tuning parameters. This sets them apart from the conventional two-way methods we discussed before and makes them harder to adopt in practice. In the case of the matrix completion, this tuning parameter is the regularization 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 finite 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 generalize the two-way methodology and do not require explicit tuning. One such proposal is [MW18], where the authors analyze the limiting version of the estimator from [ABD+21b] with λ approaching zero. They show that the resulting estimator is consistent under relatively weak assumptions, albeit it 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/DID ideas or with unconfoundedness 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 one has calculated the SC weights, the SC estimator for the treatment effect can be characterized as a weighted least squares regression,

$$(16) \qquad \min_{\beta, \tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \omega_i \left(Y_{it} - \beta_t - \tau W_{it} \right)^2 .$$

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

$$(17) \qquad \min_{\beta, \alpha, \tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \alpha_i - \beta_t - \tau W_{it} \right)^2 .$$

The two differences are that (i), the SC regression in (16) uses weights ω_i, and (ii) the TWFE regression in (17) has unit-specific fixed effects α_i.

In the light of of this comparison the omission of the unit fixed effects from the synthetic control regression may seem surprising. [AAH$^+$21] 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 ω_i, as well as analogous time weights λ_t, leading to

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

The time weights λ_t are calculated in a way similar to the unit weights,

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

subjet to the restriction that $\lambda_s \geq 0$, and $\sum_{s=1}^{T-1} \lambda_s = 1$.

7.3.2. Augmented Synthetic Control

[BMFR21] augment the SC estimator using a linear regression of the treated period outcomes on the lagged outcomes for the control units. Following [BMFR21] we use 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:

$$\hat{Y}_{NT} = \hat{\eta}_0 + \sum_{s=1}^{T-1} \hat{\eta}_s Y_{Ns} + \sum_{i=1}^{N-1} \omega_i \left(Y_{iT} - \hat{\eta}_0 - \sum_{s=1}^{T-1} \hat{\eta}_s Y_{is} \right)$$

$$= \sum_{i=1}^{N-1} \omega_i Y_{iT} + \sum_{s=1}^{T-1} \hat{\eta}_s \left(Y_{Ns} - \sum_{j=1}^{N-1} \omega_j Y_{js} \right).$$

[BMFR22] 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, [SDSY22] 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 $\hat{Y}_{NT} = \hat{\gamma}_0 + \sum_{s=1}^{T-1} \hat{\gamma}_s Y_{Ns}$. [SDSY22] show that if we drop the intercept from this regression, $\gamma_0 = 0$, and we impose the restrictions that the γ_s are nonnegative and sum to one, then the unconfoundedness imputation is identical to the SC imputation (where we also restrict the weights to be nonnegative and sum to one).

In an alternative connection, [ABD$^+$21b] show that in some cases linear versions of all four estimators can all be characterized as solutions to the same optimization problem, subject to different restrictions on parameters of that optimization problem.

Now 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

$$(18) \quad \sum_{i=1}^{N} \sum_{t=1}^{T} (1 - W_{it}) \left(Y_{it} - \sum_{r=1}^{R} U_{ir} V_{tr} - \alpha_i - \beta_t \right)^2 + \lambda \left(\sum_{i=1}^{N} \sum_{r=1}^{R} U_{ir}^2 + \sum_{t=1}^{T} \sum_{r=1}^{R} V_{tr}^2 \right).$$

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 minimizing the objective function (18) 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 $(1 - W_{it})(Y_{it} - \sum_{r=1}^{R} A_{ir} B_{tr} - \gamma_i - \delta_t) = 0$ for all (i, t), with different imputed values for Y_{NT}. The implication is that we need to add some structure to the

optimization problem. The next result shows that unconfoundedness regression, the SC estimator, the DID estimator, and the MC estimator can all be expressed as minimizing the objective function under different restrictions on, or with different approaches to regularization of, $(R, \mathbf{U}, \mathbf{V}, \alpha, \beta)$.

NUCLEAR NORM MATRIX COMPLETETION 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 characterized as the solution to minimizing (18) with the restrictions

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

SYNTHETIC CONTROL The SC estimator imposes the restrictions subject to

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

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. See [AH23, IV23] 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. 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 as long as T_0 goes to infinity fast enough. 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 [AAH$^+$21, AH23].

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

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

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 $\varepsilon_{iT_0+1} - \frac{1}{T_0} \sum_{t \leq T_0} \varepsilon_{it}$ remains correlated with D_i which introduces bias. The performance of the SC estimator is different, and the results in [AH23] show that the SC estimator is consistent and asymptotically unbiased as long as T_0 goes to infinity fast enough. The consistency properties of the matrix completion estimator and the unconfoundedness regression are not established for this setting.

[IV23] focuses 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 analyze the behavior 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.

8.   Nonlinear Models

In this section we discuss some nonlinear panel data models. By nonlinear models, we mean here models that do not specify a model for the conditional mean that is 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 analyzed. With binary outcomes, it is 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

[AI06] focus on the repeated cross-section case with two periods and two groups, one treated in the second period, and one always in the control group. 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 it cannot hold in logarithms. In fact, in some cases on 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 [AI06] 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

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

[AI06] show that the 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 [AB99]. [Ark19] discusses a similar approach to the changes-in-changes estimator in [AI06], where the role of the groups and time periods are reversed and also considers an extension for multiple outcomes. [Woo22] also studies nonlinear versions of DID and TWFE approaches. In the two-period-two-group setting this 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

[Gun23] 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 [Gun23] 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:

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

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 absence of the treatment is identical to the quantile function for the control group, and the pre-treatment distributions are immaterial.

8.3. Balancing Statistics to Control for Unit Differences

[AI22] focus on settings where the treatment can switch on and off, as in the assignment matrix in Equation (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 between 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. [AI22] 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 [AI18a]. 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 [Mun78].

[AI18a] build on the Mundlak insight, still in the clustered setting, by making the unconfoundedness assumption that

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

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 directly related to [AM05], 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 non-parametric adjustment methods, possibly in combination with inverse propensity score weighting.

[AI18a] then generalize this by assuming that

$$W_i \perp\!\!\!\perp \left(Y_i(0), Y_i(1) \right) \mid X_i, S_{C_i},$$

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.

[AI22] extend these ideas from the clustered cross-sction 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, [Vog12]. [Woo21] shows the same estimator can be obtained by through what he calls the Mundlak regression

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

[AI22] 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:

(19) $$\underline{W}_i \perp\!\!\!\perp Y_{it}(w) \mid S_i.$$

Consider the case where the balancing statistic is fraction treated periods, $S_i = \overline{W}_i$. The unconfoundedness assumption in (19) 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 [AI22] 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.

8.4. Negative Controls, Proxies, and Deconvolution

The results in [AI22] 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 emphasize the connections between this literature and economic applications, we use these two terms interchangeably. In biostatistics a recent literature focuses on nonparametric identification results for average treatment effects that are based on negative controls [SRC$^+$16, SMNT20]. See [YMST21] for an introductory article. This literature is tightly connected to econometric literature on non-parametric identification with measurement error [HS08] and the changes-in-changes model [AI06]. 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 [Dea21b].

Proxy variables have a long history in economics. In early applications [Gri77, Cha77] use data on several test scores to estimate returns to schooling accounting for unobserved ability (see also [Dea21a]). 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, [HENR88] 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 [IKM21] for a discussion).

A similar idea is used in [FHS19], 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 a 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 [BS11, AB11a] and nonlinear quantile panel data models [AB16], 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. [ACI20] 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 [GYR+22, IKM21, KM20].

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 the interpretation of the uncertainty in the estimates. 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 be see 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 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 analyze 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 realization of a random process defined on a superpopulation of alternative nations." ([MP18], p. 234).

An alternative approach to formalizing uncertainty focuses on the random assignment of causes, taking the potential outcomes as fixed. This approach has a long history in the analysis of randomized experiments [Fis37, Ney90] where the justification for viewing the causes as random is immediate. For modern discussions see [IR15, Ros23]. Recently these ideas have been used to capture uncertainty in observational studies, see

[AAIW20, AAIW23]. 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 re-unification in 1989 on West German Gross Domestic Product. A design based approach would require the researcher to contemplate an alternative world where either other countries would have joint with East Germany, or an alternative world where the re-unification with West Germany would have happened in a different year. Both are difficult to consider. On the other hand, a sampling-based approach would require the researcher to consider a world with additional countries that could experience a unification event, again not an easy task.

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

[AI18b] analyze 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. [AI18b] derive the randomization-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 writte 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:

$$(20) \qquad \tau_t^{a,a'} = \frac{1}{N} \sum_{i=1}^{N} \Big(Y_{it}(a') - Y_{it}(a) \Big),$$

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

[AI18b] 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. [RR20] 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 [Coc39] called the *rotation experiment*, and what more recently has been referred as a *switchback experiment* [BSLZ20] or *crossover experiment* [BJ80]. 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, cattled 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 design have become popular in tech company settings to deal with spillovers. For example, Lyft and Uber often randomize 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 [HHC+15, HH07, BMDC16] 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.

[XABI19] focused on the question of optimally choosing the fraction 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. [BBI+23] discuss randomization-based inference for some of these settings and present exact variances for some estimators.

9.4. Design with Dynamic Effects

[BRS21] propose unbiased estimators and derive their properties under the randomization 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. [BSLZ22] discuss optimal design from a minimax perspective, allowing for carryover 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 randomized experiments. However, in practice, applied researchers continue using conventional panel data methods, such as TWFE, even with experimental data (*e.g.,* [BP14, CP22]). 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., [BH22]).

To address these issues [AILL21] 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:

$$(21) \qquad (\hat{\tau}^{rob}, \hat{\alpha}, \hat{\beta}) = \arg \min_{\tau, \alpha, \beta} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \alpha_i - \beta_t - \tau W_{it} \right)^2 \omega_i,$$

where the weights $\{\omega_i\}_{i=1}^{n}$ are constructed using the information about the design.

10.   Open Questions for Further Research

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

10.1.   Modeling Dynamics in Potential Outcomes

The recent panel data literature has only paid limited attention to dynamic treatment effects (*e.g.,* [Han20]). 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 the follow [Rob86] and a number of follow-up studies that developed a sequential unconfoundedness approach. See also [BGK+15, BMAF+23]. Another direction is the earlier panel data literature (*e.g.,* [AH81]), with [AH07a] providing an overview.

10.2. Validation

[LaL86] has become a very 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 no data sets that can deliver the same in many other panel data settings. However, there are methods that can be used to assess the performance of proposed estimators. An early paper with suggested tests is [HH89]. Currently many approaches in panel data rely on placebo tests where the researcher pretends the treatment ocurred 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 favor of the proposed methods. See for examples [IRS01] and [ADH15].

10.3. Dynamic Treatment Effects

The modern panel literature has paid relatively little attention to dynamic effects, compared to the earlier literature [HN07], as well as compared to its importance in practice. An interesting approach, somewhat related to synthetic control methods and factor models, and taking a Bayesian perspectcive, is developed in [BGK+15]. Other recent work includes [Han20, BB23].

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 emphasized 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 [CGBS21] and [dCD22], 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. Conclusion

The recent literature has 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.

References

[AA03]   Javier Alvarez and Manuel Arellano. The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica*, 71(4):1121–1159, 2003.

[AAH$^+$21] Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021.

[AAIW20] Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296, 2020.

[AAIW23] Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35, 2023.

[AB91]   Manuel Arellano and Stephen Bond. Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297, 1991.

[AB99]   Joseph G Altonji and Rebecca M Blank. Race and gender in the labor market. *Handbook of labor economics*, 3:3143–3259, 1999.

[AB11a]   Manuel Arellano and Stéphane Bonhomme. Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies*, 79(3):987–1020, 2011.

[AB11b]   Manuel Arellano and Stéphane Bonhomme. Nonlinear panel data analysis. 2011.

[AB16] Manuel Arellano and Stéphane Bonhomme. Nonlinear panel data estimation via quantile regressions, 2016.

[Aba05] Alberto Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19, 2005.

[Aba19] Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 2019.

[ABD⁺21a] Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 2021.

[ABD⁺21b] Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

[AC85] Orley Ashenfelter and David Card. Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, 67(4):648–660, 1985.

[AC86] John M Abowd and David Card. On the covariance structure of earnings and hours changes, 1986.

[ACI20] Susan Athey, Raj Chetty, and Guido Imbens. Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*, 2020.

[ADH10] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. 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, 2010.

[ADH15] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, pages 495–510, 2015.

[AG03] Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(-):113–132, 2003.

[AG04] Orley Ashenfelter and Michael Greenstone. Using mandated speed limits to measure the value of a statistical life. *Journal of political Economy*, 112(S1):S226–S267, 2004.

[AH81] Theodore Wilbur Anderson and Cheng Hsiao. Estimation of dynamic models with error components. *Journal of the American statistical Association*, 76(375):598–606, 1981.

[AH01] Manuel Arellano and Bo Honoré. Panel data models: some recent developments. *Handbook of econometrics*, 5:3229–3296, 2001.

[AH07a] Jaap H Abbring and James J Heckman. Econometric evaluation of social programs, part iii: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. *Handbook of econometrics*, 6:5145–5303, 2007.

[AH07b] Manuel Arellano and Jinyong Hahn. Understanding bias in nonlinear panel models: Some recent developments. *Econometric Society Monographs*, 43:381, 2007.

[AH23] Dmitry Arkhangelsky and David Hirshberg. Large-sample properties of the synthetic control method under selection on unobservables. *arXiv preprint arXiv:2311.13575*, 2023.

[AI06] Susan Athey and Guido W Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006.

[AI18a] Dmitry Arkhangelsky and Guido Imbens. The role of the propensity score in fixed effect models. Technical report, National Bureau of Economic Research, 2018.

[AI18b] Susan Athey and Guido Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. 2018.

[AI21] Susan Athey and Guido W Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 2021.

[AI22] Dmitry Arkhangelsky and Guido W Imbens. Doubly robust identification for causal panel data models. *The Econometrics Journal*, 25(3):649–674, 2022.

[AILL21] Dmitry Arkhangelsky, Guido W Imbens, Lihua Lei, and Xiaoman Luo. Double robust two-way fixed effect regression for panel data. *arXiv preprint arXiv:1909.09412*, 2021.

[AIW18] Susan Athey, Guido W Imbens, and Stefan Wager. 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, 2018.

[AK91] Joshua D Angrist and Alan Krueger. Does compulsory schooling affect schooling and earnings. *Quarterly Journal of Economics*, CVI(4):979–1014, 1991.

[AKM99] John M Abowd, Francis Kramarz, and David N Margolis. High wage workers and high wage firms. *Econometrica*, 67(2):251–333, 1999.

[AL21] Alberto Abadie and Jérémy L'hour. A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, 116(536):1817–1834, 2021.

[Ald81] David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

[AM05] Joseph G Altonji and Rosa L Matzkin. Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica*, 73(4):1053–1102, 2005.

[And84] Theodore Wilbur Anderson. Estimating linear statistical relationships. *The Annals of Statistics*, 12(1):1–45, 1984.

[AP08] Joshua Angrist and Steve Pischke. *Mostly Harmless Econometrics: An Empiricists' Companion*. Princeton University Press, 2008.

[Are87] Manuel Arellano. Computing robust standard errors for within group estimators. *Oxford bulletin of Economics and Statistics*, 49(4):431–434, 1987.

[Are03] Manuel Arellano. *Panel data econometrics*. OUP Oxford, 2003.

[Ark19] Dmitry Arkhangelsky. *Dealing with a technological bias: The difference-in-difference approach*. Centro de estudios monetarios y financieros, 2019.

[AS02] Susan Athey and Scott Stern. The impact of information technology on emergency health care outcomes. *The RAND Journal of Economics*, 33(3):399–432, 2002.

[AS17] Peter M Aronow and Cyrus Samii. Estimating average causal effects under general

interference, with application to a social network experiment. *Annals of Applied Statistics*, 11(4):1912–1947, 2017.

[Ash78]   Orley Ashenfelter. Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, pages 47–57, 1978.

[ASS18]   Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *The Journal of Machine Learning Research*, 19(1):802–852, 2018.

[AVdB03] Jaap H Abbring and Gerard J Van den Berg. The nonparametric identification of treatment effects in duration models. *Econometrica*, 71(5):1491–1517, 2003.

[Bai09]   Jushan Bai. Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279, 2009.

[Bal08]   Badi Baltagi. *Econometric analysis of panel data*. John Wiley & Sons, 2008.

[BB98]   Richard Blundell and Stephen Bond. Initial conditions and moment restrictions in dynamic panel data models. *Journal of econometrics*, 87(1):115–143, 1998.

[BB23]   Nicholas Brown and Kyle Butts. Dynamic treatment effect estimation with interactive fixed effects and short panels. 2023.

[BBI+23] Patrick Bajari, Brian Burdick, Guido W Imbens, Lorenzo Masoero, James McQueen, Thomas S Richardson, and Ido M Rosen. Experimental design in marketplaces. *Statistical Science*, 1(1):1–19, 2023.

[BDM04]   Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275, 2004.

[Bek94]   Paul A Bekker. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, pages 657–681, 1994.

[BGK+15] Kay H Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L Scott. Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, pages 247–274, 2015.

[BH22]   Kirill Borusyak and Peter Hull. Non-random exposure to exogenous shocks. *NBER Working Paper*, (27845), 2022.

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[BJ80]   Byron Wm Brown Jr. The crossover experiment for clinical trials. *Biometrics*, pages 69–79, 1980.

[BJ15]   Andrew Bell and Kelvyn Jones. Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, 3(1):133–153, 2015.

[BJS21]   Kirill Borusyak, Xavier Jaravel, and Jann Spiess. Revisiting event study designs: Robust and efficient estimation. *arXiv preprint arXiv:2108.12419*, 2021.

[BLM22a] Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. Discretizing unobserved heterogeneity. *Econometrica*, 90(2):625–643, 2022.

[BLM22b] Luca Braghieri, Ro'ee Levy, and Alexey Makarin. Social media and mental health.

American Economic Review, 112(11):3660–3693, 2022.

[BLW21] Andrew Baker, David F Larcker, and Charles CY Wang. How much should we trust staggered difference-in-differences estimates? *Available at SSRN 3794018*, 2021.

[BM15] Stéphane Bonhomme and Elena Manresa. Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184, 2015.

[BMAF⁺23] Eli Ben-Michael, David Arbour, Avi Feller, Alexander Franks, and Steven Raphael. Estimating the effects of a california gun control program with multitask gaussian processes. *The Annals of Applied Statistics*, 17(2):985–1016, 2023.

[BMDC16] Daniel Barker, Patrick McElduff, Catherine D'Este, and MJ Campbell. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. *BMC medical research methodology*, 16(1):1–19, 2016.

[BMFR21] Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536):1789–1803, 2021.

[BMFR22] Eli Ben-Michael, Avi Feller, and Jesse Rothstein. Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):351–381, 2022.

[BN02] Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

[Bon12] Stéphane Bonhomme. Functional differencing. *Econometrica*, 80(4):1337–1385, 2012.

[Bon20] Stéphane Bonhomme. Econometric analysis of bipartite networks. In *The econometric analysis of network data*, pages 83–121. Elsevier, 2020.

[BP14] Christian Broda and Jonathan A Parker. The economic stimulus payments of 2008 and the aggregate demand for consumption. *Journal of Monetary Economics*, 68:S20–S36, 2014.

[BR05] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[BRS21] Iavor Bojinov, Ashesh Rambachan, and Neil Shephard. Panel experiments and dynamic causal effects: A finite population perspective. *Quantitative Economics*, 12(4):1171–1196, 2021.

[BS11] Stéphane Bonhomme and Ulrich Sauder. Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *Review of Economics and Statistics*, 93(2):479–494, 2011.

[BSLZ20] Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switch-back experiments. *Available at SSRN 3684168*, 2020.

[BSLZ22] Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switch-back experiments. *Management Science*, 2022.

[Car94] David Card. *Intertemporal labour supply: an assessment*, volume 2 of *Econometric Society Monographs*, page 49–78. Cambridge University Press, 1994.

[CCD⁺17] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

[CFT21] Matias D Cattaneo, Yingjie Feng, and Rocio Titiunik. Prediction intervals for syn-

thetic control methods. *Journal of the American Statistical Association*, 116(536):1865–1880, 2021.

[CFVHN13] Victor Chernozhukov, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey. Average and quantile effects in nonseparable panel models. *Econometrica*, 81(2):535–580, 2013.

[CGBS21] Brantly Callaway, Andrew Goodman-Bacon, and Pedro HC Sant'Anna. Difference-in-differences with a continuous treatment. *arXiv preprint arXiv:2107.02637*, 2021.

[Cha77] Gary Chamberlain. Education, income, and ability revisited. *Journal of Econometrics*, 5(2):241–257, 1977.

[Cha80] Gary Chamberlain. Analysis of covariance with qualitative data. *The review of economic studies*, 47(1):225–238, 1980.

[Cha82] Gary Chamberlain. Multivariate regression models for panel data. *Journal of econometrics*, 18(1):5–46, 1982.

[Cha84] Gary Chamberlain. Panel data. *Handbook of econometrics*, 2:1247–1318, 1984.

[Cha92] Gary Chamberlain. Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 567–596, 1992.

[Cha10] Gary Chamberlain. Binary response models for panel data: Identification and information. *Econometrica*, 78(1):159–168, 2010.

[CK94] David Card and Alan Krueger. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, 84(4):772–793, 1994.

[CLLX23] Albert Chiu, Xingchen Lan, Ziyi Liu, and Yiqing Xu. What to do (and not to do) with causal panel analysis under parallel trends: Lessons from a large reanalysis study. *Available at SSRN 4490035*, 2023.

[CM22] Denis Chetverikov and Elena Manresa. Spectral and post-spectral estimators for grouped panel data models. *arXiv preprint arXiv:2212.13324*, 2022.

[Coc39] WG Cochran. Long-term agricultural experiments. *Supplement to the Journal of the Royal Statistical Society*, 6(2):104–148, 1939.

[CP22] Emanuele Colonnelli and Mounu Prem. Corruption and firms. *The Review of Economic Studies*, 89(2):695–732, 2022.

[CR83] Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica: Journal of the Econometric Society*, pages 1281–1304, 1983.

[CRY22] David Card, Jesse Rothstein, and Moises Yi. Industry wage differentials: A firm-based approach. *Unpublished draft, University of California, Berkeley*, 2022.

[CS20] Brantly Callaway and Pedro HC Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 2020.

[CWZ21] Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864, 2021.

[DCd20] Clément De Chaisemartin and Xavier d'Haultfœuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–

2996, 2020.

[dCD22] Clément de Chaisemartin and Xavier D'haultfœuille. Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *Econometrics Journal*, 2022.

[Dea85] Angus Deaton. Panel data from time series of cross-sections. *Journal of econometrics*, 30(1-2):109–126, 1985.

[Dea21a] Ben Deaner. Many proxy controls. *arXiv preprint arXiv:2110.03973*, 2021.

[Dea21b] Ben Deaner. Proxy controls and panel data. *arXiv preprint arXiv:1810.00283*, 2021.

[DI16] Nikolay Doudchenko and Guido W Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research, 2016.

[DL19] Peng Ding and Fan Li. A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis*, 27(4):605–615, 2019.

[DW99] Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

[DW02] Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

[EHR83] Robert F Engle, David F Hendry, and Jean-Francois Richard. exogeneity. *Econometrica: Journal of the Econometric Society*, pages 277–304, 1983.

[EL96] Nada Eissa and Jeffrey B Liebman. Labor supply response to the earned income tax credit. *The quarterly journal of economics*, 111(2):605–637, 1996.

[FFJR69] Eugene F Fama, Lawrence Fisher, Michael C Jensen, and Richard Roll. The adjustment of stock prices to new information. *International economic review*, 10(1):1–21, 1969.

[FHS19] Simon Freyaldenhoven, Christian Hansen, and Jesse M Shapiro. Pre-event trends in the panel event-study design. *American Economic Review*, 109(9):3307–38, 2019.

[Fis37] Ronald Aylmer Fisher. *The design of experiments*. Oliver And Boyd; Edinburgh; London, 1937.

[FP21] Bruno Ferman and Cristine Pinto. Synthetic controls with imperfect pretreatment fit. *Quantitative Economics*, 12(4):1197–1221, 2021.

[Fre18] Joachim Freyberger. Non-parametric panel data models with interactive fixed effects. *The Review of Economic Studies*, 85(3):1824–1851, 2018.

[FVW16] Iván Fernández-Val and Martin Weidner. Individual and time effects in nonlinear panel models with large n, t. *Journal of Econometrics*, 192(1):291–312, 2016.

[FVW18] Iván Fernández-Val and Martin Weidner. Fixed effects estimation of large-t panel data models. *Annual Review of Economics*, 10:109–138, 2018.

[GB21] Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021.

[GM16] Laurent Gobillon and Thierry Magnac. Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics*, 98(3):535–551,

2016.

[Gol91] Arthur Stanley Goldberger. *A course in econometrics*. Harvard University Press, 1991.

[GP12] Bryan S Graham and James L Powell. Identification and estimation of average partial effects in "irregular"' correlated random coefficient panel data models. *Econometrica*, 80(5):2105–2152, 2012.

[Gri77] Zvi Griliches. Estimating the returns to schooling: Some econometric problems. *Econometrica: Journal of the Econometric Society*, pages 1–22, 1977.

[GSW22] Dalia Ghanem, Pedro HC Sant'Anna, and Kaspar Wüthrich. Selection and parallel trends. *arXiv preprint arXiv:2203.09001*, 2022.

[Gun23] Florian F Gunsilius. Distributional synthetic controls. *Econometrica*, 91(3):1105–1117, 2023.

[GYR+22] AmirEmad Ghassami, Alan Yang, David Richardson, Ilya Shpitser, and Eric Tchetgen Tchetgen. Combining experimental and observational data for identification and estimation of long-term causal effects. *arXiv preprint arXiv:2201.10743*, 2022.

[Han07] Christian B Hansen. Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *Journal of Econometrics*, 140(2):670–694, 2007.

[Han20] Sukjin Han. Identification in nonparametric models for dynamic treatment effects. *Journal of Econometrics*, 2020.

[Hec81] James J Heckman. Statistical models for discrete panel data. *Structural analysis of discrete data with econometric applications*, 114:178, 1981.

[HENR88] Douglas Holtz-Eakin, Whitney Newey, and Harvey S Rosen. Estimating vector autoregressions with panel data. *Econometrica: Journal of the econometric society*, pages 1371–1395, 1988.

[HH89] James J Heckman and V Joseph Hotz. 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):862–874, 1989.

[HH07] Michael A Hussey and James P Hughes. Design and analysis of stepped wedge cluster randomized trials. *Contemporary clinical trials*, 28(2):182–191, 2007.

[HH08] Michael Hudgens and Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, pages 832–842, 2008.

[HHC+15] Karla Hemming, Terry P Haines, Peter J Chilton, Alan J Girling, and Richard J Lilford. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *Bmj*, 350, 2015.

[HK02] Jinyong Hahn and Guido Kuersteiner. Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and t are large. *Econometrica*, 70(4):1639–1657, 2002.

[HN07] James J Heckman and Salvador Navarro. Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics*, 136(2):341–396, 2007.

[Hon92] Bo E Honoré. Trimmed lad and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica: journal of the Econometric Society*, pages 533–565, 1992.

[HS08] Yingyao Hu and Susanne M Schennach. Instrumental variable treatment of non-classical measurement error models. *Econometrica*, 76(1):195–216, 2008.

[HSCKW12] Cheng Hsiao, H Steve Ching, and Shui Ki Wan. A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics*, 27(5):705–740, 2012.

[Hsi22] Cheng Hsiao. *Analysis of panel data*. Number 64. Cambridge university press, 2022.

[HT06] Bo E Honoré and Elie Tamer. Bounds on parameters in panel dynamic discrete choice models. *Econometrica*, 74(3):611–629, 2006.

[IK21] Kosuke Imai and In Song Kim. On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis*, 29(3):405–415, 2021.

[IKM21] Guido Imbens, Nathan Kallus, and Xiaojie Mao. Controlling for unmeasured confounding in panel data using minimal bridge functions: From two-way fixed effects to factor models. *arXiv preprint arXiv:2108.03849*, 2021.

[IKW21] Kosuke Imai, In Song Kim, and Erik H Wang. Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 2021.

[Imb02] Guido W Imbens. Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics*, 20(4):493–506, 2002.

[Imb04] Guido Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, pages 1–29, 2004.

[IR15] Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[IRS01] Guido W Imbens, Donald B Rubin, and Bruce I Sacerdote. Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American Economic Review*, pages 778–794, 2001.

[IV23] Guido Imbens and Davide Viviano. Identification and inference for synthetic controls with confounding. 2023.

[Jak21] Pamela Jakiela. Simple diagnostics for two-way fixed effects. *arXiv preprint arXiv:2103.13229*, 2021.

[KM20] Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408*, 2020.

[KMPT21] Maxwell Kellogg, Magne Mogstad, Guillaume A Pouliot, and Alexander Torgovitsky. Combining matching and synthetic control to tradeoff biases from extrapolation and interpolation. *Journal of the American statistical association*, 116(536):1804–1816, 2021.

[KZEM21] Timo Kuosmanen, Xun Zhou, Juha Eskelinen, and Pekka Malo. Design flaw of the synthetic control method. 2021.

[LaL86] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

[Lan00] Tony Lancaster. The incidental parameter problem since 1948. *Journal of econometrics*, 95(2):391–413, 2000.

[LC20]   Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *arXiv preprint arXiv:2006.06138*, 2020.

[LJ76]   Robert E Lucas Jr. Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pages 19–46. North-Holland, 1976.

[LWX22]   Licheng Liu, Ye Wang, and Yiqing Xu. A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 2022.

[Lyn84]   James Lynch. Canonical row-column-exchangeable arrays. *Journal of multivariate analysis*, 15(1):135–140, 1984.

[LZ86]   Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[Mag04]   Thierry Magnac. Panel binary variables and sufficiency: generalizing conditional logit. *Econometrica*, 72(6):1859–1876, 2004.

[MP18]   Charles F Manski and John V Pepper. How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics*, 100(2):232–244, 2018.

[Mug22]   Martin Mugnier. Make the difference! computationally trivial estimators for grouped fixed effects models. *arXiv preprint arXiv:2203.08879*, 2022.

[Mun78]   Yair Mundlak. On the pooling of time series and cross section data. *Econometrica*, pages 69–85, 1978.

[MVD95]   Bruce D Meyer, W Kip Viscusi, and David L Durbin. Workers' compensation and injury duration: Evidence from a natural experiment. *The American Economic Review*, 85(3):322, 1995.

[MW15]   Hyungsik Roger Moon and Martin Weidner. Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, 83(4):1543–1579, 2015.

[MW18]   Hyungsik Roger Moon and Martin Weidner. Nuclear norm regularized estimation of panel regression models. *arXiv preprint arXiv:1810.10987*, 2018.

[Ney90]   Jerzey Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1923/1990.

[Nic81]   Stephen Nickell. Biases in dynamic models with fixed effects. *Econometrica: Journal of the econometric society*, pages 1417–1426, 1981.

[NS48]   Jerzy Neyman and Elizabeth L Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

[Pes06]   M Hashem Pesaran. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012, 2006.

[PT07]   Thomas Plümper and Vera E Troeger. Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects. *Political analysis*, 15(2):124–139, 2007.

[RHB00]   James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560, 2000.

[Rob86] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

[Ros02] Paul R Rosenbaum. Multiple control groups. *Observational Studies*, pages 253–275, 2002.

[Ros23] Paul R Rosenbaum. *Causal inference*. MIT Press, 2023.

[RR83]   Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[RR20]   Ashesh Rambachan and Jonathan Roth. Design-based uncertainty for quasi-experiments. *arXiv preprint arXiv:2008.00602*, 2020.

[RR23]   Ashesh Rambachan and Jonathan Roth. A more credible approach to parallel trends. *Review of Economic Studies*, page rdad018, 2023.

[RS23]   Jonathan Roth and Pedro HC Sant'Anna. When is parallel trends sensitive to functional form? *Econometrica*, 91(2):737–747, 2023.

[RSBP23] Jonathan Roth, Pedro HC Sant'Anna, Alyssa Bilinski, and John Poe. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 2023.

[Rub74] Donald B Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[Rub78] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

[SA20]   Liyang Sun and Sarah Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 2020.

[SDSY22] Dennis Shen, Peng Ding, Jasjeet Sekhon, and Bin Yu. A tale of two panel data regressions. *arXiv preprint arXiv:2207.14481*, 2022.

[SIV23] Jann Spiess, Guido Imbens, and Amar Venugopal. Double and single descent in causal inference with an application to high-dimensional synthetic control. *arXiv preprint arXiv:2305.00700*, 2023.

[SMNT20] Xu Shi, Wang Miao, Jennifer C. Nelson, and Eric J. Tchetgen Tchetgen. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 82(2):521–540, 2020.

[SRC+16] Tamar Sofer, David B Richardson, Elena Colicino, Joel Schwartz, and Eric J Tchetgen Tchetgen. On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 31(3):348, 2016.

[SSP+19] Pantelis Samartsidis, Shaun R Seaman, Anne M Presanis, Matthew Hickman, and Daniela De Angelis. Assessing the causal effect of binary interventions from observational panel data with few treated units. *Statistical Science*, 34(3):486–503, 2019.

[SW98]   James H Stock and Mark W Watson. Diffusion indexes, 1998.

[SZ20]   Pedro HC Sant'Anna and Jun Zhao. Doubly robust difference-in-differences estima-

tors. *Journal of Econometrics*, 219(1):101–122, 2020.

[Vog12] Timothy J Vogelsang. Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects. *Journal of Econometrics*, 166(2):303–319, 2012.

[Woo10] J.M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press. MIT Press, 2010.

[Woo21] Jeffrey M Wooldridge. Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Available at SSRN 3906345*, 2021.

[Woo22] Jeffrey M Wooldridge. Simple approaches to nonlinear difference-in-differences with panel data. *Available at SSRN 4183726*, 2022.

[XABI19] Ruoxuan Xiong, Susan Athey, Mohsen Bayati, and Guido Imbens. Optimal experimental design for staggered rollouts. *arXiv preprint arXiv:1911.03764*, 2019.

[Xu17] Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76, 2017.

[Xu23] Yiqing Xu. Causal inference with time-series cross-sectional data: a reflection. *Available at SSRN 3979613*, 2023.

[YMST21] Andrew Ying, Wang Miao, Xu Shi, and Eric J Tchetgen Tchetgen. Proximal causal inference for complex longitudinal studies. *arXiv preprint arXiv:2109.07030*, 2021.