Causal Inference with Time-Series Cross-Sectional Data: A Reflection*

Yiqing Xu
(Stanford University)

Prepared for
The Oxford Handbook for Methodological Pluralism

First Version: September 26, 2021
This Version: March 10, 2022

Abstract

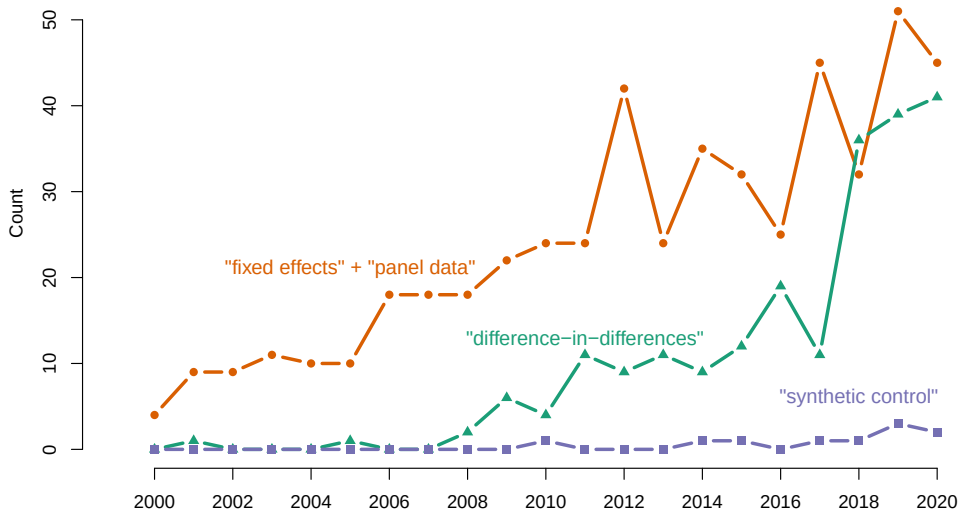
This chapter surveys new development in causal inference using time-series cross-sectional (TSCS) data. I start by clarifying two identification regimes for TSCS analysis: one under the strict exogeneity assumption and one under the sequential ignorability assumption. I then review three most commonly used methods by political scientists: the difference-in-differences approach, two-way fixed effects models, and the synthetic control method. For each method, I examine its assumptions, explain its pros and cons, and discuss its extensions. I then introduce several new methods under strict exogeneity or sequential ignorability, including the factor-augmented approach, panel matching, and marginal structural models. I conclude by providing some recommendations to applied researchers and pointing to several directions for future research.

Keywords: causal inference, time-series-cross-sectional data, panel data, difference-in-differences, two-way fixed effects, synthetic control

*Yiqing Xu, Department of Political Science, Stanford University. Email: yiqingxu@stanford.edu. I thank Avi Feller, Seth Hill, Kosuke Imai, Jonathan Katz, Luke Keele, Lihua Lei, Molly Roberts, Ye Wang, as well as members of the Causal Inference with Panel Data Reading Group at Stanford, for helpful comments. All errors are mine.

Analysis of time-series cross-sectional (TSCS) data, or long panel data, has becoming increasingly popular in empirical political science during the past decade. The popularity of TSCS data among political scientists is accompanied by an ever-rising interest in establishing causal relationships between treatment variables and outcome variables using observational data. The "credibility revolution," first started in statistics and economics and spills over to other social sciences, pushes scholars to answer causal questions;[1] TSCS data often come to the rescue when other identification strategies are unsuitable or infeasible—for example, seldomly can researchers randomly assign political institutions to different countries, or take advantage of a binding rule that determines electoral institutions of localities to allow a regression discontinuity design, or find a plausible instrumental variable for a state's nuclear capabilities.

FIGURE 1. NUMBER OF PAPERS PUBLISHED IN 5 TOP POLITICAL SCIENCE JOURNALS USING FIXED EFFECTS AND DID METHODS



Note: Data are based on Google Scholar advanced searches (accessed on September 18, 2021). The five journals are *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, *Comparative Political Studies*, and *International Organization*. I thank Ziwen Zu for helping me collect the data.

Remarkably, among all available methods for TSCS analysis, difference-in-differences

[1]Whether this obsession with causal inference is a good thing for knowledge accumulation in the social sciences is up to debate. See, for example, Deaton and Cartwright (2018) and Imbens (2010) for more discussion.

(DID) and two-way fixed effects (TWFE) are becoming the two most frequently used methods by political scientists. Figure 1 plots the numbers of papers mentioning fixed effects (in a panel setting) and DID published in five top political science journals since 2000. I suspect that their increasing popularity is mainly due to fact that popular econometrics textbooks, such as Angrist and Pischke (2009), credit DID and TWFE as trustworthy methods for establishing causal relationships using observational data. Another method that has enter political scientists' toolkit is the synthetic control method (SCM). Because of the need for establishing causality with comparative case studies, as well as the SCM's clever and intuitive design, it has gradually attracted scholars' attention since its invention.

The focus on the DID, TWFE, SCM methods has important consequences. First, it sets common practice apart from recommendations from earlier political science research—e.g., Beck and Katz (2004, 2011) argue that TSCS data are fundamentally different from panel data and the time dimension of a TSCS dataset needs to be carefully analyzed to avoid biased estimates and erroneous conclusions—TSCS analysts in recent years rarely check whether the time-series data are stationary or take advantage of their data's time dimension except for controlling for "time fixed effects." Second, many studies are susceptible to drawbacks of these popular methods, especially, TWFE models. These drawbacks have drawn much attention recently in the econometrics literature. Third, both DID and TWFE rely on a type of parallel trends assumptions, which corresponds to a particular type of directional acyclical graphs (DAGs) or hypothetical experiments. These assumptions are implausible in many dynamics-rich social science contexts. Critics of these methods argue that they should not be interpreted as methods exploiting "natural experiments" because the sources of randomness in treatment assignment in such designs is usually unclear to researchers (Keele, 2020). It is indeed strange that the treatment assignment mechanisms behind these methods have received little attention when they are frequently being used to establish causality. Fourth, because these methods operate in the linear world, few empirical research pay special attention to binary or discrete outcomes, which may lead to misleading research

outcomes (Beck and Katz, 2001).

In this chapter, I first introduce two main identification regimes for TSCS data analysis. The motivation is simple: If researchers want to claim that estimates from a TSCS model are causally interpretable, they need to be clear about the identification assumptions based on which such claims are made. I then discuss each of the three methods, including their advantages, drawbacks, and extensions. I follow up with a survey of several newly emerged methods, including factor augmented models, panel matching, and marginal structural models. I conclude by suggesting several directions for future research. One major caveat is that most of the methods covered in this chapter concerns dichotomous treatments only. Future research is needed to investigate how researchers can obtain credible causal estimates based on discrete or continuous treatments without strong modeling assumptions.

Two Identification Regimes

Broadly speaking, there are two identification regimes for TSCS analysis. By identification regimes, I mean the combination of key identification assumptions and the data generating processes (DGPs) consistent with the assumptions. Together, they allow researchers to identify some casual quantity of interest using observed data.

Strict exogeneity. The first identification regime is the one under the strict exogeneity assumption. Researchers usually invoke strict exogeneity when estimating a TWFE model. Assume a balanced TSCS dataset with N units and T periods. Strict exogeneity states that

$$\text{(Strict exogeneity)} \qquad \{Y_{it}(0), Y_{it}(1)\} \perp\!\!\!\perp D_{is} | \mathbf{X}_i^{1:T}, \alpha_i, \mathbf{f}^{1:T}, \quad \forall i, t, s, \tag{1}$$

in which D_{is} is treatment status for unit i at time s; $Y_{it}(0)$ and $Y_{it}(1)$ are potential outcomes under the treatment condition (when $D_{it} = 1$) and under the control condition (when $D_{it} = 0$); $\mathbf{X}_i^{1:T} = \{X_{i1}, X_{i2}, \cdots, X_{iT}\}$ are the entire histories of covariates; α_i and

$\mathbf{f}^{1:T} = \{f_1, f_2, \cdots, f_T\}$ are unit and time fixed effects, respectively. When $\mathbf{f}^{1:T}$ are seen as fixed parameters, this assumption can also be interpreted as *selection on observables and time-invariant unobservables* (Callaway and Karami, 2021). Strict exogeneity is often accompanied by a cross-sectional stable unit treatment value assumption (SUTVA), which requires that a unit's potential outcomes are not affected by other units' treatment status. We take the cross-sectional SUTVA to be true in this chapter; however, when spatial spillover or general equilibrium effects are major concerns, researchers should call into question the validity of this assumption. Strict exogeneity implies the parallel trends assumption (also known as the common trend assumption), which is required in a DID design. It states that the expectations of non-treatment potential outcomes, conditional on observed characteristics, follow parallel paths, i.e.,

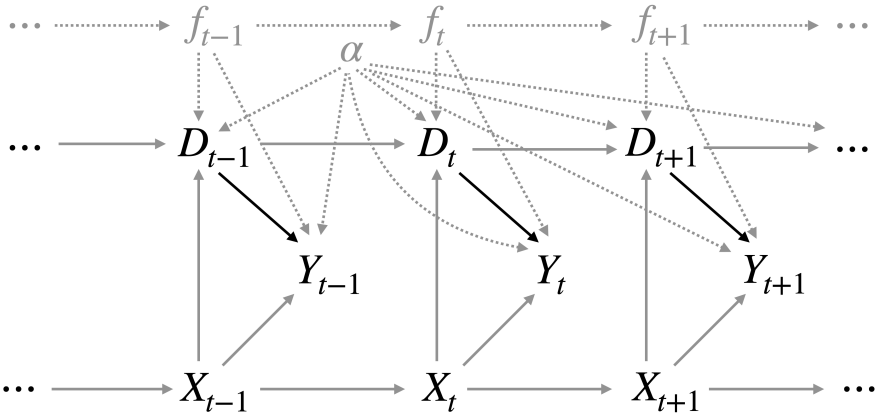
(Parallel trends)

$$\mathbb{E}[Y(0)_{it} - Y(0)_{is}|\mathbf{X}_i^{1:T}] = \mathbb{E}[Y(0)_{jt} - Y(0)_{js}|\mathbf{X}_j^{1:T}], \quad \forall i, j, t, s. \tag{2}$$

Although it is mathematically weaker than strict exogeneity, in practice, it is difficult to think of a realistic example in which the former holds while the latter fails. That is why researchers usually justify it in very similar ways as they justify strict exogeneity, e.g. by checking whether a "pre-trend" exists or by arguing that the timing of the treatment is quasi-exogenous conditional on unit fixed effects and the common trend. Hence, we treat them as qualitatively the same.

Figure 2 presents a DAG consistent with strict exogeneity. In this DAG, I omit unit subscripts for simplicity and use Y_{t-1}, Y_t, and Y_{t+1} to represent three consecutively observed outcome of a single unit. In addition, variables in gray are unobserved. Imai and Kim (2019) clarify that strict exogeneity implies that (1) no time-varying confounder exists—hence, no unobserved variables besides α and $\mathbf{f}$; (2) no direct effects from past outcomes to current outcome—hence, no arrows from Y_{t-1} to Y_t; (3) no "feedback" from past outcomes (or covariates) to current and future treatment status—hence, no arrows from Y_{t-1} to D_t or

FIGURE 2. DAG FOR DGPS UNDER STRICT EXOGENEITY



Note: Subscript i (for unit) is omitted for simplicity. Black solid arrows are the causal relationships of researchers' interest. Gray solid arrows and gray dashed arrows represent relationship among observed variables and relationships among variables some of which are unobserved, respectively.

X_{t-1} to D_t; (4) no "carryover effect" from current treatment to future outcomes—hence, no arrows from D_{t-1} to Y_t or Y_{t+1}.[2] Under strict exogeneity, we allow unit-invariant confounders α and common shocks $\mathbf{f}$ to affect both treatment assignment and the outcome. As we will see in the next section, this is achieved through a functional form requirement.

Blackwell and Glynn (2018) argue that strict exogeneity corresponds to hypothetical experiments with baseline randomization. A more precise interpretation is that the treatment assignment is pre-specified at the basedline and independent of the realization of the outcome. Hence, besides the well-known requirement of no time-varying confounders, the most crucial component of strict exogeneity is the lack of feedback from variables affected by past treatments (e.g. past outcome or covariates) to current and future treatment status. This is a key differentiating factor between strict exogeneity and sequential ignorability.

Sequential ignorability. DGPs under sequential ignorability allow most past information to directly affect treatment assignment and the outcome at the same time. Sequential ignorability states that treatment assignment is ignorable conditional on all past information,

[2]Note that we may relax (2) as long as (3) remains valid; we make relax (4) if either the carryover effect is limited or we do now aim at distinguishing contemptuous effect from carryover effects (e.g., in a staggered DID design). See below for more discussion.

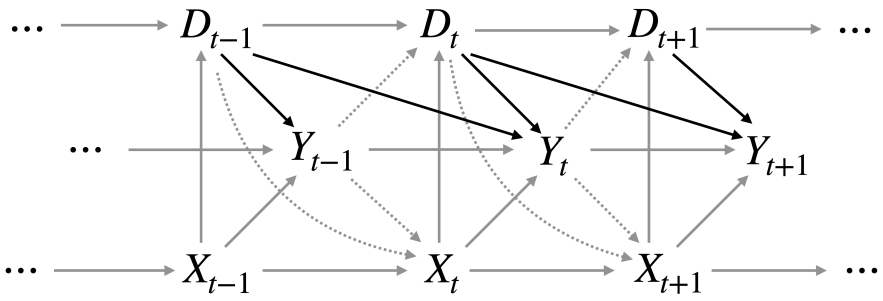
including past outcomes and covariates, i.e.,

$$\text{(Sequential ignorability)} \qquad \{Y_{it}(0), Y_{it}(1)\} \perp\!\!\!\perp D_{it} | \mathbf{X}_i^{1:t}, \mathbf{Y}_i^{1:(t-1)}, \quad \forall i, t, \tag{3}$$

in $\mathbf{X}_i^{1:t}$ denotes the histories of covariates up to time t and $\mathbf{Y}_i^{1:(t-1)} = \{Y_{i1}, Y_{i2}, \cdots, Y_{i,t-1}\}$ denotes the history of the outcome variable up to time $(t-1)$. A cross-sectional SUTVA assumption similar to that under strict exogeneity usually accompanies this assumption unless cross-sectional spillover or network effects are taken into consideration. Sequential ignorability is an extension of the strong ignorability assumption in the cross-sectional setting (Rosenbaum and Rubin, 1983).

Figure 3 presents a DAG consistent with sequential ignorability. Although the time and unit fixed effects are missing from the graph, it is obvious that most of the restrictions imposed by strict exogeneity are now relaxed. For example, sequential ignorability allows past outcomes and covariates to directly affect treatment assignment (feedback)—hence the arrows from Y_{t-1} to D_t and from X_{t-1} to D_t. It also allows past outcomes to directly affect current outcome—hence the arrows from Y_{t-1} to Y_t. Moreover, it allows covariates to be

FIGURE 3. DAG FOR DGPS UNDER SEQUENTIAL IGNORABILITY



Note: This figure is modified based on Figure 1 in Blackwell and Glynn (2018). Black solid arrows are causal relationships of researchers' interest. Gray solid arrows represent relationship among observed variables. Gray dashed arrows represent links absent from DAGs for DGPs under strict exogeneity (Figure 2).

affected by past treatments—hence the arrow from D_{t-1} to X_t. Carryover effects are allowed, too—hence the arrow from D_{t-1} to Y_t. Therefore, sequential ignorability tolerates much more dynamic relationships than strict exogeneity. It corresponds to hypothetical experiments

with sequentially randomized treatments in which treatment assignment depends on observed past outcomes and covariates (Blackwell and Glynn, 2018).

There are several important differences between the two identification regimes. First, under sequential ignorability, although researchers are still interested in estimating the direct, contemporaneous effect of D_t on Y_t, they are able to estimate the lagged indirectly effect of, say, D_{t-1} on Y_t through X_t, which is ruled out under strict exogeneity. Second, history matters under sequential ignorability: because the probability of getting treated is affected by observed outcome and covariates, the potential outcomes for a unit at the given time hinge on past treatment history as well as the history of the outcome and covariates. Third, sequential ignorability assumes away unobserved confounders, including time-invariant attributes of a unit and common time trends. Because both the outcome and treatment are path-dependent, conditional on averages of D or Y is not long a valid debiasing strategy and will cause biases in casual estimates. Typically, researchers resort to lagged dependent variable models and/or autoregressive distributed lag models in such a setting, without imposing unit fixed effects (e.g., Beck and Katz, 2011). I summarize the differences between these two identification regimes in Table 1.

TABLE 1. TWO IDENTIFICATION REGIMES

	Strict exogeneity	Sequential ignorability
Key assumption	$Y_{it}(d_{it}) \perp\!\!\!\perp D_{is} \vert \mathbf{X}_i^{1:T}, \mathbf{U}_i^{1:T}$	$Y_{it}(d_{it}) \perp\!\!\!\perp D_{it} \vert \mathbf{X}_i^{1:t}, \mathbf{Y}_i^{1:(t-1)}$
Ideal experiment	randomization at baseline	sequential randomization
Confounders handled	$\mathbf{X}_i^{1:T}$, unit/time FEs, interactive FEs	information in $\mathbf{X}_i^{1:t}$ and $\mathbf{Y}_i^{1:(t-1)}$
Identification concerns	feedback, especially from past outcomes	confounding must show in observables
Examples	DID, TWFE, LFM, SCM	LDV, ADL, MSM

Note: $\mathbf{U}_i^{1:T}$ is defined under Equation (8). LFM, LDV, ADL, MSM represent linear factor models, lagged dependent variables, autoregressive distributed lag, marginal structural models, respectively. Chad Hazlett contributes to this table.

Our discussion so far suggests that both identification regimes have pros and cons. In the next few sections, I review popular methods for TSCS data analysis that operates in one of

two identification frameworks. I will discuss drawbacks of each method while keeping in mind that an alternative view of the DGP (and especially the treatment assignment mechanism) exists.

Difference-in-Differences (DID)

The idea of DID design, which explores both within- and between-unit variability of sample, can be traced back the famous work of John Snow, the London doctor who studied the cause of cholera in the 1850s (Angrist and Pischke, 2009; Coleman, 2020). As far as I know, Ashenfelter (1978) first uses a DID approach to estimate the effect of job training programs on earnings. Card and Krueger (1994) formally introduce the classic two-group two-period DID design—hence, data are grouped into four cells—to the social sciences to study the effect of a minimum wage increase in New Jersey on employment in the fast food industry. The essence of the DID approach is to use treated units' pre-treatment data to difference out unobserved time-invariant attributes and to use the average before-and-after difference from a comparison group to adjust for the common time trend between the first and second periods.

A DID design. Without loss of generality, let's define $f_1 = 0$, $f_2 = \mathbb{E}[Y(0)_{i2} - Y_{i1}(0)|D_{i2} = 0]$, and $\alpha_i = \mathbb{E}[Y_{i1}(0)]$ for any unit i; we can rewrite the unconditional version of the parallel trends assumption in Equation (5), as

$$\mathbb{E}[Y_{it}(0)] = \alpha_i + f_t, \quad \forall i, t = 1, 2.$$

In other words, the parallel trends assumption without covariates is equivalent to assuming an outcome model in which the expected outcome is comprised of the two additive fixed effects. Therefore, a DID approach is scale-dependent, meaning that a strictly monotonic transformation of the outcome variable (e.g., from level to logarithmic terms) may render

the assumption invalid (Athey and Imbens, 2006).[3] We can further write a DID model as

$$Y_{it}(0) = \tau_{it} D_{it} + \alpha_i + f_t + \epsilon_{it}, \quad \forall i, t = 1, 2, \tag{4}$$

in which τ_{it} is the treatment effect for unit i at time t; ϵ_{it} represents idiosyncratic errors and $\mathbb{E}[\epsilon_{it}] = 0$. Equation (4) looks familiar because it resembles a TWFE model often appearing in political science journals. Many social scientists equate TWFE to DID possibly under the influence of Angrist and Pischke (2009). However, a crucial difference exists, that is, τ_{it} can be arbitrarily heterogeneous in a DID design but not in TWFE models. This difference will have important implications for the weighting problem we will discuss later.

Athey and Imbens (2018) provide a design-based perspective for the DID setting. They imagine a hypothetical experiment in which the treatment—a die indicating whether a unit gets treatment in period 2—is randomly assigned to each unit, and the probability of getting treated can depend observed or unobserved unit-specific attributes. This design is a type of baseline randomization, therefore is consistent with the strict exogeneity assumption.

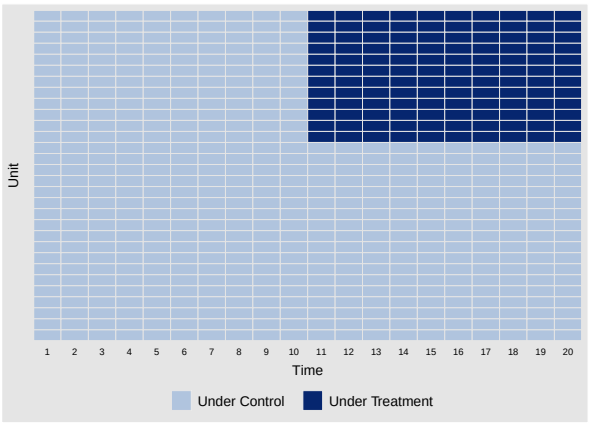
The DID estimator. Denote T_0 the number of pre-treamtment periods and G_i a dichotomous group indicator, indicating whether unit $i$belongs to the treatment group (i.e., $G_i = 1$ if $D_{i1} = 0, D_{i2} = 1$) or the control group (i.e., $G_i = 0$ if $D_{i1} = D_{i2} = 0$). The casual estimand is defined as the average treatment effect on the treated (ATT), i.e., $ATT = \mathbb{E}[\tau_{i2}|D_{it} = 1, G_i = 1]$. Under the parallel trend assumption and some regularity conditions, the DID estimator in Equation (5), which is based on four moment conditions, identifies the ATT.

$$\begin{aligned}\hat{\tau}^{DID} = (\hat{\mathbb{E}}[Y_{it}(0)|G_i = 1, t > T_0] - \hat{\mathbb{E}}[Y_{it}(0)|G_i = 1, t \leq T_0]) - \\ (\hat{\mathbb{E}}[Y_{it}(0)|G_i = 0, t > T_0] - \hat{\mathbb{E}}[Y_{it}(0)|G_i = 0, t \leq T_0]).\end{aligned} \tag{5}$$

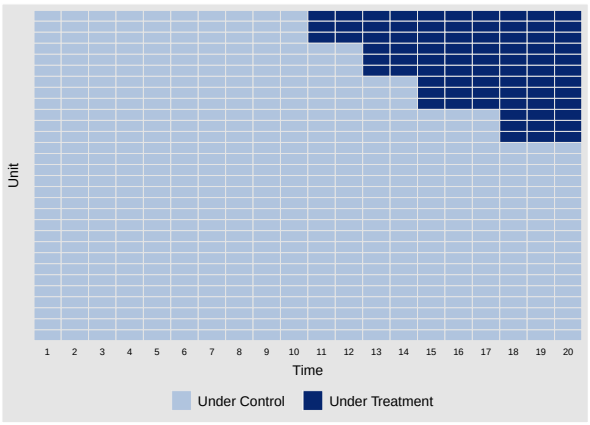
A natural extension of the classic DID design is to have multiple time periods both before

[3] Roth and Sant'Anna (2021) provide a set of sufficient conditions under which a DID design is scale independent.

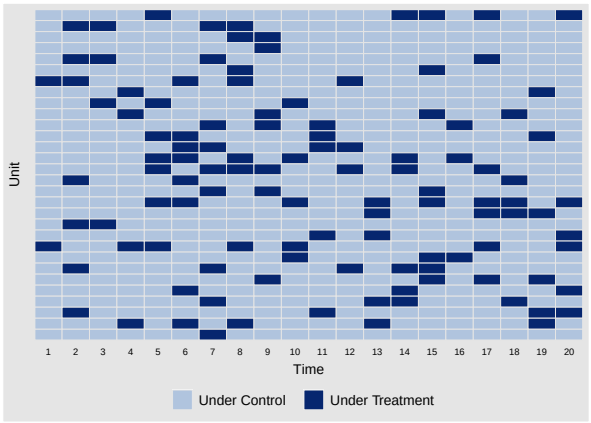
FIGURE 4. TREATMENT STATUS UNDER DIFFERENT DGPS



(a) Multi-period DID



(b) Staggered Adoption

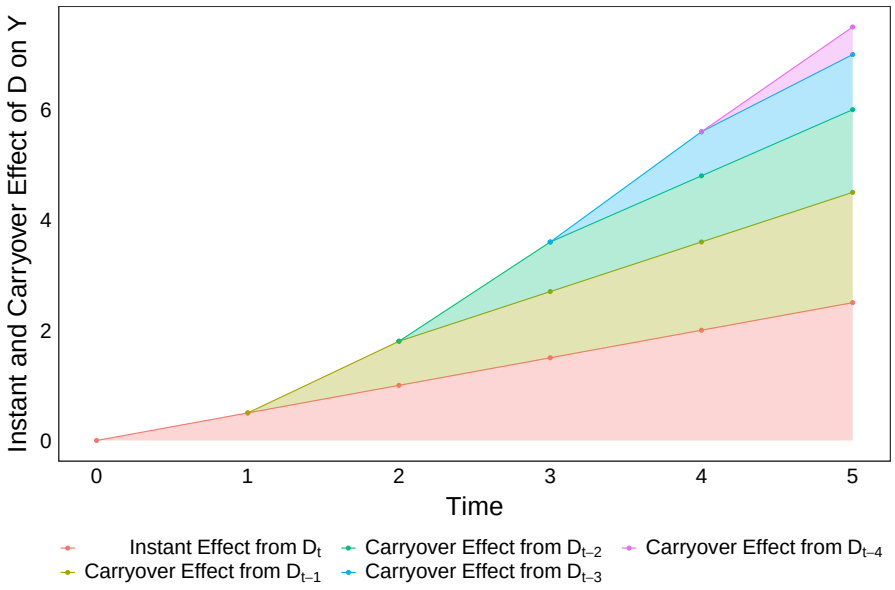


(c) General Pattern

Note: The above figures show the treatment status with three hypothetical examples: (a) multi-period DID, (b) staggered adoption and (b) a general panel treatment structure. Dark and light blue colors indicate treatment and control conditions, respectively.

and after the treatment starts, which I call a multi-period DID, as illustrated in Figure 4(a). Because the treatment starts at the same time, say, time $(T_0 + 1)$, for units in the treatment group, the same DID estimator $\hat{\tau}^{DID}$ identifies the ATT without imposing any homogeneity assumptions on τ_{it} (such as a constant treatment effect, i.e., $\tau_{it} = \tau$, which is assumed in most TWFE models). Moreover, because only two treatment histories are in comparison with each other, potential carryover effects of the treatment are allowed—the only thing a researcher needs to do is to relabel τ_{it} as the cumulative effect of all treatment before time t on unit i, as illustrated in Figure 5. One major benefit of a multi-period DID design is that it allow researchers to plot the so-call dynamic treatment effects plots, which may help

FIGURE 5. DAG FOR DGPS UNDER STRICT EXOGENEITY



Note: This figure is adapted from Figure A3 in Liu, Wang and Xu (2021). It demonstrates a decomposition of τ_{it} in a hypothetical case under a multi-period DID (or staggered adoption) design when carry over effects exist. The x-axis indicates the time relative the onset of the treatment.

researchers gauge how plausible the parallel trends assumption is. Egami and Yamauchi (2021) leverage on multiple time periods and propose a double DID based on a parallel trends-in-trends assumption to address potential failures of parallel trends.

If the treatment kicks in at different time for the treated units but never switches back, as illustrated in Figure 4(b), we call it a staggered DID design. Such cases are common in political science applications, for example, policy and institutions as treatments may be adopted at different time periods, but once adopted, they are difficult to change back. A staggered DID design can be decomposed into multiple smaller DIDs, whose ATTs can be identified using the DID estimator (Goodman-Bacon, 2021; Callaway and Sant'Anna, 2020). However, most political scientists use TWFE models to estimate treatment effects under staggered adoption. This practice has important consequences, which we will discuss in the next section.

Another direction for extensions is to incorporate pre-treatment covariates. Abadie (2005) introduces semiparametric DID which takes two steps. In the first step, it estimates

a propensity score model; in the second step, it runs a weighted DID model using estimated inverse propensity scores as weights. The paper shows that this approach reveals the ATT under a conditional parallel trends assumption (Equation 5). Strezhnev (2018) extends this approach to the staggered adoption setting.

Twoway Fixed Effects (TWFE)

TWFE models are probably the most commonly used statistical routines when it comes to TSCS/panel data analysis in the social sciences. It was formally introduced to social sciences by Mundlak (1978) and Chamberlain (1982) and popularized by famous textbooks, such as Wooldridge (2001) and Hsiao (2003). Compared with multilevel models with random effects, TWFE models are less susceptible to biases caused by correlations between random effects and regressors (Hsiao, 2003; Hazlett and Wainstein, 2020). Ever since Angrist and Pischke (2009, pp. 236–243) interpret the TWFE model as "regression DD" method that are suitable for causal inference with a more general panel data setting (such as the one illustrated in Figure 4(c) where the treatment switches on and off), it has become a workhorse model for applied TSCS analysts. A TWFE model usually takes the following form:

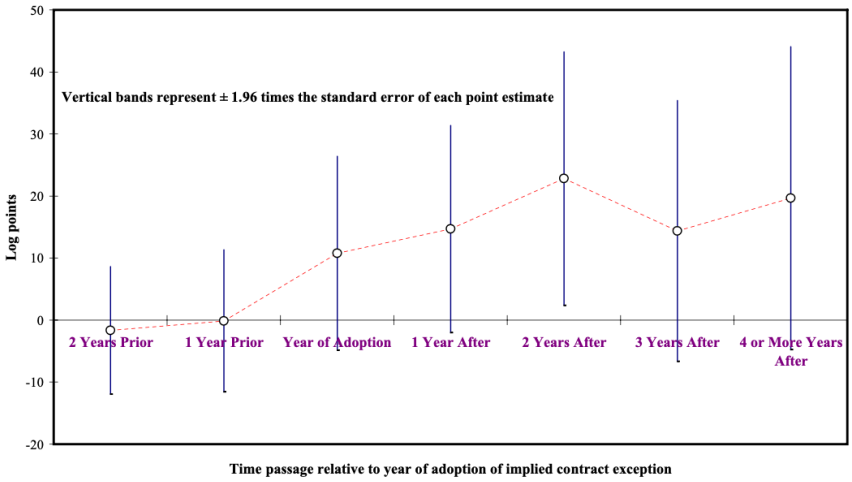
$$Y_{it} = \tau D_{it} + \mathbf{X}'_{it}\beta + \alpha_i + f_t + \epsilon_{it}, \quad \forall i, t. \tag{6}$$

in which $\mathbf{X}_{it}$ represents a vector of covariates; β is a vector of coefficients for the covariates; and τ is the casual quantity of interest. Let's denote the TWFE estimate based on this model $\hat{\tau}^{TWFE}$.

Violations of strict exogeneity, diagnostics, and design-based approaches. There are several limitations with this approach. First, assuming that the treatment effect is constant, the casual identification of τ hinges on the strict exogeneity assumption; as a result, any violation of the four requirements we discussed earlier will lead to biases in the estimate. The requirement of no time-varying confounder is well acknowledged by researchers

and they often rely the aforementioned dynamic treatment effects plot to gauge its validity (see Figure 6 for an example). Although research has shown that such estimates may be biased when the evolution of treatment effects are heterogeneous across units (Sun and Abraham, 2018) and cannot differentiate diverging trends and anticipation effect in linear forms (Borusyak, Jaravel and Spiess, 2021), these plots have been very helpful in revealing potential time-varying confounders. It can also help researchers diagnose whether potential

FIGURE 6. EXAMPLE OF A DYNAMIC TREATMENT EFFECT PLOT



Note: This figure is adapted from Angrist and Pischke (2009, p. 529). It is created based on coefficients from a TWFE model with a set of leads and lags of the treatment indicator.

feedback from past outcomes to current treatment exists (e.g., by observing whether there is a pre-trend leading toward the onset of the treatment).

The requirement for the absence of lagged dependent variables (LDVs) is often invalid with most political economy data. Theoretically speaking, as long as LDVs are not correlated with the probability of getting treated, not controlling for it will not cause biases in the TWFE estimate, but only efficiency. Beck and Katz (2011) argue that the worry about the Nickell bias (Nickell, 1981) is exaggerated with TSCS data and including both fixed effects and LDVs causes little biases when T is big. Empirical practice in political science has not yet catch up with their advice maybe due to the fact that estimated effects will be more difficult to interpret when LDVs are at present.

The requirement for no carryover effects have received much less attention. Unlike in multi-period DID or staggered DID designs where the number of treatment histories is limited, violations of no carryover effects will bias the causal estimates because it is a type of temporal interference, that is, the potential outcome at time t is affected by treatment status at time $(t-1)$ or earlier. Although this assumption are partially testable by data (e.g. through a placebo test) and can be relaxed by adding lagged terms of the treatment or by not used these periods in constructing a predictive model for $Y_{it}(0)$ (e.g., Liu, Wang and Xu, 2021), such practices are currently uncommon in applied research.

Another research frontier for TWFE models is to link the model-based approach with a design-based perspective. For example, Arkhangelsky and Imbens (2019) propose a method that breaks the connection between treatment assignment and unmeasured confounders by conditioning on a sufficient statistics guided by experimental design or economic theories; Arkhangelsky et al. (2021) develops a doubly robust TWFE estimator that weights each unit based on known or estimated treatment assignment mechanisms and obtains the average treatment effect in large samples even when the parallel trends assumption is not satisfied.

The weighting problem and solutions. The consequences of heterogeneity treatment effects in TWFE models, i.e., $\tau_{it} \neq \tau$, have received a lot of attention recently in applied statistics and econometrics. Chernozhukov et al. (2013, Theorem 1) show that a unit fixed effect model give a variance-weighted average of the treatment effect, analogous to what Aronow and Samii (2016) find in a cross-sectional setting.

Strezhnev (2018) extends this result to TWFE models under staggered adoption and shows that the highest weights are given to observations having the highest variance both cross-sectionally and temporally (around the time when treatment status changes). A series papers find similar results but use slightly different approaches (Sun and Abraham, 2018; Goodman-Bacon, 2021; Callaway and Sant'Anna, 2020; de Chaisemartin and D'Haultfœuille, 2020; Borusyak, Jaravel and Spiess, 2021). For example, Goodman-Bacon (2021) decomposes

staggered adoption data into smaller multi-periods DIDs and shows that the probability limit of $\hat{\tau}^{TWFE}$ can be interpreted as a variance-weighted ATTs when the parallel trends assumption is satisfied in each DID and when the treatment effect remain the same over time for each unit. He illustrates that the key to the problem is that a TWFE model uses units that adopt the treatment early (early adopters) as controls for units that adopt the treatment later (late adopters), which incurs biases when the treatment effect evolves over time. Using a similar logic, de Chaisemartin and D'Haultfœuille (2020) decompose $\hat{\tau}^{TWFE}$ into weighted average of $\{\hat{\tau}_{it}\}_{D_{it}=1}$. They show that some of the weights can be negative because later adopters serve as controls for early adopters. This is problematic because $\hat{\tau}^{TWFE}$ may not even be a convex combination of $\{\hat{\tau}_{it}\}_{D_{it}=1}$. Similar problem naturally extends panel/TSCS data in a more general setting. This issue also plagues fuzzy DID designs where units falling into the treatment group have higher chancesof getting treated, but not 100% (de Chaisemartin and D'Haultfœuille, 2017).

Researchers also provide several potential solutions. Strezhnev (2018) and Callaway and Sant'Anna (2020) suggest that one can estimate the ATT for each "cohort" of units that receive the treatment at the same time, thus circumventing the problem all together. A potential problem with this approach is the lack of power with two few units in each cohort. de Chaisemartin and D'Haultfœuille (2020) propose a DID_M estimate that only use observations right before and after the treatment starts. Imai and Kim (2019); Imai, Kim and Wang (2021) suggest a matching method to difference out unit fixed effects. The downside of these approaches is that they drop many observations. Wooldridge (2021) suggests that researchers can overcome the negative weighting problem by meticulously adding interaction terms to a Mundlak regression, where within-unit or within- period averages of regressors are directly included in an OLS regression. Arkhangelsky and Imbens (2019) propose a doubly-robust estimator that combines a reweighting approach with outcome modeling. It gives a consistent estimate for an estimand that is convex combination of treatment effects if either the weights or the outcome model is correctly specified. Liu, Wang and Xu (2021) and

Borusyak, Jaravel and Spiess (2021) independently recommend a counterfactual approach that uses observations under the control condition only to build a predictive model (such as TWFE) for treated counterfactual, i.e.,:

$$Y_{it}(0) = \mathbf{X}'_{it}\beta + \alpha_i + f_t + \epsilon_{it}, \quad \forall i, t, \tag{7}$$

which is a natural generalization of the DID model specified in Equation (5) because it imposes no restrictions on treatment effect homogeneity. This method has the advantage of keeping the time-series structure and accommodating more general panel data settings. They also suggest several diagnostic tests, including a new dynamic treatment effect plot, to help research gauge the validity of the strict exogeneity assumption. The main limitation is that it relies on correction model specifications.

The Synthetic Control Method (SCM)

Researchers see the SCM (Abadie and Gardeazabal, 2003; Abadie, Diamond and Hainmueller, 2010, 2015) as one of the most important innovations in the policy evaluation literature in recent years (Athey and Imbens, 2016). The SCM is suitable for comparative case studies, which use TSCS data of a small number of aggregate units; among them, one is treated with an intervention at a given time period (Abaide, 2020). The basic idea of the SCM is to use a convex combination of the control unit trajectories to serve as the counterfactual trajectory for the treated unit in the post-treatment period; the weights are chosen to achieve (approximate) balance on pre-treatment covariates and outcomes between the treated and reweighted control units. The SCM is a path-breaking innovation for casual inference because it allows users to construct counterfactuals using very limited data; it also has the great advantage of being transparent and producing easily interpretable estimates.

Although past outcomes are incorporated in matching process, from a design perspective, I see the SCM as a method operating under strict exogeneity because the timing of the treatment is seen as either fixed or quasi-random, and not directly affected by past

outcomes—hence, the error terms are time-exchangeable. The authors motivate the method with a linear factor model (LFM) to and justifies matching on pre-treatment outcomes as a way to difference out the heterogeneous impact of time-varying factors.

Challenges. Applying the SCM faces several challenges. First, the SCM algorithm may not be able to find a solution if the trajectory of the treated unit is outside the convex hull of the control unit trajectories. This is not necessarily a weakness of the SCM because it rules out negative weights and ensures overlap, but it does limit the applicability of the method (Hollingsworth and Wing, 2020). In my view, the real problem in applying the SCM is that it allows a lot of user discretion—research has shown that cherry-picking on the covariates (including the pre-treatment outcomes) to make the pre-treatment fit look better leads to spurious findings (Ferman, Pinto and Possebom, 2020). Cattaneo, Feng and Titiunik (2019) shows that researchers tend to be over-confident in their SCM results when the data in use are nonstationary.

The second challenge of applying SCM is to obtain valid inference. Because the intervention is usually not randomly assigned across units, one cannot interpret results from the leave-one-out placebo test recommended by the authors as results of a permutation test (Hahn and Shi, 2017). Recent research has made progress on inferential problem of the SCM based on sensitivity analysis (Firpo and Possebom, 2018), exact test via permutation over time (Chernozhukov, Wuthrich and Zhu, 2017), predictive inference (Cattaneo, Feng and Titiunik, 2019), or Bayesian inference (Kim, Lee and Gupta, 2020).

Extensions. A group of new methods intend to improve the SCM on both fronts. On the algorithmic side, Doudchenko and Imbens (2016) provide a framework to nest several extensions of the SCM. They argue that these new algorithm differ in the constraints they put on the control group weights and in whether they allow an intercept shift. As a benchmark, the SCM does not allow intercept shift and force the weights to be non-negative and add up to 1 (convexity). On the other hand, the best subset approach (Hsiao, Steve Ching and

Wan, 2012), as well as the elastic net approach proposed by the authors, allows intercept shift and relaxes the convexity constraint. The balancing approach proposed by Robbins, Saunders and Kilmer (2017) imposes convexity but is more computational efficient than the SCM when there are multiple treated units. Hazlett and Xu (2018) replace mean-balancing with kernel balancing to take into account higher-order terms of the pre-treatment variables and make the parallel trends assumption more credible.

Researchers have also attempted to combine the SCM (as a reweighting method) with outcome modeling, thus achieving double robustness (or having smaller biases than either the SCM or the outcome model alone). For example, Ben-Michael, Feller and Rothstein (2018) combines a ridge regression with an SCM applied to the residual time series of the ridge. Arkhangelsky et al. (2018) propose a weighted DID model, in which observation-specific weights are obtained by conducting the SCMs in both the temporal and cross-sectional dimensions.

Another strand of the literature uses factor-augmented models to extend the SCM to large-N, large-T settings. We will discuss these studies in the next section.

The Factor-Augmented Approach

When both N and T are large, the factor-augmented approach relaxes the conventional strict exogeneity assumption, Equation (1), to the following general form:

(Strict exogeneity: general form)

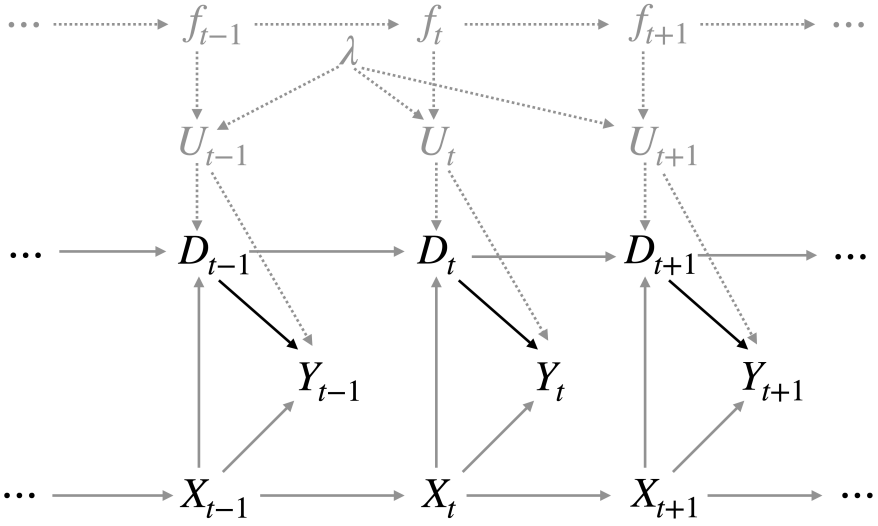
$$\{Y_{it}(0), Y_{it}(1)\} \perp\!\!\!\perp D_{is} \mid \mathbf{X}_i^{1:T}, \mathbf{U}_i^{1:T}, \quad \forall i, t, s, \tag{8}$$

in which $\mathbf{U}_i^{1:T} = \{\mathbf{U}_{i1}, \cdots, \mathbf{U}_{it}\}$ is the histories of a vector of unobserved confounders. Accordingly, we introduce a more general form of the outcome model:

$$Y_{it} = \tau_{it} D_{it} + g(\mathbf{X}_{it}) + h(\mathbf{U}_{it}) + \epsilon_{it}, \ \forall i, t \tag{9}$$

in which $g(.)$ and $h(.)$ are two functions of known forms. Note that this model still assumes additive separability between the four items. The main challenge to estimate this model under the more general form of strict exogeneity is that $\mathbf{U}_{it}$ is unobserved. We get around this problem in DID and TWFE models by decomposing it into two additive fixed effects, i.e., $h(\mathbf{U}_{it}) = \alpha_i + f_t$. Using a factor-augmented approach, we assume $h(\mathbf{U}_{it}) = \lambda_i' f_t$, in which f_t and λ_i are a $(r \times 1)$ vectors of factors and factor loadings. $\lambda_i' f_t$, referred to as interactive fixed effects (IFE) in the econometrics literature (e.g., Bai and Ng, 2002; Bai, 2009), is a lower-rank representation of $\mathbf{U}_{it}$ and, thus, is estimable from data under mild conditions. It is easy to see that additive fixed effects α_i and f_t are its special cases.

FIGURE 7. DAG FOR DGPS UNDER STRICT EXOGENEITY: AUGMENTED BY FACTORS



Note: Subscript i (for unit) is omitted for simplicity. Black solid arrows are causal relationships of researchers' interest. Gray solid arrows and gray dashed arrows represent relationship among observed variables and relationships involving unobserved variables, respectively.

Figure 7 shows a DAG consistent with Equations (8) and (9). It shows that if researchers do not condition on λ_i and f_t, $\mathbf{U}_{it}$ will act as time-varying confounders in estimating the causal effect of D_{it} on Y_{it}. It is also important to note the lack of arrows from, say, Y_{t-1} to U_t or Y_{t-1} to X_t, which would otherwise result in feedback from past outcomes to current treatment. Because λ_i is pre-determined (just like α_i in a TWFE model) and $\mathbf{f}^{1:T}$ is not affect by the history of Y, treatment assignment can be seen as pre-specified. Hence, it is

still under the general framework of strict exogeneity.

The factor-augmented approach was first introduced to the DID setting by Gobillon and Magnac (2016). Xu (2017) proposes a slightly different version of the method that uses never-treated units only to learn the factors and develops automated tuning parameter selection and methods for inference. Athey et al. (2021) generalize this approach with the matrix-completion method from the computer science literature and introduce regularization on the nuclear norm (i.e., the sum of singular values) of error-term matrix. Liu, Wang and Xu (2021) propose diagnostics tests under this framework and provide practical advice on implementation. Researchers have also developed Bayesian versions of this method to accommodate multiple outcomes (Samartsidis et al., 2020), to facilitate variable and factor selection (Pang, Liu and Xu, 2021), to model time series of factors using Gaussian Processes (Arbour et al., 2021), while at the same time offering valid Bayesian inference. Researchers have also extended this approach to combine factor analysis with sparse regressions (Fan, Masini and Medeiros, 2020) or to incorporate possibly nonlinear factors (Feng, 2020). Imbens, Kallus and Mao (2021) propose a GMM estimator to control for unmeasured confounders in TWFE and IFE models using bridge functions which does not require a large T, but requires the post-treatment error terms to be independent of error terms in previous periods. Using time-invariant covariates as instruments (whose effects on the outcome are assumed to be constant), Callaway and Karami (2021) develop a method to estimate treatment effects in IFE models with short panels.

Methods under Sequential Ignorability

We now move onto methods under sequential ignorability. The key advantage of this identification regime is that it allows treatment assignment to be determined sequentially in response to changes in the outcome and covariates which may be affected by past treatment, as illustrated in Figure (3). So far, social scientists have largely ignored such dynamics.

In the biomedical literature, however, researchers have paid more attention to this issue because, for example, treatment plans are often designed dynamically based on patients' medical conditions. The main limitation of the sequential ignorability approach is that it usually does not allow the presence of observed confounders.[4]

Identification under sequential ignorability face the challenge of adjusting for the dynamics without causing bias. Using an example adapted from Blackwell and Glynn (2018), we assume a researcher estimate the following regression model consistent with Figure (3):

$$Y_{it} = \beta_0 + \alpha Y_{t-1} + \beta_1 D_{it} + \beta_2 D_{i,t-1} + X_{it} + \epsilon_{it}.$$

Even if the linear functional form is correct, β_2 will be inconsistently estimated because X_{it} is affected by $D_{i,t-1}$ (thouhg β_1 will be consistently estimated). The literature has proposed several methods to deal with this problem. Below I discuss two: panel matching and marginal structural models (MSMs).

Panel matching. Imai, Kim and Wang (2021) propose panel matching, which focus on estimating the marginal effect of the treatment D_{it} on current and future outcomes $Y_{i,t+F}$ ($F \geq 0$). The method assumes no cross-sectional spillover but allow treatment reversal and limited carryover effects. The procedure takes three steps. First, for a transition from the control condition at time $(t-1)$ to the treatment condition at time t, the algorithm finds a set of matched units sharing the exact same treatment history from $(t-L)$ to $(t-1)$, in which L is a small integer. Second, it refines the matched set using a matching or reweighting method such that the units in the comparison group share similar pre-treatment covariates (including lagged outcomes) with the unit that experiences the transition. Third, it computes the ATT using the refined set using either a difference-in-means estimator or a DID estimator. Inference is made using block bootstrap or theoretical approximation.

The biggest advantages of panel matching are its weak assumptions (compared with methods under strict exogeneity), its applicability to many user cases, and its ease of im-

[4]See Blackwell and Yamauchi (2021) for an exception, in which the authors assume a sequential ignorability assumption that condition on time-invariant unobservables but not past outcomes.

plementation. It main limitations, in my view, are two-fold: (1) unlike MSMs, it does not disentangle the direct and indirect effects of past treatment on current and future outcomes; and (2) its breaks the TSCS data structure and drops a lot of data. An additional concern, though it applies to many other panel methods, is that many aspects of the procedure require user discretion, e.g., what covariates to enter the balancing scheme, how to weight these covariates, and how much imbalance should be tolerated.

MSM. MSM is a popular method to estimate causal effects of time-varying treatments in epidemiology and biomedical sciences (Robins, Hernán and Brumback, 2000; Hernán, Brumback and Robins, 2001). MSMs were introduced to political science by Blackwell (2013) and Blackwell and Glynn (2018), and since then, have been applied in several social science settings (e.g., Ladam, Harden and Windett, 2018; Kurer, 2020). The MSMs are "marginal" because they model for the marginal means of the potential outcomes as a function of treatment histories.

To illustrate the MSM approach, denote $d_{1:t}$ and $d'_{1:t}$ two treatment histories and define a causal estimand of interest $\tau(d_{1:t}, d'_{1:t}) = \mathbb{E}[Y_{it}(d_{1:t}) - Y_{it}(d'_{1:t})]$ the average causal effect of treatment history $d_{1:t}$ as compared to $d'_{1:t}$. An MSM attempts to flexibly estimate $\mathbb{E}[Y_{it}(d_{1:t})$ (and $\mathbb{E}[Y_{it}(d'_{1:t})]$) using a flexible function $g(d_{1:t}; \beta)$, in which β is a set of coefficients. The main challenge is that the relationship between $d_{1:t}$ and Y_{it} is confounded by time-varying covariates and past outcomes, which are post-treatment, as explained earlier. One solution is to use inverse propensity of treatment weighting (IPTW) to balance out their influences across different values of the the treatments, thus removing biases caused by these confounders (Robins, Hernán and Brumback, 2000; Blackwell and Glynn, 2018). The limitations of this approach, in addition to not accommodating unit fixed effects, include strong modeling assumptions for in the propensity scores estimation and unstable weights, which are to be addressed in future research.

Discussion

Causal inference with TSCS data is one of the fastest-developing subfields in the causal inference literature in recent years. Researchers have been focusing on understanding the nature of commonly used estimators, such as DID and TWFE, relaxing the functional form assumptions (e.g. to accommodate heterogeneous treatment effects), and to more formally incorporate a design-based perspective. They have made significant progress on these fronts. Below I provide some practical advice based on existing research and then discuss several open questions.

Recommendations for empirical research. As mentioned earlier in this chapter, the biggest pitfall with current common practice is the lack of attention paid to treatment assignment mechanisms. Therefore, my first and foremost suggestions is for researchers to adopt a design-based perspective and think about the hypothetical experiments when using panel/TSCS to establish causality. If, for example, we believe that the dynamic feedback effect from past outcomes and covariates to current treatment assignment is particularly strong, we should opt for methods operating under sequential ignorability instead of strict exogeneity or its variants. This is no easy task because researchers using observational data usually lack the precise knowledge of how the treatment is assigned. One approach is to exploit the "bracketing relationship" between the fixed effects approach and LDV adjustments (e.g., Angrist and Pischke, 2009; Ding and Li, 2019) and check whether the empirical findings are robust to different identification assumptions and model specifications. Alternatively, carefully examining raw data helps researcher spot obvious issues. Hence, my second recommendation is to plot raw data using data visualization tools, especially for the evolution of treatment status and outcome variables, as well as their relationships.[5] Third, after applying an estimator researchers deem appropriate for their data under presumed assump-

[5]My collaborators and I developed a package called `panelView` in both `Stata` and `R` for that purpose. See https://yiqingxu.org/software/ for more details.

tions, they should present the results in a transparent and accessible way—again, a visual check will be extremely helpful. My fourth and last recommendation is to conduct diagnostic tests, including placebo tests and sensitivity analysis, which are crucial for boosting readers' confidence in the identifying assumptions.

Despite the recent progress, many questions remain in this burgeoning field. Next, I discuss several issues that have received relative little attention and present research opportunities.

Taking time and dynamics more seriously. Political methodologists have long emphasized the importance of the temporal dimension of TSCS data and argued the big T compared with N in such data presents unique challenges to researchers (Beck and Katz, 1995; Beck, Katz and Tucker, 1998; Beck, 2008). Although properly standard errors adjustment has become standard practice in the social sciences, few researchers pay attention to stationarity issues in a TSCS setting. Moreover, with only a few exceptions (e.g., Carlson, 2018; Tyler, 2021; Chen et al., 2021; Arbour et al., 2021), most newly emerged methods do not take sufficient advantage of the time-series aspect of the data in an effort to model either the outcome or treatment assignment more precisely—for example, the SCM, as well as most of its extensions, primarily uses cross-sectional correlation of units to predict counterfactuals for the treated. Last not but least, methods under sequential ignorability, such as the MSMs, have not entered the toolkit of most political scientists possibly due to a lack of understanding of the challenges they address.

SUTVA violations. Interference across both the temporal and cross-section dimensions are prevalent in social science applications. Examples of temporal interference includes the anticipation effect, i.e., potential outcome affected by the anticipation of the arrival of a future treatment, and the carryover effect. Examples of cross-sectional interference includes spatial spillover and general equilibrium effects. In spatial experimental setting, Aronow, Samii and Wang (2020) propose a method that estimates average treatment effects along

distance to experimental locations to account for potential spatial interference. Wang (2021) extends this idea to observational setting using a IPTW method and considers both temporal and spatial inference. Sanford (2021) studies the effect of more secure property rights on agricultural outcomes using remote sensing data, addressing similar issues with geographic spillover effects. More research is needed to fully explore both types of interference in TSCS data.

A second type of SUTVA violation is multiple treatment types being lumped together as one treatment.[6] For example, students of voter ID laws in the United States often see these laws as a dichotomous treatment, e.g., whether a state has passed such a law. In reality, however, states adopting voter ID laws implement different policies with varying degrees of enforcement (e.g., Mycoff, Wagner and Wilson, 2009). Treating them as the same policy violates SUTVA, a problem often neglected in the current empirical literature.

More complex treatment assignment. In many social science setting, the treatment assignment mechanisms are more complex than what is covered in this chapter. The most obvious case is continuous treatments, or treatment with incremental intensities. Few studies have formally dealt with continuous treatment from a design-based perspective—with a few exceptions (e.g., Callaway, Goodman-Bacon and Sant'Anna, 2021; Squires et al., 2020). Another example is policy diffusion, in which case a locality's treatment status affects its neighbor treatment status (e.g., Egami, 2018). Researchers need to develop new methods are to accommodate these important scenarios.

In this chapter, I clarify two identification regimes under which researchers commonly establish causal relationships using observational TSCS data: One is based on the strict exogeneity assumption, which corresponds to baseline randomization; the other is based on the sequential ignorability assumption, which corresponds to sequential randomization. I discuss three most widely used methods, including DID, TWFE, and SCM, as well as several alternatives,

[6] I thank Jonathan Katz for raising this point.

such as LFM, MSM and PanelMatch. Future research is needed to accommodate more complex user cases, such as dynamics in the treatment assignment, continuous treatment, and various types of SUTVA violations.

References

Abadie, A and J Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *The American Economic Review* 93(1):113–132.

Abadie, Alberto. 2005. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies* 72(1):1–19.

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105(490):493–505.

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59(2):495–510.

Abaide, Alberto. 2020. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." *Journal of Economic Literature* .

Angrist, Josh D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arbour, David, Eli Ben-Michael, Avi Feller, Alex Franks and Steven Raphael. 2021. "Using Multitask Gaussian Processes to Estimate the Effect of a Targeted Effort to Remove Firearms.".

Arkhangelsky, Dmitry and Guido W Imbens. 2019. "Double-Robust Identification for Causal Panel Data Models.".

Arkhangelsky, Dmitry, Guido W Imbens, Lihua Lei and Xiaoman Luo. 2021. "Double-Robust Two-Way-Fixed-Effects Regression For Panel Data.".

Arkhangelsky, Dmitry, Susan Athey, David A Hirshberg, Guido W Imbens and Stefan Wager. 2018. "Synthetic Difference in Differences.".

Aronow, Peter M and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science* 60(1):250–267.

Aronow, Peter M, Cyrus Samii and Ye Wang. 2020. "Design-Based Inference for Spatial Experiments with Interference.".

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *The Review of Economics and Statistics* 60(1):47–57.

Athey, Susan and Guido Imbens. 2016. "The State of Applied Econometrics - Causality and Policy Evaluation.".

Athey, Susan and Guido W Imbens. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica: journal of the Econometric Society* 74(2):431–497.

Athey, Susan and Guido W Imbens. 2018. "Design-based Analysis in Difference-in-Differences Settings with Staggered Adoption.".

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens and Khashayar Khosravi. 2021. "Matrix Completion Methods for Causal Panel Data Models." *Journal of the American Statistical Association* forthcoming.

Bai, Jushan. 2009. "Panel Data Models with Interactive Fixed Effects." *Econometrica: journal of the Econometric Society* .

Bai, Jushan and Serena Ng. 2002. "Determining the Number of Factors in Approximate Factor Models." *Econometrica: journal of the Econometric Society* .

Beck, Nathaniel. 2008. Time-Series Cross-Section Methods. In *Oxford Handbook of Political Methodology*, ed. Janet M Box-Steffensmeier, Henry E Brady and David Collier. Oxford University Press.

Beck, Nathaniel and Jonathan N Katz. 1995. "What to Do (and Not to Do) with Time-Series Cross-Section Data." *The American political science review* 89(3):634–647.

Beck, Nathaniel and Jonathan N Katz. 2001. "Throwing Out the Baby with the Bath Water: A Comment on Green, Kim, and Yoon." *International organization* 55(2):487–495.

Beck, Nathaniel and Jonathan N Katz. 2004. Time-Series–Cross-Section Issues: Dynamics, 2004. In *Annual Meeting of the Society for Political Methodology*.

Beck, Nathaniel and Jonathan N Katz. 2011. "Modeling Dynamics in Time-Series–Cross-Section Political Economy Data." *Annual Review of Political Science* 14(1):331–352.

Beck, Nathaniel, Jonathan N Katz and Richard Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* 42(4):1260–1288.

Ben-Michael, Eli, Avi Feller and Jesse Rothstein. 2018. "The Augmented Synthetic Control Method.".

Blackwell, Matthew. 2013. "A Framework for Dynamic Causal Inference in Political Science." *American journal of political science* 57(2):504–520.

Blackwell, Matthew and Adam Glynn. 2018. "How to Make Causal Inferences with Time-Series Cross-Sectional Data Under Selection on Observables." *The American political science review* 112(2):1067–1082.

Blackwell, Matthew and Soichiro Yamauchi. 2021. "Adjusting for Unmeasured Confounding in Marginal Structural Models with Propensity-Score Fixed Effects.".

Borusyak, Kirill, Xavier Jaravel and Jann Spiess. 2021. "Revisiting Event Study Designs: Robust and Efficient Estimation.".

Callaway, Brantly, Andrew Goodman-Bacon and Pedro H C Sant'Anna. 2021. "Difference-in-Differences with a Continuous Treatment.".

Callaway, Brantly and Pedro H C Sant'Anna. 2020. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* .

Callaway, Brantly and Sonia Karami. 2021. "Treatment Effects in Interactive Fixed Effects Models." Mimeo, University of Georgia.

Card, David and Alan B Krueger. 1994. "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.

Carlson, David George. 2018. Advanced Methods in Comparative Politics: Modeling without Conditional Independence PhD thesis Washington University in St. Louis.

Cattaneo, Matias D, Yingjie Feng and Rocio Titiunik. 2019. "Prediction Intervals for Synthetic Control Methods.".

Chamberlain, Gary. 1982. "Multivariate Regression Models for Panel Data." *Journal of Econometrics* 18(1):5–46.

Chen, Yehu, Roman Garnett, Jacob M Montgomery and Annamaria Prati. 2021. "Difference in Differences with Parralel"ish" Trends via Joint Gaussian Processes." Mimeo, Washington University in St. Louis.

Chernozhukov, Victor, Ivan Fernandez-Val, Jinyong Hahn and Whitney Newey. 2013. "Average and Quantile Effects in Nonseparable Panel Models." *Econometrica* 81(2):535–580.

Chernozhukov, Victor, Kaspar Wuthrich and Yinchu Zhu. 2017. "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls.".

Coleman, Thomas. 2020. "John Snow and Cholera: Revisiting Difference-in-Differences and Randomized Trials in Research." Mimeo, University of Chicago.

de Chaisemartin, Clément and Xavier D'Haultfœuille. 2017. "Fuzzy Differences-in-Differences." *The Review of economic studies* 85(2):999–1028.

de Chaisemartin, Clément and Xavier D'Haultfœuille. 2020. "Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects." *The American Economic Review* 110(9):2964–2996.

Deaton, Angus and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social science & medicine* 210:2–21.

Ding, Peng and Fan Li. 2019. "A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment." *Political Analysis* 27(4):605–615.

Doudchenko, Nikolay and Guido W Imbens. 2016. "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis.".

Egami, Naoki. 2018. "Identification of causal diffusion effects under structural stationarity.".

Egami, Naoki and Soichiro Yamauchi. 2021. "Using Multiple Pre-treatment Periods to Improve Difference-in-Differences and Staggered Adoption Designs." *Political analysis: an annual publication of the Methodology Section of the American Political Science Association* forthcoming.

Fan, Jianqing, Ricardo P Masini and Marcelo C Medeiros. 2020. "Do We Exploit all Information for Counterfactual Analysis? Benefits of Factor Models and Idiosyncratic Correction.".

Feng, Yingjie. 2020. "Causal Inference in Possibly Nonlinear Factor Models.".

Ferman, B, C Pinto and V Possebom. 2020. "Cherry Picking with Synthetic Controls." *Journal of Policy Analysis and Management* .

Firpo, Sergio and Vitor Possebom. 2018. "Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets." *Journal of Causal Inference* 6(2).

Gobillon, Laurent and Thierry Magnac. 2016. "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls." *Review of Economics and Statistics* 98(3):535–551.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics* .

Hahn, Jinyong and Ruoyao Shi. 2017. "Synthetic Control and Inference." *Econometrics* 5(4):52.

Hazlett, Chad and Leonard Wainstein. 2020. "Understanding, Choosing, and Unifying Multilevel and Fixed Effect Approaches." *Political analysis: an annual publication of the Methodology Section of the American Political Science Association* .

Hazlett, Chad and Yiqing Xu. 2018. "Trajectory Balancing: A General Reweighting Approach to Causal Inference With Time-Series Cross-Sectional Data." Mimeo, Stanford University.

Hernán, Miguel A, Babette Brumback and James M Robins. 2001. "Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments." *Journal of the American Statistical Association* 96(454):440–448.

Hollingsworth, Alex and Coady Wing. 2020. "Tactics for Design and Inference in Synthetic Control Studies: An Applied Example Using High-Dimensional Data.".

Hsiao, Cheng. 2003. *Analysis of Panel Data*. Cambridge University Press.

Hsiao, Cheng, H Steve Ching and Shui K I Wan. 2012. "A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Economic Integration of Hong Kong with Mainland China." *Journal of Applied Economics* 27:705–740.

Imai, Kosuke and In Song Kim. 2019. "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science* 63(2):467–490.

Imai, Kosuke, In Song Kim and Erik Wang. 2021. "Matching Methods for Causal Inference with Time-Series Cross-Section Data." *American Journal of Political Science* (forthcoming).

Imbens, Guido, Nathan Kallus and Xiaojie Mao. 2021. "Controlling for Unmeasured Confounding in Panel Data Using Minimal Bridge Functions: From Two-Way Fixed Effects to Factor Models.".

Imbens, Guido W. 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48(2):399–423.

Keele, Luke. 2020. Differences-in-Differences: Neither Natural nor an Experiment. In *The SAGE Handbook of Research Methods in Political Science and International Relations*, ed. Luigi Curini and Robert Franzese. SAGE Publications Ltd.

Kim, Sungjin, Clarence Lee and Sachin Gupta. 2020. "Bayesian Synthetic Control Methods." *JMR, Journal of marketing research* 57(5):831–852.

Kurer, Thomas. 2020. "The Declining Middle: Occupational Change, Social Status, and the Populist Right." *Comparative Political Studies* 53(10-11):1798–1835.

Ladam, Christina, Jeffrey J Harden and Jason H Windett. 2018. "Prominent Role Models: High-Profile Female Politicians and the Emergence of Women as Candidates for Public Office." *American Journal of Political Science* 62(2):369–381.

Liu, Licheng, Ye Wang and Yiqing Xu. 2021. "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data." Mimeo, Stanford University.

Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica: journal of the Econometric Society* 46(1):69–85.

Mycoff, Jason D, Michael W Wagner and David C Wilson. 2009. "The empirical effects of voter-ID laws: Present or absent?" *PS: Political Science & Politics* 42(1):121–126.

Nickell, Stephen. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica: journal of the Econometric Society* 49(6):1417–1426.

Pang, Xun, Licheng Liu and Yiqing Xu. 2021. "A Bayesian Alternative to Synthetic Control for Comparative Case Studies." *Political Analysis* (forthcoming).

Robbins, Michael W, Jessica Saunders and Beau Kilmer. 2017. "A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention." *Journal of the American Statistical Association* 112(517):109–126.

Robins, J M, M A Hernán and B Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11(5):550–560.

Rosenbaum, Paul R and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41–55.

Roth, Jonathan and Pedro H C Sant'Anna. 2021. "When is Parallel Trends Sensitive to Functional Form?".

Samartsidis, Pantelis, Shaun R Seaman, Silvia Montagna, André Charlett, Matthew Hickman and Daniela De Angelis. 2020. "A Bayesian Multivariate Factor Analysis Model for Evaluating an Intervention by Using Observational Time Series Data on Multiple Outcomes." *Journal of the Royal Statistical Society, Series A* 183(4):1437–1459.

Sanford, Luke. 2021. "Geospatial Synthetic Controls for agricultural impact evaluation." Mimeo, Yale University.

Squires, Chandler, Dennis Shen, Anish Agarwal, Devavrat Shah and Caroline Uhler. 2020. "Causal Imputation via Synthetic Interventions.".

Strezhnev, Anton. 2018. "Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs." Mimeo, New York University.

Sun, Liyang and Sarah Abraham. 2018. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects.".

Tyler, Matthew. 2021. "Counterfactual Forecasting: Causal Inference without Simultaneous Controls." Mimeo, Stanford University.

Wang, Ye. 2021. "Causal Inference under Temporal and Spatial Interference." Mimeo, New York University.

Wooldridge, Jeffrey M. 2001. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Wooldridge, Jeffrey M. 2021. "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators." Mimeo.

Xu, Yiqing. 2017. "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models." *Political Analysis* 25(1):57–76.