A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data*

Licheng Liu (MIT) Ye Wang (NYU) Yiqing Xu (Stanford)

First version: 11th July 2019

This version: 25th October 2020

Abstract

This paper introduces a unified framework of counterfactual estimation for time-series cross-sectional data, which estimates the average treatment effect on the treated by directly imputing treated counterfactuals. Its special cases include several newly developed methods, such as the fixed effects counterfactual estimator, interactive fixed effects counterfactual estimator, and matrix completion estimator. These estimators provide more reliable causal estimates than conventional two-way fixed effects models when the treatment effects are heterogeneous or unobserved time-varying confounders exist. Under this framework, we propose two sets of diagnostic tests, tests for (no) pre-trend and placebo tests, accompanied by visualization tools, to help researchers gauge the validity of the no-time-varying-confounder assumption. We illustrate these methods with two political economy examples and develop an open-source package, `fect`, in both `R` and `Stata` to facilitate implementation.

Keywords: counterfactual methods, two-way fixed effects, parallel trends, interactive fixed effects, matrix completion, equivalence test, placebo test, time-series cross-sectional data, panel data

Word Count: 9,936

*Licheng Liu, Department of Political Science, Massachusetts Institute of Technology. Email: liulch@mit.edu. Ye Wang, Wilf Family Department of Politics, New York University. Email: yw1576@nyu.edu. Yiqing Xu, Department of Political Science, Stanford University. Email: yiqingxu@stanford.edu. We thank Naoki Egami, Avi Feller, Neal Beck, Bernie Black, Dan de Kadt, Erin Hartman, Chad Hazlett, Danny Hidalgo, Apoorva Lal, Kosuke Imai, In Song Kim, Jeff Lewis, Marc Ratkovic, seminar participants at NYU, UCSD, UCLA, and MIT, as well as participants of PolMeth 2019 for helpful comments.

1. Introduction

The linear two-way fixed effects model is one of the most commonly used statistical routines in the social sciences to establish causal relationships using observational time-series cross-sectional (TSCS) data (or long panel data). Researchers rely on two-way fixed effects models because they can control for a potentially large set of unobserved unit- and time-invariant confounders. Two crucial assumptions underlie this approach. First, the linearity assumption requires that the treatment effect changes in a constant rate with the treatment for all units at all time periods. When the treatment is dichotomous, it reduces to the assumption of *constant treatment effect*. The second assumption is the *absence of time-varying confounders*. It means that no unobserved, time-changing factors correlate with both the treatment and the potential outcomes. When there are only two periods and two treatment groups, this assumption implies "parallel trends," that is, the average untreated potential outcomes of the treated and control groups follow parallel paths (e.g., Angrist and Pischke 2009). Failures of these two assumptions lead to biases in the estimates for researchers' causal quantities of interest.

In this paper, we introduce a simple framework of treatment effect estimation that relaxes both assumptions in TSCS settings with dichotomous treatments. Estimators in this framework, which we call *counterfactual estimators*, take observations under the treatment condition as missing data and directly estimate their counterfactuals. They correct biases induced by treatment effect heterogeneity, which has recently much attention in the literature (e.g., Chernozhukov et al. 2013; Goodman-Bacon 2018; Strezhnev 2018; de Chaisemartin and D'Haultfœuille 2018). Because counterfactual estimators do not use treated observations in the modeling stage, they allow the treatment effects to be arbitrarily heterogeneous. As a result, the constant treatment effect assumption is relaxed.

We discuss three examples that have recently emerged in the causal inference literature, including (1) the fixed effects counterfactual (FEct) estimator, of which difference-in-

differences (DiD) is a special case, (2) the interactive fixed effects counterfactual (IFEct) estimator (Gobillon and Magnac 2016; Xu 2017) and (3) the matrix completion (MC) estimator (Athey et al. 2018; Kidziński and Hastie 2018). They differ from each other in the underlying outcome model that predicts treated counterfactuals.

Both the IFEct and MC estimators use a latent factor approach to adjust for potential time-varying confounders. Mathematically, these estimator are designed to construct a lower-rank approximation of the outcome data matrix using information of untreated observations but differ in the way of regularizing the latent factor model. Previous work and results in this paper suggest that they can provide more reliable causal estimates than conventional methods if potential time-varying confounders can be expressed as interactions between time-varying common shocks and unit-specific intercepts.

This paper aims to provide practical guidance to social scientists for analyzing TSCS data with a dichotomous treatment. An increasingly popular practice among researchers when evaluating the validity of the no-time-varying-confounder assumption is to draw a plot of the so-called "dynamic treatment effects," which are coefficients of a series of interactions between a dummy variable indicating the treatment group—units that are exposed to the treatment for at least one period during the observed time window—and a set of time dummies indicating the time period relative to the onset of the treatment using a two-way fixed effects model. If these coefficients exhibit a monotonic trend leading toward the onset of the treatment, or a "pre-trend," the assumption is deemed problematic. However, this method relies on parametric assumptions and the statistical tests derived from it are informal and often underpowered (Roth 2020). Taking advantage of the counterfactual estimation framework, we improve the practice of estimating and plotting the dynamic treatment effects, or the average treatment effects on the treated (ATT) over different periods, without assuming treatment effect homogeneity of any kind.

In addition to visual inspections, we develop two sets of diagnostic tests to gauge the validity of the no-time-varying-confounder assumption. The first one tests the presence of

a pre-trend, or whether the error term is uncorrelated with the timing of the treatment in the pretreatment periods. We use both a difference-in-means (DIM) approach, which tests against the null of no difference, and an equivalence approach, which flips the null and tests against a pre-specified difference. Consistent with the literature on equivalence tests (Hartman and Hidalgo 2018; Hartman 2020), we show that the equivalence approach has advantages over the DIM approach when limited power is a concern.

However, for more complex estimators such as the IFEct and MC, the tests for (no) pre-trend may suffer from overfitting because it is based on in-sample residual averages. Therefore, we complement these tests with a placebo test. For each unit that receives the treatment, we hide a few periods of observations right before the onset of the treatment and use the same counterfactual estimators to estimate the "treatment effects" for those periods, which are presumably zero under the identifying assumptions. Here as well, if the estimated average treatment effects in the placebo periods are statistically different from zero (with a conventional DIM approach), researchers should be concerned that the assumptions are likely invalid. Rejecting the null that the magnitude of average treatment effects in the placebo periods is beyond a given threshold (using an equivalence approach) provides another piece of evidence to support the no-time-varying-confounder assumption. The placebo tests have the merits of being intuitive and robust to model misspecification, but it does not make use of information in all pretreatment periods and can be underpowered. Using both the sets of tests and applying the visualization tools will give researchers a better idea whether the no-time-varying-confounder assumption likely holds.

This paper makes two main contributions to the literature. First, it introduces a simple framework for treatment effect estimation that covers a variety of novel estimators. This new estimation approach fixes the weighting issue that leads to biases when treatment effects are heterogeneous and are suitable for most TSCS datasets with dichotomous treatments if researchers deem the identifying assumptions reasonable. Moreover, under this framework, the IFEct and MC estimators can deal with potential time-varying confounders that can

be approximated by a lower-rank matrix of the error matrix. Our second contribution is to develop a set of visualization and diagnostic tools to assist researchers in choosing the most suitable estimator for their applications. Even if researchers in the end opt for a simple model such as the FEct, these tools make it much more convenient for researcher to evaluate the validity of no-time-varying confounders transparently.

Our approach builds on an emerging literature on causal inference with TSCS data and has advantages over existing methods under various circumstances. Compared with conventional two-way fixed effects models, the counterfactual estimators relax the constant treatment effect assumption with only minor efficiency loss. Compared with existing factor-augmented methods (e.g., Gobillon and Magnac 2016; Xu 2017), which are also counterfactual estimators, our framework can accommodate more complex TSCS designs such as treatment reversal. Compared with TSCS methods based on matching and reweighting (e.g., Abadie 2005; Imai and Kim 2018; Imai, Kim and Wang 2018; Hazlett and Xu 2018; Strezhnev 2018), our model-based approach can accommodate more complex data structure and incorporate covariates and time trends more conveniently.

These advantages come with costs. First, the strict exogeneity assumption the counterfactual estimators rely on implies that (1) past outcomes do not directly affect current treatment (*no feedback*) and (2) past treatments do not directly affect current outcome (*no carryover effect*), in addition to the assumption of no time-varying confounders (Imai and Kim 2018). The alternative approach is to assume sequential ignorability and to condition on covariates and past outcomes. Second, the model-based approach, although more flexible and often more efficient than matching or reweighting methods, is more likely to suffer from biases due to model dependency (Ho et al. 2007). Just as with fixed effects models, researchers must pay these costs if they believe unit-level time-invariant heterogeneities are important confounding factors.

The rest of the paper is organized as follows. Section 2 introduces the framework and estimation strategy, as well as three novel estimators as examples. We then introduce the

diagnostic tools, including the dynamic treatment effects plot and two sets of tests, in Section 3. In Section 4, we apply these methods to two empirical examples in political economy. The last section summaries the diagnostic tests and provides practical recommendations to researchers. Due to space limitations, we present results from Monte Carlo exercises in Supplementary Information (SI).

2. Counterfactual Estimators

2.1. A Unified Framework

Setup. Our approach can accommodate both balanced and imbalanced panels. For notational convenience, we consider a balanced panel with N units and T periods and denote Y_{it} the outcome of unit i in period t, D_{it} the treatment status, $\mathbf{X}_{it}$ a vector of the covariates, $\mathbf{U}_{it}$ the unobservable attributes, and ε_{it} the idiosyncratic error term. We assume the following class of outcome models:

Assumption 1 *(Functional form)* $Y_{it} = \delta_{it} D_{it} + f(\mathbf{X}_{it}) + h(\mathbf{U}_{it}) + \varepsilon_{it}$,

in which $f(\cdot)$ and $h(\cdot)$ are functions with known forms and δ_{it} is the treatment effect of D_{it} on unit i in period t. Under the potential outcomes framework, $Y_{it}(0) = f(\mathbf{X}_{it}) + h(\mathbf{U}_{it}) + \varepsilon_{it}$ and $Y_{it}(1) = \delta_{it} + f(\mathbf{X}_{it}) + h(\mathbf{U}_{it}) + \varepsilon_{it}$. Note that this specification assumes additive separability of the four right hand side terms. As a result, this class of models is scale-dependent (Athey and Imbens 2006), i.e., transforming the outcome from levels to logarithms may render the identification assumptions discussed below invalid. Assumption 1 also rules out treatment effect spillover both across units and over time.[1]

Estimands. The primary causal quantity of interest is the average treatment effect on the treated (ATT), whose treatment status has changed during the observed time window, i.e.,

$$ATT = \mathbb{E}[\delta_{it} | D_{it} = 1, \forall i \in \mathcal{T}, \forall t], \quad \mathcal{T} \coloneqq \{i \mid \exists t, t' \text{ s.t. } D_{it} = 0, D_{it'} = 1\}.$$[2]

[1] An exception is a staggered adoption design, with which the treatment's carryover effect on a unit is allowed.
[2] This quantity of interest is also referred to as the average treatment effect on the changed (ATC) in the

For units that have never been exposed to the treatment condition, it is difficult to compute their treated potential outcomes without strong structural assumptions; similarly, it is difficult to estimate causal effects on units that are always treated. In empirical work, researchers may be also interested in following two estimands. The first is the average treatment effect on the treated at sth ($s > 0$) periods since the treatment's onset:

$$ATT_s = \mathbb{E}[\delta_{it} | D_{i,t-s} = 0, \underbrace{D_{i,t-s+1} = D_{i,t-s+2} = \cdots = D_{it} = 1}_{s \text{ periods}}, \forall i \in \mathcal{T}], \quad s > 0.$$

For the purpose of the diagnostic tests we will introduce later, we define $ATT_s = 0$ for any $s \leq 0$. The second is the cohort average treatment effects

$$CATT_c = \mathbb{E}[\delta_{it} | D_{it} = 1, \forall i \in \mathcal{T}_c, \forall t],$$

in which a cohort is defined by $\mathcal{T}_c \coloneqq \{i \mid D_{t,t<c} = 0, D_{it',t' \geq c} = 1\}$, units that receive the treatment from period c and stay treated for the rest of the observed time window (Strezhnev 2018). This estimand is meaningful when treatment reversal never occurs. In addition to Assumption 1, we make the following assumptions necessary for causal identification:[3]

Assumption 2 (*Low-dimensional decomposition*) There exists a low-dimensional decomposition of $h(\mathbf{U}_{it})$: $h(\mathbf{U}_{it}) = L_{it}$, and $rank(\mathbf{L}_{N \times T}) \ll \min\{N, T\}$. For example, $\mathbf{L} = \mathbf{\Lambda F}$, in which $\mathbf{\Lambda}$ is a $(N \times r)$ matrix of factor loadings and $\mathbf{F}$ is a $(r \times T)$ matrix of factors and $r \ll \min\{N, T\}$.

To give a concrete example, if $\mathbf{U}_{it} = f_t \cdot \lambda_i$ is one dimensional, we can understand it as the impact of a common time trend f_t having a heterogeneous impact on each unit, whose heterogeneity is captured by λ_i. Moreover, when f_t is constant, $\mathbf{U}_{it}$ reduces to a set of unit fixed effects; when λ_i is constant, it reduces to time fixed effects. Because treatment assignment is dependent on observed untreated outcomes, we are operating under a special case of missing not at random (MNAR, Rubin 1976). Assumption 2 allows us to estimate unobserved $\mathbf{U}_{it}$ using untreated data to break this dependency.

literature (Imai and Kim 2018). To avoid confusion with the acronym's more commonly used meaning, the average treatment effect on the controls, we use ATT instead.

[3]We require additional technical assumptions, such as the assumptions on the error term and regularity conditions for each specific model. We discuss the details of these assumptions in SI.

Assumption 3 (*Strict exogeneity*) $e_{it} \perp \{D_{js}, \mathbf{X}_{js}, \mathbf{U}_{js}\}$ for any $i, j \in \{1, 2, \ldots, N\}$ and $s, t \in \{1, 2, \ldots, T\}$.

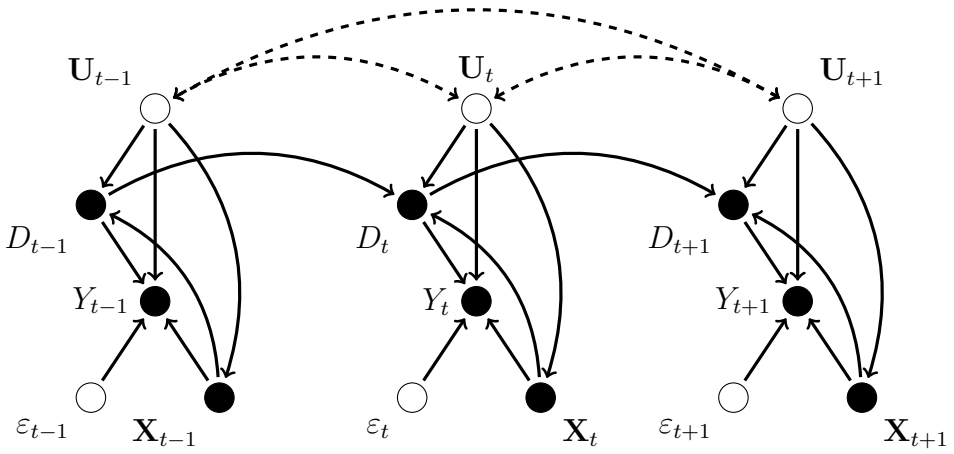
Assumption 3 implies a variant of the parallel trends assumption, i.e.,

$$\mathbb{E}[Y(0)_{it}|\mathbf{X}_{it}, \mathbf{U}_{it}] - \mathbb{E}[Y(0)_{is}|\mathbf{X}_{is}, \mathbf{U}_{is}] = \mathbb{E}[Y(0)_{jt}|\mathbf{X}_{jt}, \mathbf{U}_{jt}] - \mathbb{E}[Y(0)_{js}|\mathbf{X}_{js}, \mathbf{U}_{js}],$$

$i \in \mathcal{T}, j \notin \mathcal{T}, \forall t, s$, which states that conditional on observed covariates and unobserved attributes, the average changes in untreated potential outcome from period s to period t of the treatment group is the same as that of the control group. It is more general than the conventional parallel trends assumption in that we may condition on potentially time-varying but decomposable attributes $\mathbf{U}$.

In Figure 1, we illustrate what Assumptions 1-3 entail using a directed acyclic graph (DAG). It shows that the strict exogeneity assumption rules out both the carryover effect and feedback and that the treatment effects are separable from the influences of $\mathbf{U}_{it}$ and $\mathbf{X}_{it}$ on Y_{it}.[4] It also shows that the above setup nests many existing models for TSCS data analyses,

FIGURE 1. A DAG ILLUSTRATION



Note: Unit indices are dropped for simplicity.

including two-way FE and IFE models, both of which are under the restriction of constant treatment effect, i.e, $\delta_{it} = \delta$. Since previous research has thoroughly investigated these models, in this paper we focus on the cases in which the treatment effects are heterogeneous.

[4]Note that there are no arrows between Y_{t-1} and Y_t or between Y_t and Y_{t+1}, because such relationships violate the strict exogeneity assumption (Assumption 3).

Estimation strategy. We define the observations under control and treatment conditions as $\mathcal{O} = \{(i,t)|D_{it} = 0\}$ and $\mathcal{M} = \{(i,t)|i \in \mathcal{T}, D_{it} = 1\}$, respectively, in which $\mathcal{O}$ stands for "observed" and $\mathcal{M}$ stands for "missing." Although the outcome model researchers employ may vary, the estimation method proceeds in a similar fashion with the following steps:

Step 1. On the subset of untreated observations ($\mathcal{O}$), fit a model of the response surface Y_{it}, obtaining: $\hat{f}$ and $\hat{h}$. This step relies on the functional form assumptions on $f(\mathbf{X}_{it})$ and $h(\mathbf{U})$, as well as a lower-rank representation of $\mathbf{U}$.

Step 2. Predict the counterfactual outcome $Y_{it}(0)$ for each treated observation using $\hat{f}$, $\hat{h}(\mathbf{U})$, i.e., $\hat{Y}_{it}(0) = \hat{f}(\mathbf{X}_{it}) + \hat{h}(\mathbf{U}_{it})$, for all $(i,t) \in \mathcal{M}$.

Step 3. Estimate the individual treatment effects δ_{it} using $\hat{\delta}_{it} = Y_{it} - \hat{Y}_{it}(0)$ for each treated observation $(i,t) \in \mathcal{M}$. Note that δ_{it} is not individually identified because of idiosyncratic errors.

Step 4. Take averages of $\hat{\delta}_{it}$ to produce estimates for the quantities of interest. For example, for the ATT, $\widehat{ATT} = \frac{1}{|\mathcal{M}|}\sum_{\mathcal{M}} \hat{\delta}_{it}$; for the ATT at time period s since the treatment occurred $\widehat{ATT}_s = \frac{1}{|\mathcal{S}|}\sum_{(i,t)\in\mathcal{S}} \hat{\delta}_{it}$, in which $\mathcal{S} = \{(i,t)|D_{i,t-s} = 0, D_{i,t-s+1} = D_{i,t-s+2} = \cdots = D_{it} = 1\}$. $|\mathcal{A}|$ denotes the number of elements in set $\mathcal{A}$.

Recent research has shown that estimates from fixed effects or generalized DiD models are inconsistent for the ATT when the treatment effects are heterogeneous (Chernozhukov et al. 2013; Goodman-Bacon 2018; Strezhnev 2018; de Chaisemartin and D'Haultfœuille 2018). In fact, they converge to a weighted average of individual treatment effects on the treated, where the weights are proportional to the treatment's conditional variances. The counterfactual approach addresses this weighting problem by not using treated observations at the modeling stage and assigning uniform weights on $\hat{\delta}_{it}$ to produce an ATT estimate. Because our approach preserves the TSCS data structure, it also makes diagnostic tests easier, compared to matching and reweighting methods.

2.2. Examples

In this subsection, we review three estimators as examples of this framework. They are conceptually similar because they all construct treated counterfactuals at the individual observation level, i.e., $\hat{Y}_{it}(0)$ for all $(i,t) \in \mathcal{M}$.

a) The fixed effects counterfactual estimator. We start by introducing the *two-way fixed effects counterfactual* (FEct) estimator, in which Y_{it}, $(i,t) \in \mathcal{O}$ is fitted by a two-way fixed effects model:

$$Y_{it}(0) = \mathbf{X}'_{it}\beta + \mu + \alpha_i + \xi_t + \varepsilon_{it}, \forall i, t, D_{it} = 0.$$

We impose two linear constraints over the fixed effects $\sum_{D_{it}=0} \alpha_i = 0$, $\sum_{D_{it}=0} \xi_t = 0$ to achieve identification. With this model, we assume $f(\mathbf{X}_{it}) = \mathbf{X}'_{it}\beta + \mu$ and $h(\mathbf{U}_{it}) = \alpha_i + \xi_t$.

FEct has an advantage over the conventional fixed effects approach in that it produces unbiased and consistent causal estimates, such as ATT, ATT_s, and $CATT_c$, when the treatment effect δ_{it} is heterogeneous. When no covariates exist, we can rewrite FEct as as a weighting estimator, that is, each treated observation is matched with its predicted counterfactual $\hat{Y}_{it}(0) = \mathbf{W}\mathbf{Y}_{(i,t)\in\mathcal{O}}$, which is weighted sum of untreated observations. We provide the proofs for both results in SI.

Proposition 1 (*Unbiasedness and consistency of FEct*) Under Assumptions 1-3, as well as regularity conditions,

$$\mathbb{E}[\widehat{CATT}_c] = CATT_c;\ \mathbb{E}[\widehat{ATT}_s] = ATT_s;\ \mathbb{E}[\widehat{ATT}] = ATT;$$
$$\widehat{CATT}_c \xrightarrow{p} CATT_c,\ \widehat{ATT}_s \xrightarrow{p}, ATT_s; \text{ and } \widehat{ATT} \xrightarrow{p} ATT \text{ as } N \to \infty.$$

Proposition 2 (*FEct as a weighting estimator*): Under Assumptions 1-3 and when there are no covariates, we have:

$$\widehat{ATT}_t = \frac{1}{\sum_i D_{it}} \sum_{i, D_{it}=1} [Y_{it} - \hat{Y}_{it}(0)],$$

where $\hat{Y}_{it}(0) = \mathbf{W}\mathbf{Y}_{(i,t)\in\mathcal{O}}$ is a weighted average of untreated observations.

Remark 1: Special cases. It is easy to see that in a classic DiD setup with two groups and two periods but without covariates, the FEct estimator is the DiD estimator. Moreover, FEct and the weighted fixed effects estimator (Imai and Kim 2018) are mathematically equivalent when only unit fixed effects exist. Though nonparametric methods have the advantage of minimizing the risks of severe extrapolation (Ho et al. 2007), our model-based approach makes it easier to conduct diagnostic tests for the assumption of no time-varying confounders (see next section).

Remark 2: Including lagged dependent variables (LDVs). Although Assumption 3 rules out LDVs, in many social science applications, including LDVs in fixed effects models only induces small biases when T is large (Nickell 1981) and substantially improves the precision of estimates (Beck and Katz 2011). Chen, Chernozhukov and Fernández-Val (2019) propose a simple procedure to correct the Nickell bias. FEct incorporating LDVs behaves similarly. Moreover, in the absence of unit fixed effects, the counterfactual estimation framework can be applied to identification with sequential ignorability (e.g., Ding and Li 2019), though proofs need to change accordingly under such assumptions.

b) The interactive fixed effects counterfactual (IFEct) estimator. FEct will lead to biased estimates when unobserved time-varying confounders exist. A couple of authors have proposed using factor-augmented models to relax the no-time-varying-confounder assumption when confounders can be decomposed into time-specific factors interacted with unit-specific factor loadings (Bai 2009; Gobillon and Magnac 2016; Xu 2017; Bai and Ng 2020). IFEct models the response surface of untreated observations using a factor-augmented model:

$$Y_{it}(0) = \mathbf{X}'_{it}\beta + \alpha_i + \xi_t + \lambda'_i f_t + \varepsilon_{it}, \text{ for all } (i,t) \in \mathcal{O}.$$

When the model is correctly specified, i.e., as in Assumption 1, $f(\mathbf{X}_{it}) = \mathbf{X}'_{it}\beta$ and $h(\mathbf{U}_{it}) = \alpha_i + \xi_t + \lambda'_i f_t$ (which satisfies Assumption 2), IFEct is consistent. We provide the algorithm, as well as the proof of the following proposition, in SI.

Proposition 3 (*Consistency of IFEct*) Under Assumptions 1-3, as well as some regularity conditions, $\widehat{ATT} \xrightarrow{p} ATT$ as $N, T \to \infty$.

c) The matrix completion (MC) estimator. Athey et al. (2018) introduce the matrix completion method from the computer science literature as a generalization of factor-augmented models. Similar to FEct and IFEct, it treats a causal inference problem as a task of completing a $(N \times T)$ matrix with missing entries, where missing occurs when $D_{it} = 1$. Mathematically, MC assumes that the $(N \times T)$ matrix of $[h(\mathbf{U}_{it})]_{i=1,2,...,N,t=1,2,...,T}$ can be approximated by a lower rank matrix $\mathbf{L}_{(N\times T)}$, i.e.,

$$\mathbf{Y}(\mathbf{0}) = \mathbf{X}\beta + \mathbf{L} + \boldsymbol{\varepsilon},$$

in which $\mathbf{Y}$ is a $(N \times T)$ matrix of untreated outcomes; $\mathbf{X}$ is a $(N \times T \times k)$ array of covariates; and $\boldsymbol{\varepsilon}$ represents a $(N \times T)$ matrix of idiosyncratic errors. As with IFEct, $\mathbf{L}$ can be expressed as the product of two r-dimension matrices: $\mathbf{L} = \boldsymbol{\Lambda}\mathbf{F}$. Unlike IFEct, however, the MC estimator does not explicitly estimate $\mathbf{F}$ and $\boldsymbol{\Lambda}$; instead, it seeks to directly estimate $\mathbf{L}$ by solving the following minimization problem:

$$\widehat{\mathbf{L}} = \arg\min_{\mathbf{L}} \left[\sum_{(i,t)\in\mathcal{O}} \frac{(Y_{it} - L_{it})^2}{|\mathcal{O}|} + \lambda_L \|\mathbf{L}\| \right],$$

in which $\mathcal{O} = \{(i,t)|D_{it} = 0\}$, $\|\mathbf{L}\|$ is the chosen matrix norm of $\mathbf{L}$, and λ_L is a tuning parameter. Athey et al. (2018) propose an iterative algorithm to obtain $\widehat{\mathbf{L}}$ and show that $\widehat{\mathbf{L}}$ is an asymptotically unbiased estimator for $\mathbf{L}$. We summarize the algorithm in SI.

Remark 3: The difference between IFEct and MC. The main difference between IFEct and MC is how they regularize the singular values when decomposing the residual matrix. IFEct uses a "best subset" approach that selects the biggest r singular values, in which r is a fixed number and $r < \min\{N, T\}$, while MC imposes an $L1$ penalty on all singular values with a tuning parameter λ_L (Athey et al. 2018).[5]

[5]In the machine learning literature, they are referred to as *hard impute* and *soft impute*, respectively. See SI for the mathematical details.

Whether IFEct or MC performs better depends on context. In SI, we provide Monte Carlo evidence to show that when the factors are strong and sparse, IFEct out-performs MC. In practice, researchers may choose between the two models based on how they behave under the diagnostic tests we introduce in the next section. If both pass the tests, researchers can rely on their relative predictive power over untreated outcomes (e.g., as measured by mean-squared prediction error, or MSPE). When $r = 0$ or when λ_L is bigger than the biggest singular value of the residual matrix, no factors are included in the model; as a result, IFEct or MC reduces to FEct.

The IFEct estimator is first proposed by Gobillon and Magnac (2016) in a DiD setting where the treatment take place at the same time for a subset of units. It is also closely related to the generalized synthetic control method (Xu 2017), in which factors are estimated using the control group data only. In this paper, we accommodate scenarios in which the treatment can occur at any time period for all units or arbitrarily switch on and off (treatment reversal). In other words, the generalized synthetic control method can be seen as a special case of IFEct when the treatment does not switch back.

Remark 4: Choosing the tuning parameters. In order to choose r for IFEct, we repeat Step 2 on a training set of untreated observations until $\hat{\beta}$ converges. The optimal r is then chosen based on model performance measured by MSPE using a k-fold cross-validation scheme. To preserve temporal correlations in the data, the test set consists of a number of triplets (three consecutive untreated observations of the same unit) from the treatment group. Similarly, for the MC estimator, we use k-fold cross-validation to select the λ_L that minimizes the MSPE. The test set is constructed in the same way as in IFEct.

2.3. Inferential Methods

We rely on nonparametric block bootstrap and jackknife—both clustered at the unit level—to obtain uncertainty estimates for the treatment effect estimates. In the bootstrap procedure,

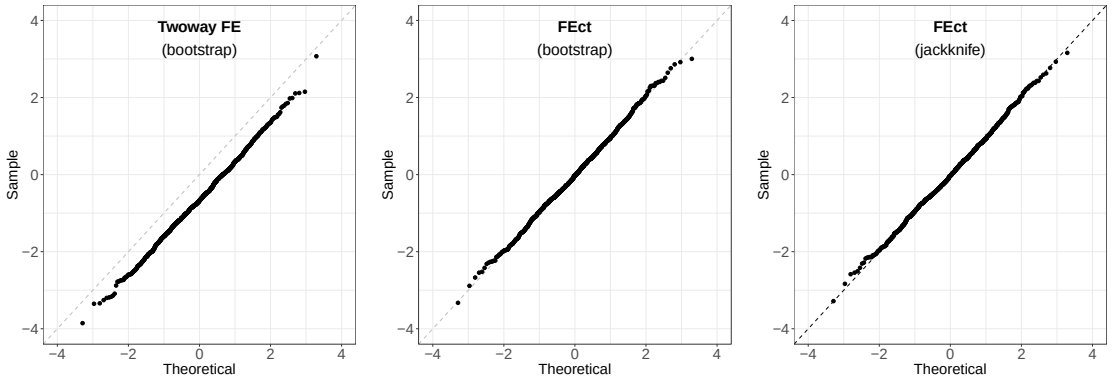
we resample, with replacement, an equal number of units from the original sample. When a unit is drawn, its entire time series of data, including the outcomes, treatment status, and covariates, are replicated. We obtain standard errors and confidence intervals of treatment effect estimates using conventional standard deviation and percentiles methods, respectively (Efron and Tibshirani 1993).

When the number of treated units is small (but bigger than one), jackknife resampling is an appealing alternative to bootstrapping (Miller 1974; Efron and Stein 1981). In each run, the procedure drops one unit (again, with its entire time series) and re-estimates the treatment effects. The variance of the ATT estimate is produced by $\widehat{\text{Var}}(\widehat{ATT}) = \frac{N-1}{N}\sum_{i=1}^{N}(\widehat{ATT}^{-i} - \overline{\widehat{ATT}})^2$, in which $\widehat{ATT}^{-i}$ is the ATT estimate from the sample in which the i'th unit is dropped and $\overline{\widehat{ATT}}$ is the average of jackknife estimates. We obtain confidence internals and p-values using a standard normal distribution. To address Bertrand, Duflo and Mullainathan (2004)'s well-known critique, both methods allow the error terms to be serially correlated but assume homoscedasticity of the errors across units. Both methods require the number of units N to be large.

We study the finite sample properties of the bootstrap and jackknife variance estimators using simulations. We simulate samples with $T = 20$ and $N = 100$ and a staggered adoption treatment assignment mechanism. We assume that no time-varying confounders exist while the treatment effects are heterogeneous, hence, FEct is consistent for the ATT while the two-way fixed effects model is not. Following Arkhangelsky et al. (2019), we plot the quantiles of the distribution for standardized errors of the ATT estimates, i.e., $(\widehat{ATT} - ATT)/\widehat{\text{Var}}(\widehat{ATT})^{1/2}$, based on 1,000 simulated samples against the quantiles of the standard normal distribution—a Quantile-Quantile plot (QQ plot)—using three combinations of estimators and inferential methods: (1) two-way fixed effects with block bootstrapped standard errors; (2) FEct with block bootstrapped standard errors; and (3) FEct with jackknife standard errors. If the ATT estimator is consistent and asymptotically normal and the chosen variance estimator precisely estimates its variance, the QQ plot should be

very close to 45-degree line.

FIGURE 2. QQ PLOTS FOR BOOTSTRAPPED AND JACKKNIFE STANDARD ERRORS



Note: The above figures show the standard Gaussian QQ plot of the standardized errors $(\widehat{ATT} - ATT)/\widehat{\text{Var}}(\widehat{ATT})^{1/2}$ for the following combination of estimators and inferential methods: (1) two-way fixed effects with block bootstrapped standard errors; (2) FEct with block bootstrapped standard errors; and (3) FEct with jackknife standard errors, each aggregated from 1,000 simulations using samples with dimension $T = 20$ and $N = 100$. The 45-degree line indicates the benchmark: consistent point estimates with perfectly calibrated Gaussian standard errors.

Figure 2 presents the results. First, because the two-way fixed effectss estimator is inconsistent, the QQ plot does not pass point $(0, 0)$ in Figure 2(a). Second, when we apply the FEct estimator, both bootstrap and jackknife procedures precisely estimate the variance of an ATT estimate: both QQ plots are almost exactly on the 45-degree lines. Our finding suggests that the literature's recommendation to use block bootstrap or jackknife for variance estimation for panel models (Bertrand, Duflo and Mullainathan 2004; Cameron and Miller 2015) can be extended to counterfactual estimators, such as FEct, IFEct, and MC.[6]

2.4. Limitations of the Model-based Approach

Throughout the paper, we take a model-based perspective for causal inference, which has a long tradition in the literature of TSCS analysis and remains popular among practitioners. This perspective requires correct specification for the outcome model, including the latent factor structure; as a result, all variables not included in the model are conditionally uncorrelated with the treatment and considered as random noises in the sampling process.

[6]Figure A3 in SI shows the performance of both variance estimators with even smaller sample sizes ($N = 30$ or $N = 50$). The results are very similar. Results are similar, too, with the IFEct and MC estimators.

In other words, the main source of causal identification is the outcome model, and model misspecification will lead to biases in the causal estimates.

Recently, researchers have made efforts to alleviate this concern by proposing doubly robust estimators that combine an outcome model with a reweighting method to take into account the treatment assignment mechanism (Ben-Michael, Feller and Rothstein 2019; Arkhangelsky et al. 2019). This paper does not incorporate these innovations because doing so would limit our the applicability of our methods a great deal (for example, they do not allow the treatment to switch on and off) and render the diagnostic tests infeasible. With the flexible models of our current approach, we attempt to achieve a balance among applicability, robustness of estimation results, and strength of causal interpretations.[7]

3. Diagnostics

In this section, we introduce a set of diagnostic tools to assist researchers probing the validity of Assumption 3. We first introduce a plot for dynamic treatment effects based on counterfactual estimators. We then propose two sets of statistical tests for the implications of the no-time-varying-confounder assumption.

3.1. A Plot for Dynamic Treatment Effects

In applied research with TSCS data, it is common for researchers to plot the so-called "dynamic treatment effects," which are coefficients of the interaction terms between the treatment indicator and a set of dummy variables indicating numbers of periods relative to the onset of the treatment—for example, $s = -4, -3, \cdots, 0, 1, \cdots, 5$ with $s = 1$ representing the first period a unit receives the treatment—while controlling for unit and time fixed effects. Researcher often gauge the plausibility of the no-time-varying-confounder assumption by

[7]For example, the canonical two-group two-period DiD model under the parallel trends assumption has a strong causal interpretation, but it is not widely applicable; the conventional two-way fixed effects model has high applicability, but only obtains a causal interpretation under very strong assumptions. We thank Naoki Egami for pointing this out.

eyeballing whether the coefficients in the pretreatment periods (when $s \leq 0$) exhibit an upward or a downward trend—often known as a "pre-trend"—or are statistically significant from zero.[8]

We improve the dynamic treatment effect plot by taking advantage of the counterfactual estimators. Instead of plotting the interaction terms, we plot the averages of the differences between Y_{it} and $\hat{Y}_{it}(0)$ for units in the treatment group ($i \in \mathcal{T}$), re-indexed based on the time relative to the onset of the treatment. Specifically, we define $\hat{\delta}_{it} = Y_{it} - \hat{Y}_{it}(0)$, for all $t, i \in \mathcal{T}$. When Assumption 3 is correct, it is easy to see that average pretreatment residuals will converge to zero, i.e., $\widehat{ATT}_s \xrightarrow{p} 0$ for all $s \leq 0$.[9] Therefore, we should expect pretreatment residual averages to be bouncing around zero, i.e., no strong pre-trend.

This method has two primary advantages over the traditional approach. First, it relaxes the constant treatment effect assumption. Even though the conventional dynamic treatment effect plot allows the treatment effects to be different across time, it assumes a constant effect for all treated units in a given time period (relative to the start of the treatment). Second, because a unit's untreated average has already been subtracted from $\hat{\delta}_{it}$, it is no longer necessary for researchers to choose a base category; to put it differently, the base category is set at a unit's untreated average after time effects are partialled out.

We illustrate the dynamic treatment effects plot using a simulated panel dataset of 200 units and 35 time periods based on the following DGP with two latent factors, f_{1t} and f_{2t}:

$$Y_{it} = \delta_{it} D_{it} + 5 + 1 \cdot X_{it,1} + 3 \cdot X_{it,2} + \lambda_{i1} \cdot f_{1t} + \lambda_{i2} \cdot f_{2t} + \alpha_i + \xi_t + \varepsilon_{it},$$

where the heterogeneous treatment effects is governed by $\delta_{it,t>T_{0i}} = 0.2(t - T_{0i}) + e_{it}$, in which e_{it} is i.i.d. $N(0,1)$. This means the expected value of the treatment effect gradually increases as a unit takes up the treatment. f_{1t} is an AR(1) process plus a deterministic upward trend and f_{2t} is an i.i.d. $N(0,1)$ white noise. Treatment assignment follows staggered adoption:

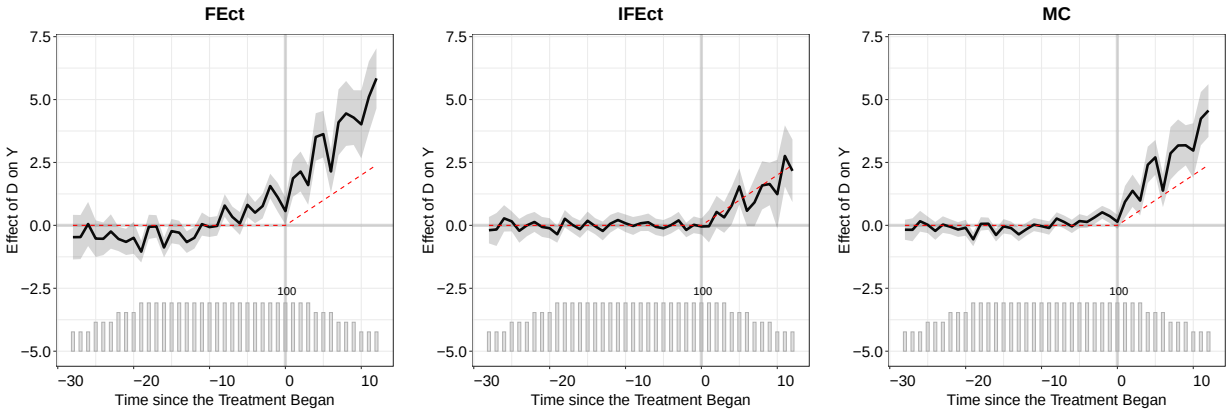
[8] The magnitudes of the coefficients and corresponding p-values often depend on the baseline category researchers choose, which varies from case to case.

[9] With some abuse of the terminology, we call the residual averages $\widehat{ATT}_s$ when $s \leq 0$.

once a unit adopts the treatment, it remains treated in the rest of the time periods (Athey and Imbens 2018). Moreover, units that have larger factor loadings (λ_{i1} and λ_{i2}) and unit fixed effects (α_i) are more likely receive the treatment early on. The selection on the factor loadings and unit fixed effects will lead to biases in the causal estimates if the latent factors and unit fixed effects are not accounted for in the estimation. This DGP satisfies Assumptions 1-3.

Figure 3 shows the estimated dynamic treatment effects with 95% confidence intervals based on block bootstraps of 1,000 times using the aforementioned counterfactual estimators. They are benchmarked against the true ATTs, which we depict with red dashed lines. From

FIGURE 3. DYNAMIC TREATMENT EFFECT FOR THE SIMULATED EXAMPLE



Note: The above figures show the dynamic treatment effects estimates from the simulated data using three different estimators: FEct, IFEct, and MC. The bar plot at the bottom of each figure illustrates the number of treated units at the given time period relative to the onset of the treatment (the number decreases as time goes by because there are fewer and fewer units that are treated for a sustained period of time). The red dashed lines indicate the true ATT.

the left panel of Figure 3, we see that using the FEct estimator, (1) a strong pre-trend leads towards the onset of the treatment and multiple "ATT" estimates (residual averages) in the pretreatment periods are significantly different from zero and (2) there are sizable positive biases in the ATT estimates in the posttreatment periods. We see a similar pattern in the posttreatment periods in the right panel where the MC estimator is applied, though with smaller biases. However, when using the IFEct estimator, the ATT estimates in both pretreatment and posttreatment periods are very close to the truth. This is expected because the DGP is generated by an IFE model with two latent factors and our cross-validation

scheme picks the correct number of factors. To help researchers gauge the effective sample size, we plot the number of treated units at a given time period beneath the corresponding ATT estimate.

In short, the plot for dynamic treatment effects displays the temporal heterogeneity of treatment effects in an intuitive way. It is also a powerful visual tool for researchers to evaluate how plausible the no-time-varying-confounder assumption is. Next, we introduce two sets of statistical procedures that formally test the implications of this assumption.

3.2. Statistical Tests

To test the presence of potential time-varying confounders, we propose two sets of statistical tests that complement each other. The first one directly tests the presence of a pre-trend, often seen as a symptom of the failure of the no-time-varying-confounder assumption. The second one is a placebo test.

Tests for no pre-trend. A natural approach is to jointly test a set of null hypotheses that the average of residuals for any pretreatment period is zero, i.e., $ATT_s = 0$ for all $s \leq 0$. Define $\boldsymbol{\delta}_{(-m:0)}$ as $(ATT_{-m}, ATT_{-(m-1)}, \ldots, ATT_0)'$, the vector of ATT between the first pretreatment period $-m$ and 0. We can construct the following variant of an F test statistic:

$$F = \frac{N_{tr}(N_{tr}-m-1)}{(N_{tr}-1)(m+1)} \boldsymbol{\delta}'_{(-m:0)} \Sigma_{\boldsymbol{\delta}_{(-m:0)}} \boldsymbol{\delta}_{(-m:0)}$$

where $\Sigma_{\boldsymbol{\delta}_{(-m:0)}}$ is the covariance matrix of $\boldsymbol{\delta}_{(-m:0)}$ and $N_{tr} = |\mathcal{T}|$, the number of treated units. Under the joint null hypothesis, the statistic converges to an F-distribution $F(m+1, N_{tr} - m - 1)$ as N_{tr} grows (Wellek 2010). We reject the null hypothesis if the statistic's value is larger than the F-distribution's 95th percentile (given the test's size $\alpha = 0.05$). However, because a test for no pre-trend is a test for equivalence, as Hartman and Hidalgo (2018) point out, this approach may suffer from limited power; that is, when the number of observations is small, failing to reject the null of joint zeros does not mean equivalence holds. Moreover, when the sample size is large, a small confounder (or a few outliers) that only contributes

to a neglectable amount of bias in the causal estimates will almost always cause rejection of the null hypothesis of joint zero means.

To address these concerns, we introduce a variant of the equivalence test proposed in experimental and regression discontinuity settings (Hartman and Hidalgo 2018; Hartman 2020).[10] The null hypothesis is reversed:

$$ATT_s < -\theta_2 \text{ or } ATT_s > \theta_1, \quad \forall s \leq 0,$$

in which $-\theta_2 < 0 < \theta_1$ are pre-specified parameters, or equivalence thresholds. Rejection of the null hypothesis implies the opposite holds with a high probability, i.e., $-\theta_2 \leq ATT_s \leq \theta_1$ for any $s \leq 0$. In other words, if we collect sufficient data and show that the pretreatment residual averages fall within a pre-specified narrow range, we obtain a piece of evidence to support the validity of the no-time-varying-confounder assumption. $[-\theta_2, \theta_1]$ is therefore called the *equivalence range*. We use the two one-sided test (TOST) to check the equivalence of ATT_s to zero for each $s < 0$. The null is considered rejected (hence, equivalence holds) only when the tests for all pretreatment periods generate significant results.[11] Following Hartman and Hidalgo (2018), we set $\theta_1 = \theta_2 = 0.36\sigma_\varepsilon$ based on simulation results. σ_ε is the standard deviation of residualized untreated outcome.[12] Given this choice, each of the TOST rejects the null of inequivalence (hence, equivalence holds) when the bootstrapped one-sided confidence interval of ATT_s falls within $[0.36\hat{\sigma}_\varepsilon, 0.36\hat{\sigma}_\varepsilon]$, the equivalence range. In addition, we also calculate the *minimum range*, the smallest symmetric bound within which we can reject the null of inequivalence using our sample. In other words, the minimal range is determined by the largest absolute value of the range of the 90% confidence intervals of $\widehat{ATT}_{s,s\leq 0}$ in the pretreatment periods if we control the size $\alpha = 0.05$. A rule of thumb is

[10] Egami and Yamauchi (2020) propose a similar test in a multi-period DiD setup.

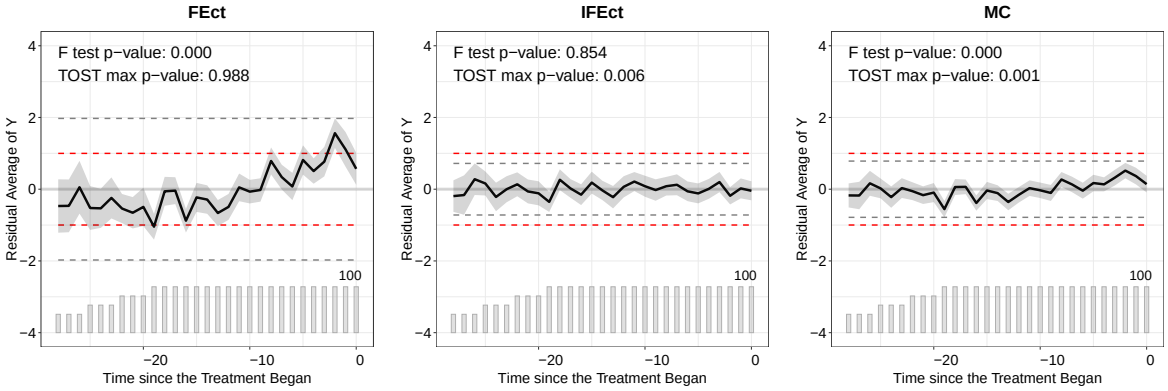
[11] This is clearly a conservative standard, as we are simultaneously testing multiple hypotheses; as a result, the Type-I error will be smaller than the test size (e.g., 0.05). In SI, we present an alternative approach: an equivalence F test, which addresses the multiple testing issue. However, its main drawback is that researchers cannot easily link the equivalence range with a substantive effect size. Because the goal of an equivalence test is control the Type-I error, multiple testing, which makes the test more conservative, is not a major concern. See Hartman (2020) (footnote 11) for a discussion.

[12] Specifically, we run a two-way fixed effects model with time-varying covariates using untreated data only and calculate the standard deviation of the residuals.

that when the minimum range is within the equivalence range, the test is considered passed. In SI, we compare the performance of the F test and the equivalence test using simulations.

Figure 4 demonstrates the results of the equivalence test based on FEct, IFEct, and MC using the simulated dataset. With FEct, the trend leading towards the onset of the treatment goes beyond the equivalence range and results in a wide minimum range. Therefore, we cannot reject the null that the pretreatment residual averages are beyond a narrow range—in other words, we cannot say that equivalence holds with high confidence. However, both IFEct and MC pass the test. The 90% confidence intervals of pretreatment residual averages are within the equivalence ranges and the minimum ranges are narrower than the equivalence range. Note that the F test p-value for MC is 0.000, which points to potential model misspecification.

FIGURE 4. TESTS FOR NO PRE-TREND: THE SIMULATED EXAMPLE



Note: The above figures show the results of the equivalence tests based on three different estimators: FEct, IFEct, and MC. Pretreatment residual averages and their 90% confidence intervals are drawn. The red dashed lines mark the equivalence range, while the gray dashed line marks the minimum range. The bar plot at the bottom of each figure illustrates the number of treated units at the given time period relative to the onset of the treatment.

The main shortcoming of the equivalence approach is that researchers need to prespecify the equivalence range. $[0.36\hat{\sigma}_\varepsilon, 0.36\hat{\sigma}_\varepsilon]$ may be too lenient when the effect size is small relative to the variance of the residualized outcome. An alternative the literature suggests is to benchmark the minimum range against a reasonable guess of the effect size based on previous studies (e.g., Wiens 2001). However, such information is often unavailable. Because the ATT estimates from a TSCS analysis can be severely biased due to failures of

the identification assumptions, unlike in experimental settings, they cannot provide valuable information for the true effect size, either. Moreover, setting the equivalence range in a post hoc fashion can lead to problematic results (Campbell and Gustafson 2018). The best practice would be for researchers to pre-register a plausible effect size and use it to set the equivalence range before analyzing data, as is a common practice in clinical trials. Another drawback of this test is that it may suffer from over-fitting: as a complex model (e.g., IFEct or MC) fits the untreated data better, the variance of the residuals becomes smaller, which makes it easier to pass the equivalence test. To guard against such risks, next, we introduce an (out-of-sample) placebo test, which employs both a DIM approach and an equivalence approach as well.

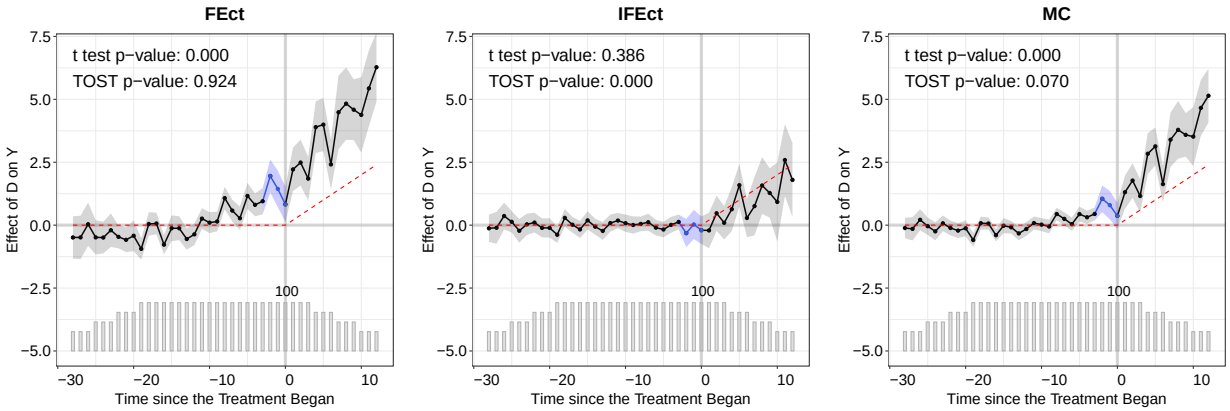
A placebo test. The basic idea for the placebo test is straightforward: we assume that the treatment starts S periods earlier than its actual onset for each unit in the treatment group and apply the same counterfactual estimator to obtain estimates of ATT_s for $s = -(S-1), \ldots, -1, 0$. We can also estimate the overall ATT for the S pretreatment periods. If the no-time-varying-confounder assumption holds, we should be able to reject the null that the magnitude of this fake "ATT" estimate is above a certain equivalence threshold. On the other hand, if this "ATT" estimate is statistically different from zero, we obtain a piece of evidence that the identifying assumption is likely to be invalid. In practice, S should not be set too large because the larger S is, the fewer pretreatment periods will remain for estimating the model. If both S and N_{tr} are too small, however, the test may be underpowered. In this and the following examples, we set $S = 3$. An important property of the proposed placebo test is that it is robust to model misspecification and immune from over-fitting because it relies on out-of-sample predictions of $Y(0)$ during the placebo periods. In this example, we use the estimated ATT to set the equivalence range. In real-world applications, we recommend researchers use an effect size they deem substantively reasonable.

Figure 5 shows the results from the placebo tests based on the three counterfactual

estimators. We see that for FEct and MC, we are able to reject the null that the placebo effect is zero but unable to reject the null that the effect is outside the equivalence range—hence, equivalence does not hold—while IFEct behaves exactly in the opposite way: the placebo effect is statistically indistinguishable from zero ($p = 0.386$), and we can reject the null hypothesis that the placebo effect is bigger than the true ATT ($p = 0.000$). Although the MC method fits the pretreatment periods well, it does not pass the placebo test using either the DIM approach ($p = 0.000$) or the equivalence approach ($p = 0.070$). This finding confirms that achieving better model fit in the pretreatment periods does not guarantee reduced biases; in fact, model misspecification and over-fitting may exacerbate the biases in causal estimates.

Because both the tests for pre-trend and the placebo test have drawbacks—the former may suffer from overfitting while the latter may be severely underpowered—we recommend researchers use both sets of tests to justify the identifying assumption.

FIGURE 5. PLACEBO TESTS FOR THE SIMULATED EXAMPLE



Note: The above figures show the results of the placebo tests based on three different estimators: FEct, IFEct, and MC. The bar plot at the bottom of each figure illustrates the number of treated units at the given time period relative to the onset of the treatment. The red dashed lines indicate the true ATT. Three pretreatment periods ($s = -2, -1, 0$) serving as the placebo are painted in blue. The p-values for the t test of the placebo effect and for the TOST are shown at the top-left corner of each figure. The equivalence range is set as [-eff, eff], in which eff is the true ATT.

4. Empirical Examples

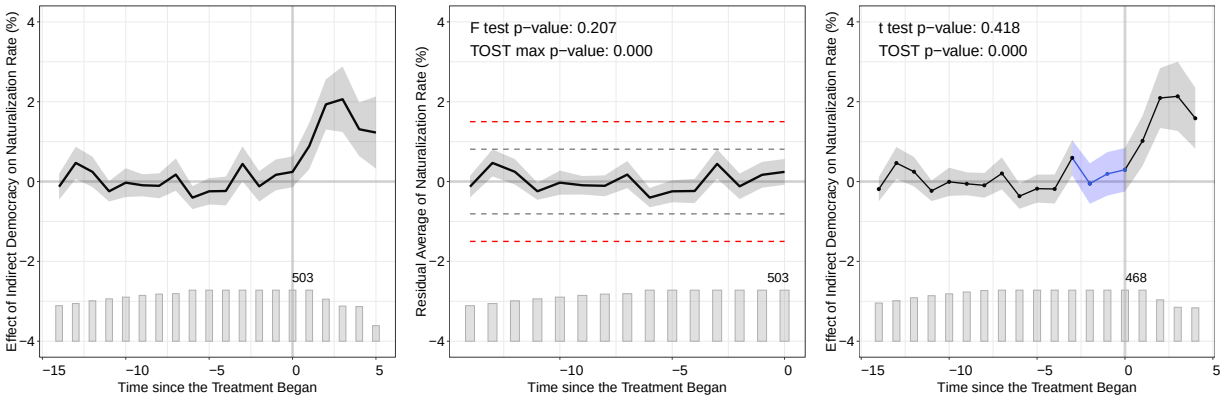
Finally, we apply the counterfactual estimators, as well as the proposed diagnostic tests, to two empirical examples in political economy. We start with FEct. If the results from FEct pass both the diagnostic tests, there is little need for more complex methods except for potential efficiency gains. If, however, the tests or visual inspection suggest the no-time-varying-confounder assumption is likely to fail, we apply IFEct and MC and conduct diagnostic tests again. In both applications, we set $S = 3$ in the placebo tests. All uncertainty estimates are obtained using clustered bootstrap at the unit level 1,000 times.

Direct democracy and naturalization rates. Hainmueller and Hangartner (2015) study whether switching from direct democracy to indirect democracy increases naturalization rates for minority immigrants in Swiss municipalities using a generalized DiD design. The outcome variable is minorities' naturalization rate in municipality i during year t. The treatment is a dummy variable indicating whether naturalization decisions are made by popular referendums. The dataset includes 1,211 Swiss municipalities over 19 years, from 1991 to 2009. The authors report that the naturalization rate increases by 1.339 percent on average after a municipality shifts the decision-making power from popular referendums to elected officials using a two-way fixed effects model. The result is replicated in Table 1 (Panel A, column 1).

We then apply the FEct estimator and obtain an estimate of 1.767 with a standard error 0.197 (column 2), even larger than the original estimate. Plots for the dynamic treatment effects and placebo test are shown in Figure 6. We find that, first, the residual averages in the pretreatment periods are almost flat and around zero and the effect gradually takes off after the treatment begins. Second, the FEct estimates pass the F test ($p = 0.207$) and the TOST ($p = 0.000$ using $\pm 0.36\hat{\sigma}_\varepsilon$ as thresholds). Third, with the placebo test, we cannot reject the null of zero placebo effect ($p = 0.418$) while we can reject the null that its magnitude is bigger than the estimated ATT ($p = 0.000$). In this case, we use the estimated

ATT to set the equivalence range because a visual inspection strongly points to a valid DiD design.[13] We also apply both IFEct and MC estimators to this example. It turns out that

FIGURE 6. THE EFFECT OF INDIRECT DEMOCRACY ON NATURALIZATION



(a) Dynamic Treatment Effects (b) Testing No Pre-Trend (c) Placebo Test

Note: The above figures show the results from applying `FEct` to data from Hainmueller and Hangartner (2015), who investigate the effect of decisions made by municipal councilors (vs. popular referendums) on naturalization rate of immigrant minorities in Swiss municipalities. The left figure shows the estimated dynamic treatment effects using FEct. The middle figure shows the results of an equivalence test for no pre-trend, in which the red and gray dashed lines mark the equivalence range and the minimum range, respectively. The right figure shows the results from a placebo test using the "treatment" in three pretreatment periods as a placebo. The bar plot at the bottom of each figure illustrates the number of treated units at a given time period relative to the onset of the treatment.

the cross-validation schemes find zero factors, in the case of IFEct, and a tuning parameter bigger than the first singular value of the residual matrix, in the case of MC, both of which imply maximum regularization (no factors). Hence, both methods reduce to FEct and give the exact same estimates as FEct.

In short, results from FEct are substantively the same as those from conventional two-way fixed effects models. However, counterfactual estimators like FEct allow us to check the validity of the no-time-varying-confounder assumption in a convenient and transparent fashion.

Official visits and firms' access to loans. The second example is based on an empirical investigation by one of the authors.[14] The research question is, in the aftermath of the 2008

[13] Using $\pm 0.36\hat{\sigma}_\varepsilon$ as thresholds gives even more favorable results.

[14] The project is suspended because of the apparent failure of the parallel trends assumption. The authors thank two former collaborators, Yunxia Bai and Ninghua Zhong, for generously sharing the data.

Table 1. Results from The Empirical Examples

Panel A: Hainmueller and Hangartner (2015)

Outcome: Naturalization Rate (%)	Two-way (1)	FEct (2)	IFEct (FEct) (3)	MC (FEct) (4)
Councilor (vs. Referendums)				
Coefficient	1.339	1.767	1.767	1.767
Standard Error	(0.161)	(0.197)	(0.197)	(0.197)
95% Confidence Interval	[1.023, 1.655]	[1.381, 2.163]	[1.381, 2.163]	[1.381, 2.163]
Unit & Period FEs	Yes	Yes	Yes	Yes
r (IFEct) / λ_L (MC)	N/A	N/A	0	$> \sigma_1$
Pre-trend F test p-value	N/A	0.207	0.207	0.207
Pre-trend TOST maximum p-value	N/A	0.000	0.000	0.000
Placebo t test p-value	N/A	0.418	0.418	0.418
Placebo TOST p-value	N/A	0.000	0.000	0.000

Panel B: Officials' Visits and Access to Loans

Outcome: Total Outstanding Loans (log)	Two-way (1)	FEct (2)	IFEct (3)	MC (4)
Visited by Officials				
Coefficient	0.526	0.557	0.290	0.422
Standard Error	(0.137)	(0.141)	(0.136)	(0.125)
95% Confidence Interval	[0.262, 0.793]	[0.333, 0.852]	[0.076, 0.556]	[0.211, 0.654]
Unit & Period FEs	Yes	Yes	Yes	Yes
r (IFEct) / λ_L (MC)	N/A	N/A	2	$0.075\sigma_1$
Pre-trend F test p-value	N/A	0.411	0.654	0.229
Pre-trend TOST maximum p-value	N/A	0.629	0.000	0.000
Placebo t test p-value	N/A	0.005	0.122	0.007
Placebo TOST p-value	N/A	0.576	0.023	0.122

Notes: The uncertainty estimates are based on cluster bootstraps at the unit level (municipality and firm, respectively) 1,000 times. σ_1 represents the first singular value of the residual matrix. If $\lambda_L > \sigma_1$, MC is reduced to IFEct.

global financial crisis, whether a high-level Chinese Communist Party official's public visit to a listed firm in China on average increased the firm's access to loans. If the answer is yes, a plausible explanation is that an official's visit to a firm signals to banks the government's backing of the firm, which makes the firm's borrowing easier.[15] We use quarterly financial data of 655 firms publicly listed in China's stock market from 2005 to 2018. The treatment is a dichotomous variable indicating that a provincial or higher level official has visited the firm since the financial crisis broke out based on firms' press releases and newspaper data. Among the 655 firms, 151 firms are treated at different time periods; the remaining 504 firms

[15]This is because most banks in China are directly or indirectly controlled by the Chinese government. The monetary policy was substantially loosened in the fourth quarter of 2008.

are never treated. The outcome of interest is a firm's log total outstanding loans.[16]

We report the treatment effect estimates in Panel B in Table 1. Columns (1) shows that a two-way fixed effects model gives an effect estimate of 0.526 (an almost 70% increase in outstanding loans) with a standard error of 0.137 under clustered bootstrap. The coefficient is statistically significant at the 1% level. When we apply the FEct estimator, the estimated effect becomes even bigger: 0.557, with a standard error of 0.141 (column (2)). However, Figure 7(a) shows that, when FEct is applied, a strong upward pre-trend leads towards the visit. One explanation for this pattern is that officials chose to visit firms that grew faster than others. We also see that the minimum range of the pre-trend is beyond the equivalence range (Figure 7(b)), which means the model fails the equivalence test for no pre-trend. Third, the estimated placebo effect is highly statistically different from 0 ($p = 0.005$) test and cannot pass the equivalence test (Figure 7(c)).[17] We then apply the IFEct and MC estimators. Figure 7 shows that, with IFEct the pre-trend mostly disappears, and we can declare equivalence for both the test for no pre-trend and the placebo test. The ATT estimate from IFEct is 0.290 with a standard error of 0.117, only 52% of that from FEct (Table 1, Panel B). This means that a large proportion of the FEct estimate is due to biases from the selection effect. The MC estimator passes the equivalence test for no pre-trend but fails the placebo tests, indicating potential model over-fitting. It is worth noting that in this application, the F tests are generally underpowered, which further demonstrates the usefulness of the equivalence approach.

In summary, our diagnostic plot and tests suggest that IFEct is a better approach for these data and the conventional two-way fixed effects approach will results in significant bias in the causal estimates.

[16] We drop firms that were visited by officials before the financial crisis. Because other firm-level variables, such as asset and market cap, are posttreatment, we do not include them in the model.

[17] In this application, we use the estimated ATT from IFEct to set the equivalence range for the placebo effect, which we deem a reasonable effect size.

5. Conclusion

This paper aims to improve practices in estimating treatment effects using observational TSCS data. Although the commonly used two-way fixed effects model requires strong assumptions to produce interpretable causal estimates, the DiD design behind these models remains a highly valuable tool for causal inference with TSCS data. Hence, we seek to make improvements based on the current practices.[18]

We focus on TSCS applications with dichotomous treatment and introduce a simple framework of estimating the average treatment effect on treated observations by directly imputing their counterfactuals. We show that, even when the treatment effects are arbitrarily heterogeneous, a counterfactual estimator is consistent for the ATT as long as the outcome model is correctly specified. We discuss three estimators under this framework, including the fixed effects counterfactual (FEct) estimator, the interactive fixed effects (IFEct) estimator, and the matrix completion (MC) estimator. IFEct and MC can account for time-varying confounders if the error matrix can be decomposed into two low-rank matrices. We recommend researchers start with the simplest estimator, FEct, and conduct the diagnostic tests. If results of these tests suggest invalid assumptions, researchers can consider applying IFEct and MC, or other more complex estimators, and conduct the diagnostic tests again.

We unify these estimators in a simple framework, allowing researchers to evaluate each model's assumptions and make an informed decision on which one to use. To do so, we improve the existing practice of estimating and plotting dynamic treatment effects and develop two sets of statistical tests—tests for (no) pre-trend and placebo tests—based on the new plot. Table 2 summarizes the usages of these tests. Using simulations (see SI), we show that the equivalence approach has advantages over conventional difference-in-means approach when limited power is a concern. We recommend researchers use these tests, as

[18] An alternative is to switch to the sequential ignorability framework, which is a promising direction (e.g., Blackwell and Glynn 2015; Imai, Kim and Wang 2018). This requires researchers to assume away unit-level time-invariant heterogeneities.

well as visual inspections, in a holistic manner to gauge the validity of the no-time-varying-confounder assumption, as we do with two empirical examples.

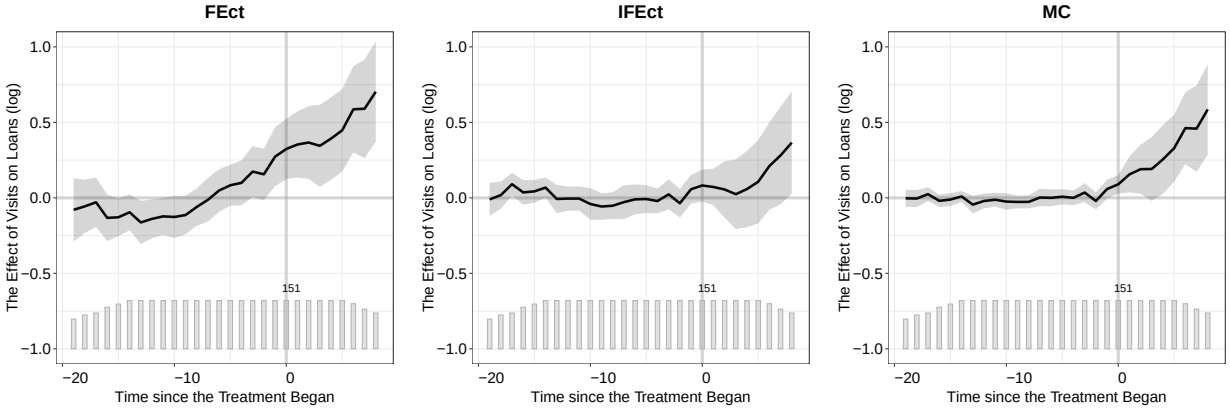
TABLE 2. DIAGNOSTIC TESTS SUMMARY

	Testing (no) pre-trend		Placebo test	
	F test	TOST	t test	TOST
Null	$ATT_s = 0, \forall s \leq 0$	$\lvert ATT_s \rvert > \theta, \exists s \leq 0$	$ATT^p = 0$	$\lvert ATT^p \rvert > \theta$
If Rejecting the Null	Invalidate Assumption	Support Assumption	Invalidate Assumption	Support Assumption
Equivalence threshold θ		$0.36\hat{\sigma}_\varepsilon$ or eff		$0.36\hat{\sigma}_\varepsilon$ or eff

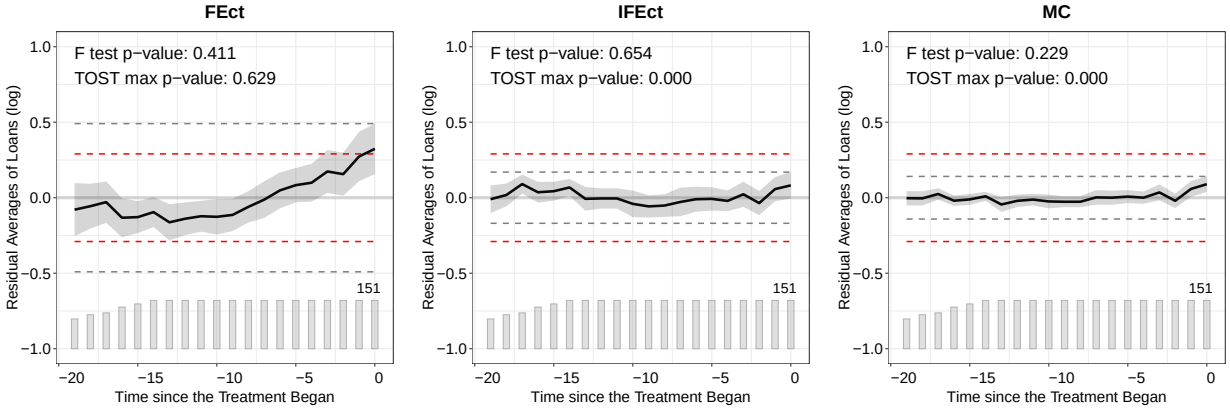
Note: Both the F and t tests are conventional difference-in-means tests, testing against the null of no difference (Hartman and Hidalgo 2018). "Assumption" refers to the no-time-varying-confounder assumption. $\hat{\sigma}_\varepsilon$ is the standard deviation of the residuals after two-way fixed effects are partialled out using untreated data only. ATT^p denotes the average placebo treatment effect on the treated. "eff" represents an effect size that researchers deem reasonable.

Our methods have several limitations. First, although we provide flexible modeling options, such as IFEct and MC, they are no panacea for all TSCS applications. Failure of the identifying assumptions and model misspecification will lead to biases in the causal estimates. Second, both IFEct and MC require both a large N and a large T; so do the diagnostic tests. Last but not least, the equivalence test approach requires users to specify an equivalence range, which may leave room for post hoc model justification. Despite these drawbacks, we believe that counterfactual estimation is a promising framework for TSCS analysis and can be extended to support a wide range of models.

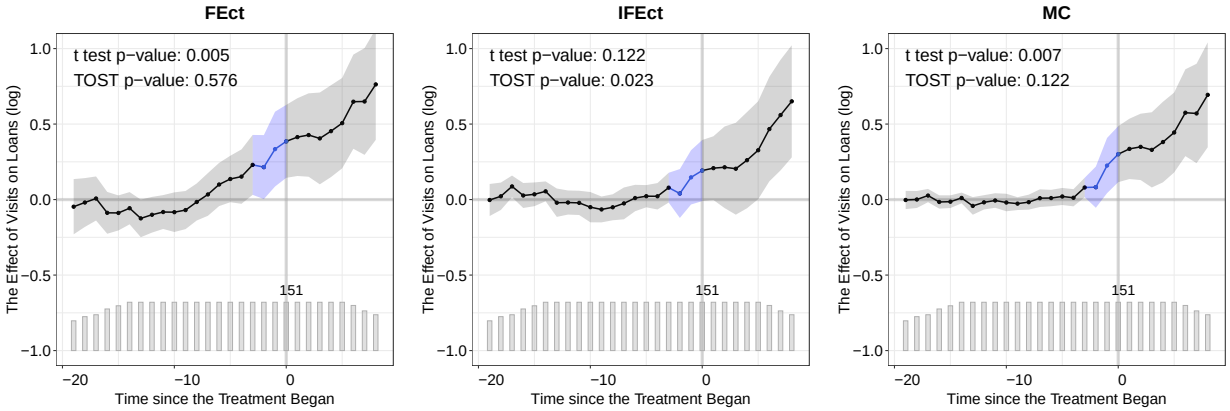
FIGURE 7. THE EFFECT OF OFFICIALS' VISITS TO FIRMS ON LOANS



(a) Dynamic Treatment Effects



(b) Testing No Pre-Trend



(c) Placebo Test

Note: The above figures show the results from applying the counterfactual estimators to the application of officials' visit to firms on firms' access to loans. The bar plot at the bottom of each figure illustrates the number of treated units at a given time period relative to the onset of the treatment. In plot (b), pretreatment residual averages and their 90% confidence intervals are drawn; the red and gray dashed lines mark the equivalence range and the minimum range, respectively.

References

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

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

Arkhangelsky, Dmitry, Susan Athey, David A Hirshberg, Guido W Imbens and Stefan Wager. 2019. "Synthetic Difference in Differences.".

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

Athey, Susan and Guido W Imbens. 2018. Design-based analysis in difference-in-differences settings with staggered adoption. Technical report National Bureau of Economic Research.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens and Khashayar Khosravi. 2018. Matrix completion methods for causal panel data models. Technical report National Bureau of Economic Research.

Bai, Jushan. 2009. "Panel Data Models with Interactive Fixed Effects." *Econometrica* 77:1229–1279.

Bai, Jushan and Serena Ng. 2020. "Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data." arXiv [econ.EM].

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

Ben-Michael, Eli, Avi Feller and Jesse Rothstein. 2019. The Augmented Synthetic Control Method. Technical report University of California, Berkeley.

Bertrand, Marianne, Ester Duflo and Sendhil Mullainathan. 2004. "How Much Should We Trust Difference-in-Differences Estimates." *The Quarterly Jounral of Economics* 119(1):249–275.

Blackwell, Matthew and Adam Glynn. 2015. "How to Make Causal Inferences with Time-Series Cross-Sectional Data." Mimeo, Harvard Uniersity.

Cameron, A. Colin and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50(2):317–372.

Campbell, Harlan and Paul Gustafson. 2018. "What to Make of Non-inferiority and Equivalence Testing with a Post-specified Margin?" Mimeo, University of British Columbia.

Chen, Shuowen, Victor Chernozhukov and Iván Fernández-Val. 2019. "Mastering Panel 'Metrics: Causal Impact of Democracy on Growth.".

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

de Chaisemartin, Clément and Xavier D'Haultfœuille. 2018. "Two-way fixed effects estimators with heterogeneous treatment effects.".

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

Efron, Brad and Rob Tibshirani. 1993. *An Introduction to the Bootstrap*. New York, NY: Chapman & Halll.

Efron, Bradley and Charles Stein. 1981. "The Jackknife Estimate of Variance." *The Annals of Statistics* 9(3):586—-596.

Egami, Naoki and Soichiro Yamauchi. 2020. "How to Improve the Difference-in-Differences Design with Multiple Pre-treatment Periods"." Columbia, Stanford University.

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

Goodman-Bacon, Andrew. 2018. "Difference-in-Differences with Variation in Treatment Timing.".

Hainmueller, Jens and Dominik Hangartner. 2015. "Does Direct Democracy Hurt Immigrant Minorities? Evidence from Naturalization Decisions in Switzerland." *American Journal of Political Science* pp. 14–38.

Hartman, Erin. 2020. "Equivalence Testing for Regression Discontinuity Designs." *Political Analysis* (forthcoming).

Hartman, Erin and F Daniel Hidalgo. 2018. "An Equivalence Approach to Balance and Placebo Tests." *American Journal of Political Science* 62(4):1000–1013.

Hazlett, Chad and Yiqing Xu. 2018. "Trajectory Balancing: A General Reweighting Approach to Causal Inference with Time-Series Cross-Sectional Data." Working Paper, UCLA and UCSD.

Ho, Daniel E, Kosuke Imai, Gary King and Elizabeth A Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3):199–236.

Imai, Kosuke and In Song Kim. 2018. "When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data." Mimeo, Massachusetts Institute of Technology.

Imai, Kosuke, In Song Kim and Erik Wang. 2018. "Matching Methods for Causal Inference with Time-Series Cross-Section Data." Working Paper, Princeton University.

Kidziński, Łukasz and Trevor Hastie. 2018. "Longitudinal data analysis using matrix completion." *arXiv preprint arXiv:1809.08771* .

Li, Kathleen. 2018. "Inference for Factor Model Based Average Treatment Effects.".

Miller, Rupert G. 1974. "The Jackknife—A Review." *Biometrika* 61(1):1–15.

Moon, Hyungsik Roger and Martin Weidner. 2015. "Dynamic Linear Panel Regression Models with Interactive Fixed Effects." *Econometric Theory* (forthcoming).

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

Roth, Jonathan. 2020. "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends." Mimeo, Harvard University.

Rubin, Donald B. 1976. "Inference and missing data." *Biometrika* 63(3):581–592.

Strezhnev, Anton. 2018. "Semiparametric weighting estimators for multi-period difference-in-differences designs." Working Paper, Harvard University.

Wansbeek, Tom and Arie Kapteyn. 1989. "Estimation of the error-components model with incomplete panels." *Journal of Econometrics* 41(3):341–361.

Wellek, Stefan. 2010. *Testing statistical hypotheses of equivalence and noninferiority*. Chapman and Hall/CRC.

Wiens, Brian L. 2001. "Choosing an Equivalence Limit for Noninferiority or Equivalence Studies." *Controlled Clinical Trials* 23:2–14.

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

A. Supplementary Information (SI)

Table of Contents

A.1. Algorithms

A.1.1. The IFEct Algorithm

The IFEct algorithm takes for the following four steps.

Step 1. Assuming in round h we have $\hat{\mu}^{(h)}$, $\hat{\alpha}_i^{(h)}$, $\hat{\xi}_t^{(h)}$, $\hat{\lambda}_i^{(h)}$, $\hat{f}_t^{(h)}$ and $\hat{\beta}^{(h)}$. Denote $\dot{Y}_{it}^{(h)} \coloneqq Y_{it} - \hat{\mu}^{(h)} - \hat{\alpha}_i^{(h)} - \hat{\xi}_t^{(h)} - \hat{\lambda}_i^{(h)\prime}\hat{f}_t^{(h)}$ for the untreated ($D_{it}=0$):

Step 2a. Update $\hat{\beta}^{(h+1)}$ using the untreated data only (we can set $\hat{\lambda}_i^{(0)} = \mathbf{0}$, $\hat{f}_t^{(0)} = \mathbf{0}$ in round 0 and run a two-way fixed effects model to initialize μ, α_i and ξ_t):

$$\hat{\beta}^{(h+1)} = \left(\sum_{D_{it}=0} \mathbf{X}_{it}\mathbf{X}_{it}'\right)^{-1} \sum_{D_{it}=0} \mathbf{X}_{it}\dot{Y}_{it}^{(h)}$$

Note that matrix $\left(\sum_{D_{it}=0} \mathbf{X}_{it}\mathbf{X}_{it}'\right)^{-1}$ is fixed and does not need to be updated every time.

Step 2b. For all i, t, define

$$W_{it}^{(h+1)} \coloneqq \begin{cases} = Y_{it} - \mathbf{X}_{it}'\hat{\beta}^{(h+1)}, \ D_{it}=0 \\ = \hat{\mu}^{(h)} + \hat{\alpha}_i^{(h)} + \hat{\xi}_t^{(h)} + \hat{\lambda}_i^{(h)\prime}\hat{f}_t^{(h)}, \ D_{it}=1 \end{cases}$$

For all untreated observations (i.e., $D_{it}=0$) , calculate $W_{it}^{(h)}$. For all treated observations (i.e., $D_{it}=1$), calculate its conditional expectation:

$$\mathbb{E}\left(W_{it}^{(h+1)} \middle| \hat{\lambda}_i^{(h)}, \hat{F}_t^{(h)}\right) = \hat{\mu}^{(h)} + \hat{\alpha}_i^{(h)} + \hat{\xi}_t^{(h)} + \hat{\lambda}_i^{(h)}\hat{F}_t^{(h)}$$

Step 2c. Denote $W_{\cdot\cdot}^{(h+1)} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}^{(h+1)}}{NT}$, $W_{i\cdot}^{(h+1)} = \frac{\sum_{t=1}^{T} W_{it}^{(h+1)}}{T}, \forall i$, $W_{\cdot t}^{(h+1)} = \frac{\sum_{i=1}^{N} W_{it}^{(h+1)}}{N}, \forall t$ and $\tilde{W}_{it}^{(h+1)} = W_{it}^{(h+1)} - W_{i\cdot}^{(h+1)} - W_{\cdot t}^{(h+1)} + W_{\cdot\cdot}^{(h+1)}$. With restrictions: $\sum_{i=1}^{N}\alpha_i = 0$, $\sum_{t=1}^{T}\xi_t = 0$, $\sum_{i=1}^{N}\lambda_i = \mathbf{0}$ and $\sum_{t=1}^{T} f_t = \mathbf{0}$.

Step 2d. Update estimates of factors and factor loadings by minimizing the least squares objective function using the complete data of $\mathbf{W}^{(h+1)} = [\tilde{W}_{it}^{(h+1)}]_{\forall i,t}$:

$$(\hat{\mathbf{F}}^{(h+1)}, \hat{\mathbf{\Lambda}}^{(h+1)}) = \underset{(\tilde{\mathbf{F}},\tilde{\mathbf{\Lambda}})}{\arg\min}\, \mathbf{tr}\left[(\mathbf{W}^{(h+1)} - \tilde{\mathbf{F}}\tilde{\mathbf{\Lambda}}')'(\mathbf{W}^{(h+1)} - \tilde{\mathbf{F}}\tilde{\mathbf{\Lambda}}')\right]$$
$$s.t. \qquad \tilde{\mathbf{F}}'\tilde{\mathbf{F}}/T = \mathbf{I}_r, \tilde{\mathbf{\Lambda}}'\tilde{\mathbf{\Lambda}} = \text{diagonal}$$

Step 2e. Update estimates of grand mean and two-way fixed effects:

$$\hat{\mu}^{(h+1)} = W_{\cdot\cdot}^{(h+1)}$$
$$\hat{\alpha}_i^{(h+1)} = W_{i\cdot}^{(h+1)} - W_{\cdot\cdot}^{(h+1)}$$
$$\hat{\xi}_t^{(h+1)} = W_{\cdot t}^{(h+1)} - W_{\cdot\cdot}^{(h+1)}$$

Step 3. Estimate treated counterfactual, obtaining:

$$\hat{Y}_{it}(0) = \mathbf{X}_{it}'\hat{\beta} + \hat{\alpha}_i + \hat{\xi}_t + \hat{\lambda}_i'\hat{f}_t, \text{ for all } i,\ t,\ D_{it}=1$$

Step 4. Obtain the ATT and ATT_s as in FEct.

A.1.2. The MC Algorithm

We summarize the algorithm for the matrix completion (MC) method below. First, define $P_{\mathcal{O}}(\mathbf{A})$ and $P_{\mathcal{O}}^{\perp}(\mathbf{A})$ for any matrix $\mathbf{A}$:

$$P_{\mathcal{O}}(\mathbf{A}) = \begin{cases} \mathbf{A}_{it}, & \text{if}(i,t) \in \mathcal{O}. \\ 0, & \text{if}(i,t) \notin \mathcal{O}. \end{cases} \quad \text{and} \quad P_{\mathcal{O}}^{\perp}(\mathbf{A}) = \begin{cases} 0, & \text{if}(i,t) \in \mathcal{O}. \\ \mathbf{A}_{it}, & \text{if}(i,t) \notin \mathcal{O}. \end{cases}$$

Conduct Singular Value Decomposition (SVD) on matrix $\mathbf{A}$ and obtain $\mathbf{A} = \mathbf{S}\mathbf{\Sigma}\mathbf{R}^{\mathbf{T}}$. The matrix shrinkage operator is defined as $\text{shrink}_{\theta}(\mathbf{A}) = \mathbf{S}\tilde{\mathbf{\Sigma}}\mathbf{R}^{\mathbf{T}}$, where $\tilde{\mathbf{\Sigma}}$ equals to $\mathbf{\Sigma}$ with the i-th singular value $\sigma_i(A)$ replaced by $\max(\sigma_i(A) - \theta, 0)$, which is called "soft impute" in the machine learning literature. The MC algorithm takes the following iterative steps:

Step 0. Given a tuning parameter θ, we start with the initial value $\mathbf{L}_0(\theta) = P_{\mathcal{O}}(\mathbf{Y})$.

Step 1. For $h = 0, 1, 2, \ldots$, we use the following formula to calculate $\mathbf{L}_{h+1}(\theta)$:

$$\mathbf{L}_{h+1}(\theta) = \text{shrink}_{\theta}\left\{P_{\mathcal{O}}(\mathbf{Y}) + P_{\mathcal{O}}^{\perp}(\mathbf{L}_h(\theta))\right\}$$

Step 2. Repeat Step 1 until the sequence $\{\mathbf{L}_h(\theta)\}_{h \geq 0}$ converges.

Step 3. Given $\hat{Y}_{it}(0) = \hat{L}_{it}^{*}$, and $\hat{\delta}_{it} = Y_{it} - \hat{Y}_{it}(0)$, compute ATT and ATT_s as before.

If we replace $\sigma_i(A)$ by $\sigma_i(A)\mathbf{1}\{\sigma_i(A) \geq \theta\}$, which is called "hard impute," the algorithm will produce estimates almost identical to the IFEct algorithm. Below is an illustration of the the difference between hard impute—which selects two factors—and soft impute, adapted from Athey et al. (2018), in which hard compute (left) selects two factors:

Hard Impute ("best subset"/IFEct)

$$\begin{pmatrix} \sigma_1 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}_{N \times T}$$

Soft Impute

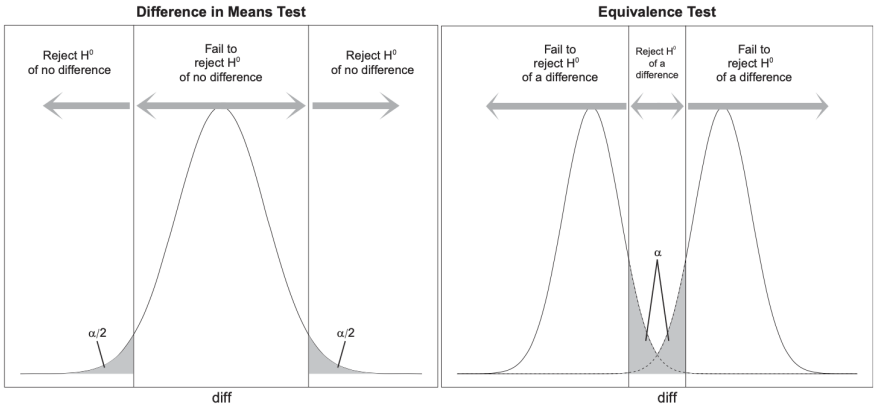
$$\begin{pmatrix} |\sigma_1 - \lambda_L|_+ & 0 & 0 & \cdots & 0 \\ 0 & |\sigma_2 - \lambda_L|_+ & 0 & \cdots & 0 \\ 0 & 0 & |\sigma_3 - \lambda_L|_+ & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ & & & & |\sigma_T - \lambda_L|_+ \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}_{N \times T}$$

Note: $|a|_+ = \max(a, 0)$

A.1.3. The Difference-in-Means Tests and Equivalence Tests

The equivalence test we introduce in the main text takes the form of two one-sided tests (TOST) for each pretreatment period s. We declare equivalence only when the test rejects the null hypothesis in all the periods of interest. Figure A1 is adapted from Hartman and Hidalgo (2018) and shows the difference between the usual t test and the TOST. The null hypothesis is rejected when the test statistic's value falls in the shaded region on both panels. But only the shaded region on the right reflects the magnitude of Type-I error that we are trying to control for.

FIGURE A1. t DISTRIBUTION UNDER THE TWO TESTS



Note: The left panel depicts the logic of tests of difference under the null hypothesis of no difference. The right panel depicts the logic of one type of equivalence test—the two one-sided t-test (TOST)—under the null hypothesis of difference.

An alternative approach is an equivalence F test, which uses the same statistic as the F test:

$$F = \frac{Ntr(Ntr - m - 1)}{(Ntr - 1)(m + 1)} \boldsymbol{\delta}'_{(-m:0)} \Sigma_{\boldsymbol{\delta}_{(-m:0)}} \boldsymbol{\delta}_{(-m:0)}$$

where $\boldsymbol{\delta}_{(-m:0)} = (ATT_{-m}, ATT_{-(m-1)}, \ldots, ATT_0)'$ and $\Sigma_{\boldsymbol{\delta}_{(-m:0)}}$ is the covariance matrix of $\boldsymbol{\delta}_{(-m:0)}$. The key difference lies in that we impose a reversed null hypothesis for the equivalence test:

$$H_0 : \quad \boldsymbol{\delta}'_{(-m:0)} \Sigma_{\boldsymbol{\delta}_{(-m:0)}} \boldsymbol{\delta}_{(-m:0)} > \kappa,$$

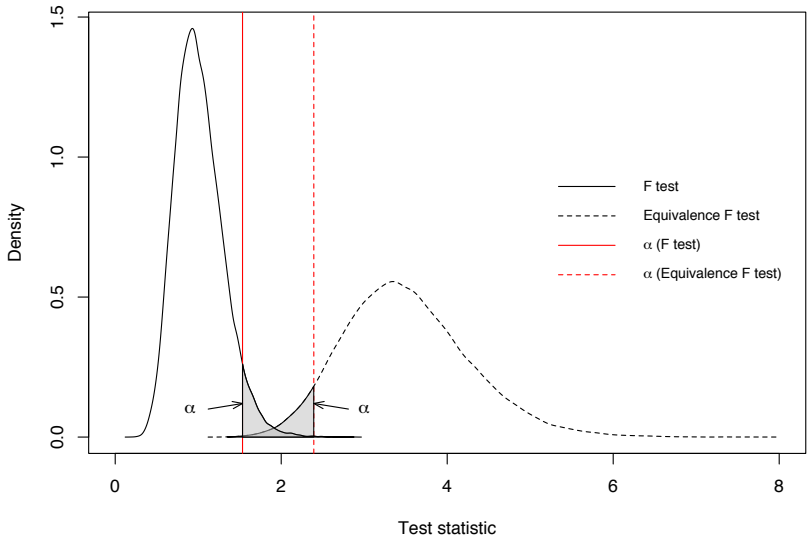
Wellek (2010) shows that under this hypothesis, the statistic converges to a non-central F-distribution $F(m + 1, N_{tr} - m - 1, N_{tr}\kappa^2)$, where $N_{tr}\kappa^2$ is the distribution's centrality parameter. The null is considered rejected (hence, equivalence holds) when the statistic's value is smaller than the 5th percentile of the distribution. When the absolute values of all the ATT_s are smaller, there will be a higher chance to reject the null . Based on the discussion in Wellek (2010) and simulation results, we recommend to set $\kappa = 0.6$.

We compare the distribution of the F test with that of the equivalence test in Figure A2 above. The solid black curve represents the distribution of the test statistic under the null of the F test. The dotted black curve represents its distribution under the null of the equivalence test. We reject the null under the former if the value of the test statistic falls on the right side of the solid red

line and reject the null under the latter if it falls on the left side of the dotted red line, i.e., the equivalence threshold.

In the above case (with a chosen equivalence threshold of 0.6), the equivalence test is more lenient than the F test: when the test statistic falls between the two red lines, we reject the null under the F test (suggesting inequivalence), but also reject the null under the equivalence test (declaring equivalence). Therefore, the equivalence test has the same advantages as the TOST does. However, because its threshold is less intuitive and harder to interpret, we choose the TOST as the primary approach to conduct the equivalence test.

FIGURE A2. F DISTRIBUTION UNDER THE TWO TESTS



Note: The above figure plots the distribution of the test statistic under the null of the F test and its distribution under the null of the equivalence test. The shaded areas represent the size of the two tests ($\alpha = 0.05$).

A.2. Proofs

A.2.1. Unbiasedness and Consistency of FEct and IFEct

Denote the number of all observations, the number of observations with $D_{it} = 1$, and the number of observations with $D_{it} = 0$ as n, n_{tr}, and n_{co}, respectively. Under FEct, our Assumptions 1, 2, and 3 lead to the following model specification:

$$\begin{aligned} Y_{it} &= \mathbf{X}_{it}'\beta + \mu + \alpha_i + \xi_t + \varepsilon_{it},\ D_{it} = 0, \\ &\sum_{D_{it}=0} \alpha_i = 0, \sum_{D_{it}=0} \xi_t = 0, \\ \varepsilon_{it} &\perp \{D_{js}, \mathbf{X}_{js}, \alpha_j, \xi_s\} \text{ for any } i, j \in \{1, 2, \ldots, N\} \text{ and } s, t \in \{1, 2, \ldots, T\}. \end{aligned} \tag{A1}$$

The data we use to estimate these parameters constitute an unbalanced panel since we are not using observations whose $D_{it} = 1$. Following Wansbeek and Kapteyn (1989), we rearrange the observations so that data on N units "are ordered in T consecutive sets", thus the index t "runs slowly" and i "runs quickly." Denote the number of untreated units in period t as N_t, then $N_t \leq N$ and $\sum_{t=1}^{T} N_t = n_{co}$, the number of untreated observations in the dataset. Similarly, denote the number of periods in which unit i is untreated as T_i. Then $T_i \leq T$ and $\sum_{i=1}^{N} T_i = n_{co}$. Let M_t be the $N_t \times N$ matrix where row i equals to the corresponding row in the unit matrix I_N if i is observed in period t. Then we can rewrite Equation (A1) in the matrix form:

$$Y = \mathbf{X}\beta + (\iota_n, \Delta)(\mu, \alpha, \xi)' + \varepsilon$$

where $\mathbf{X} = (\mathbf{x}_{11}, \mathbf{x}_{21}, \cdots, \mathbf{x}_{NT})'$ is a $n_{co} \times K$ matrix, ι_n denotes the n_{co}-dimension vector consisted of 1s, $\Delta = (\Delta_1, \Delta_2)$, $\Delta_1 = \begin{pmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \\ \vdots \\ \mathbf{M}_T \end{pmatrix}$, and $\Delta_2 = \begin{pmatrix} \mathbf{M}_1\iota_N & & & \\ & \mathbf{M}_2\iota_N & & \\ & & \ddots & \\ & & & \mathbf{M}_T\iota_N \end{pmatrix}$.

We further denote $\mathbf{D} = (D_{it})_{N\times T} = (\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_T)$. It is easy to see that $\Delta_1 * \Delta_2' = \frac{\iota_N \iota_T'}{N+T} - \mathbf{D}$.

Under IFEct, the model specification that satisfies Assumptions 1, 2, and 3 has the following form:

$$\begin{aligned} Y_{it} &= \mathbf{X}_{it}'\beta + \lambda_i' f_t + \alpha_i + \xi_t + \varepsilon_{it},\ D_{it} = 0, \\ &\sum_{D_{it}=0} \alpha_i = 0, \sum_{D_{it}=0} \xi_t = 0, \mathbf{\Lambda}'\mathbf{\Lambda} = diagnal, \mathbf{F}'\mathbf{F}/T = \mathbf{I}_r, \\ \varepsilon_{it} &\perp \{D_{js}, \mathbf{X}_{js}, \alpha_j, \xi_s, \lambda_j, f_s\} \text{ for any } i, j \in \{1, 2, \ldots, N\} \text{ and } s, t \in \{1, 2, \ldots, T\}. \end{aligned} \tag{A2}$$

in which $\mathbf{\Lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_N]'$ and $\mathbf{F} = [f_1, f_2, \ldots, f_T]'$. From now on we denote the projection matrix of matrix $\mathbf{A}$ as $P_{\mathbf{A}}$ and the corresponding residual-making matrix as $Q_{\mathbf{A}}$.

Proving the ATT estimator's consistency requires some regularity conditions. First, following Bai (2009) and Xu (2017), we assume that the error terms have weak serial dependence:

Weak serial dependence:

1. $E\left[\varepsilon_{it}\varepsilon_{is}\right]=\sigma_{i,ts}, |\sigma_{i,ts}|\leq\bar{\sigma}_i$ for all (t,s) such that $\frac{1}{N}\sum_i^N\bar{\sigma}_i<M$.
2. For every (t,s), $E\left[N^{-1/2}\sum_i^N\varepsilon_{it}\varepsilon_{is}-E\left[\varepsilon_{it}\varepsilon_{is}\right]\right]^4\leq M$.
3. $\frac{1}{NT^2}\sum_{t,s,u,v}\sum_{i,j}|cov\left[\varepsilon_{it}\varepsilon_{is},\varepsilon_{ju}\varepsilon_{jv}\right]|\leq M$ and
$\frac{1}{N^2T}\sum_{t,s}\sum_{i,j,k,l}|cov\left[\varepsilon_{it}\varepsilon_{jt},\varepsilon_{ks}\varepsilon_{ls}\right]|\leq M$.
4. $E\left[\varepsilon_{it}\varepsilon_{js}\right]=0$ for all $i\neq j$, (t,s).

These assumptions imply assumption 2 in Moon and Weidner (2015) that $\frac{||\varepsilon||}{NT}\to 0$ as N,T go to infinity. We also need some restrictions on parameters in the models:

Restriction on parameters:

1. For each t, $\frac{N_t}{N}\to p_t$ as $N\to\infty$, where p_t is a constant that varies with t.
2. All entries of the matrix $E\left[\mathbf{x_{it}}\mathbf{x_{it}}'\right]$ is bounded by M.
3. For each unit i, all the covariates have weak serial dependence: $\sum_t^{T_i}X_{it,j}\times\sum_t^{T_i}X_{it,k}\leq M$ for any (k,j).
4. Define $W(\lambda)$ as $\{\frac{1}{N}tr\left(\mathbf{x}_{k_1}'Q_\lambda\mathbf{x}_{k_2}Q_{\mathbf{F}}\right)\}_{K\times K}$ and $w(\lambda)$ as the smallest eigenvalue of $W(\lambda)$. Define $W(f)$ as $\{\frac{1}{N}tr\left(\mathbf{x}_{k_1}Q_f\mathbf{x}_{k_2}'Q_{\mathbf{\Lambda}}\right)\}_{K\times K}$ and $w(f)$ as the smallest eigenvalue of $W(f)$. Then either $\lim_{N,T\to\infty}min_\lambda\ w(\lambda)>0$, or $\lim_{N,T\to\infty}min_f\ w(f)>0$ holds.

The last restriction comes from Moon and Weidner (2015) for the consistency of the IFEct model.

Lemma 1 *Under model specification (A1) and regularity conditions, all the following limits*[A1] *exist:* $(a)\lim_{N\to\infty}\frac{\mathbf{X}'\mathbf{X}}{N}$, $(b)\lim_{N\to\infty}\frac{\mathbf{X}'\epsilon}{N}$, $(c)\lim_{N\to\infty}\frac{\mathbf{X}'\Delta_2}{N}$, $(d)\lim_{N\to\infty}\frac{\Delta_2'\Delta_2}{N}$, $(e)\lim_{N\to\infty}\frac{\Delta_1'\Delta_1}{N}$, $(f)\lim_{N\to\infty}\frac{\mathbf{X}'\Delta_1 diag\{\frac{1}{T_i}\}\mathbf{X}\Delta_1'}{N}$, *where* $diag\{\frac{1}{T_i}\}$ *is a diagonal matrix with* $\frac{1}{T_i}$ *being the ith entry on the diagonal.*

Proof: We start from proving (a). When the regularities conditions are satisfied, we can apply the weak law of large numbers:

$$\lim_{N\to\infty}\frac{\mathbf{X}'\mathbf{X}}{N}=\lim_{N\to\infty}\frac{\sum_i^N\sum_t^{T_i}\mathbf{x}_{it}\mathbf{x}_{it}'}{N}=\frac{\sum_i^N\sum_t^{T_i}E\left[\mathbf{x}_{it}\mathbf{x}_{it}'\right]}{N}=\bar{T}_iE\left[\mathbf{x}_{it}\mathbf{x}_{it}'\right]$$

which is bounded by $\bar{T}_iM$. Similarly,

$$\lim_{N\to\infty}\frac{\mathbf{X}'\varepsilon}{N}=\bar{T}_iE\left[\mathbf{x}_{i,t}\varepsilon_{\mathbf{i,t}}'\right]=\mathbf{0}_{NT\times 1}$$

For (c), we know that

$$\lim_{N\to\infty}\frac{\mathbf{X}'\Delta_2}{N}=\lim_{N\to\infty}\frac{\sum_i^N\sum_t^{T_i}\mathbf{x}_{it}\Delta_{2,it}'}{N}=\frac{\sum_i^N E\left[\sum_t^{T_i}\mathbf{x}_{it}\Delta_{2,it}'\right]}{N}=\frac{\sum_i^N E\left[\mathbf{A}_i\right]}{N}$$

where $\mathbf{A}_i$ is a $K\times T$ matrix, and the tth column of $\mathbf{A}_i$ equals to $\mathbf{0}_{K\times 1}$ when $D_{it}=1$ and equals to $\mathbf{x}_{it}$ when $D_{it}=0$. Clearly the limit exists under regularity conditions.

[A1] All the convergences here are convergence in probability.

(d) and (e) are obvious. For (f),

$$\begin{aligned}
&\lim_{N\to\infty} \frac{\mathbf{X}'\Delta_1 diag\{\frac{1}{T_i}\}\mathbf{X}\Delta_1'}{N} \\
=&\lim_{N\to\infty} \frac{\sum_i^N \{\sum_t^{T_i} X_{it,j} \times \sum_t^{T_i} X_{it,k}/T_i\}_{K\times K}}{N} \\
=&\frac{\sum_i^N E\left[\frac{\mathbf{B}_i}{T_i}\right]}{N}
\end{aligned}$$

where $\mathbf{B}_i$ is a $K \times K$ matrix and the (j,k)th entry of $\mathbf{B}_i$ is $\sum_t^{T_i} X_{it,j} \times \sum_t^{T_i} X_{it,k}$. It is bounded by $\frac{M}{T_i}$. (g) can be similarly proven. ■

Lemma 2 *Under model specification (A1) and regularity conditions, a. estimates of β, μ, α_i, and ξ_t from equations (1) to (3), i.e. $\hat{\beta}$, $\hat{\mu}$, $\hat{\alpha}_i$, and $\hat{\xi}_t$, are unbiased, and b. $\hat{\beta}$, $\hat{\mu}$, and $\hat{\xi}_t$ are consistent as $N_{co} \to \infty$.*

Proof: Under the two constraints on α_i and ξ_t (equations (2) and (3)), we have: $\bar{Y} = \bar{X}\beta + \mu$. Denote $\tilde{Y} = Y - \bar{Y}$ and $\tilde{\mathbf{X}} = \mathbf{X} - \bar{X}$. As shown in Wansbeek and Kapteyn (1989), β in this case can still be estimated using the within estimator. Multiplying both sides of demeaned equation (4) with $Q_{[\Delta]}$, we have $Q_{[\Delta]}\tilde{Y} = Q_{[\Delta]}\tilde{\mathbf{X}}\beta + Q_{[\Delta]}\tilde{\varepsilon}$, then it is easy to show that:

$$\begin{aligned}
\hat{\beta} &= (\tilde{\mathbf{X}}'Q_{[\Delta]}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'Q_{[\Delta]}\tilde{Y} = (\tilde{\mathbf{X}}'Q_{[\Delta]}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'Q_{[\Delta]}[\tilde{\mathbf{X}}\beta + \Delta(\alpha,\xi)' + \tilde{\varepsilon}] \\
&= \beta + (\tilde{\mathbf{X}}'Q_{[\Delta]}\tilde{\mathbf{X}})^{-1}\tilde{X}'Q_{[\Delta]}\tilde{\varepsilon}
\end{aligned}$$

Hence, $E[\hat{\beta}] = \beta + E[(\tilde{\mathbf{X}}'Q_{[\Delta]}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'Q_{[\Delta]}\tilde{\varepsilon}] = \beta$, and $E[\hat{\mu}] = E[\bar{Y} - \bar{X}\hat{\beta}] = \bar{Y} - \bar{X}\beta = \mu$.

Similarly,

$$Q_{[\tilde{\mathbf{X}}]}\tilde{Y} = Q_{[\tilde{\mathbf{X}}]}\Delta(\alpha,\xi)' + Q_{[\tilde{\mathbf{X}}]}\tilde{\varepsilon}$$

The level of fixed effects, $(\alpha,\xi)'$, can also be estimated using ordinary least squares under the two constrains (2) and (3), which is equivalent to the following constrained minimization problem:

$$\begin{aligned}
&Min_\gamma \quad (Q_{[\tilde{\mathbf{X}}]}\tilde{Y} - Q_{[\tilde{\mathbf{X}}]}\Delta\gamma)'(Q_{[\tilde{\mathbf{X}}]}\tilde{Y} - Q_{[\tilde{\mathbf{X}}]}\Delta\gamma) \\
&with \quad \Pi\gamma = 0
\end{aligned}$$

where $\gamma = (\alpha,\xi)'$, and $\Pi_{2\times(N+T)} = \begin{pmatrix} T_1, T_2, \ldots, T_N, 0, 0, \ldots, 0 \\ 0, 0, \ldots, 0, N_1, N_2, \ldots, N_T \end{pmatrix}$.
The solution to the minimization problem is given by the following equation:

$$\Phi\begin{pmatrix}\hat{\gamma}\\ \hat{\lambda}\end{pmatrix} = \begin{pmatrix}\Delta'Q_{[\tilde{\mathbf{X}}]}\Delta, & \Pi' \\ \Pi, & 0\end{pmatrix}\begin{pmatrix}\hat{\gamma}\\ \hat{\lambda}\end{pmatrix} = \begin{pmatrix}\Delta'Q_{[\tilde{\mathbf{X}}]}\tilde{Y}\\ 0\end{pmatrix}$$

where λ represents the corresponding Lagrangian multipliers. Finally, $\hat{\gamma} = (\hat{\alpha},\hat{\xi})' = \Phi_{11}^{-1}\Delta'Q_{[\tilde{\mathbf{X}}]}\tilde{Y}$.

Here Φ_{11}^{-1} is the upper-left block of Φ^{-1}. For unbiasedness of these estimates, notice that

$$\begin{aligned} E(\hat{\alpha},\hat{\xi})' &= E[\Phi_{11}^{-1}\Delta' Q_{[\tilde{\mathbf{X}}]}\tilde{Y}] \\ &= E[\Phi_{11}^{-1}\Delta' Q_{[\tilde{\mathbf{X}}]}\Delta(\alpha,\xi)'] \\ &= E[(I-\Phi_{12}^{-1}\Pi)(\alpha,\xi)'] \\ &= (\alpha,\xi)'. \end{aligned}$$

The second equality uses the fact that $Q_{[\tilde{\mathbf{X}}]}\tilde{\mathbf{X}}=0$. The third equality builts upon the definition of Φ_{11}^{-1} and Φ_{12}^{-1}: $\Phi_{11}^{-1}\Delta' Q_{[\tilde{\mathbf{X}}]}\Delta+\Phi_{12}^{-1}\Pi=I$. The last equality exploits the constraint $\Pi\gamma=\Pi(\alpha,\xi)'=0$.

The consistency of $\hat{\mu}$ is obvious. For $\hat{\beta}$ and $\hat{\xi}$, it is easy to show that:

$$\begin{aligned} \lim_{N\to\infty}\begin{pmatrix}\hat{\beta}\\ \hat{\xi}\end{pmatrix} = \begin{pmatrix}\beta\\ \xi\end{pmatrix} + \lim_{N\to\infty}\left[\begin{pmatrix}\tilde{\mathbf{X}}'\\ \Delta_2'\end{pmatrix} Q_{[\Delta_1]}(\tilde{\mathbf{X}},\Delta_2)\right]^{-1}\left[\begin{pmatrix}\tilde{\mathbf{X}}'\\ \Delta_2'\end{pmatrix} Q_{[\Delta_1]}\varepsilon\right] \\ \begin{pmatrix}\beta\\ \xi\end{pmatrix} + \lim_{N\to\infty}\left[\begin{pmatrix}\tilde{\mathbf{X}}'\\ \Delta_2'\end{pmatrix} Q_{[\Delta_1]}(\tilde{\mathbf{X}},\Delta_2)/N\right]^{-1}\left[\begin{pmatrix}\tilde{\mathbf{X}}'\\ \Delta_2'\end{pmatrix} Q_{[\Delta_1]}\varepsilon/N\right] \end{aligned}$$

And,

$$\begin{aligned} \begin{pmatrix}\tilde{\mathbf{X}}'\\ \Delta_2'\end{pmatrix} Q_{[\Delta_1]}(\tilde{\mathbf{X}},\Delta_2) &= \begin{pmatrix}\tilde{\mathbf{X}}'\tilde{\mathbf{X}} & \tilde{\mathbf{X}}'\Delta_2\\ \Delta_2'\tilde{\mathbf{X}} & \Delta_2'\Delta_2\end{pmatrix} - \begin{pmatrix}\tilde{\mathbf{X}}'\Delta_1\\ \Delta_2'\Delta_1\end{pmatrix}\left(\Delta_1'\Delta_1\right)^{-1}(\tilde{\mathbf{X}}\Delta_1', \Delta_2\Delta_1') \\ &= \begin{pmatrix}\tilde{\mathbf{X}}'\tilde{\mathbf{X}} & \tilde{\mathbf{X}}'\Delta_2\\ \Delta_2'\tilde{\mathbf{X}} & \Delta_2'\Delta_2\end{pmatrix} - \begin{pmatrix}\tilde{\mathbf{X}}'\Delta_1\\ \Delta_2'\Delta_1\end{pmatrix} diag\{\frac{1}{T_i}\}(\tilde{\mathbf{X}}\Delta_1', \Delta_2\Delta_1') \end{aligned}$$

Using Lemma 1, we know that as $N_{co}=N-N_{tr}\to\infty$, each term in the expression above will converge in probability to a fixed matrix.[A2] Using Slutsky's theorem, $\left[\begin{pmatrix}\tilde{\mathbf{X}}'\\ \Delta_2'\end{pmatrix} Q_{[\Delta_1]}(\tilde{\mathbf{X}},\Delta_2)/N\right]^{-1}$ also converges to a fixed matrix. Similarly, we can show that $\left[\begin{pmatrix}\tilde{\mathbf{X}}'\\ \Delta_2'\end{pmatrix} Q_{[\Delta_1]}\varepsilon/N\right]$ converges to $\mathbf{0}_{N_{co}\times 1}$ as $N_{co}\to\infty$, which leads to the consistency result.

On the contrary, $\hat{\alpha}_i$ is inconsistent when only $N_{co}\to\infty$ as the number of parameters changes accordingly.[A3] ■

Lemma 3 *Under model specification (A2) and regularity conditions, a. estimates of β, μ, α_i, ξ_t, λ_i, and f_t from equations (5) to (9), i.e. $\hat{\beta}$, $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\xi}_t$, $\hat{\lambda}_i$, and $\hat{f}_t$ are a. unbiased, and b. consistent as $N,T\to\infty$.*

Proof: Moon and Weidner (2015) show that all the coefficients of an IFE model can be estimated via a quasi maximum likelihood estimator and the estimates are unbiased as well as consistent when both N and T increase to infinity. We also know that estimates obtained from the EM algorithm

[A2] As $N_{co}\to\infty$, N also goes to infinity. And replacing $\mathbf{X}$ with $\tilde{\mathbf{X}}$ won't change the basic results.

[A3] It is easy to show that $\hat{\alpha}_i$ is inconsistent using the same algebra. As $\Delta_1'\Delta_1$ goes to the zero matrix when N_{co}, the error term does not vanish.

converge to the quasi-MLE solution since it is the unique extrema. Hence the lemma holds due to properties of QMLE. ■

Proposition 1 (Unbiasedness and Consistency of FEct) *: Under model specification (A1), as well as regularity conditions,*

$$\mathbb{E}[\widehat{ATT}_s] = ATT_s \text{ and } \mathbb{E}[\widehat{ATT}] = ATT;$$
$$\widehat{ATT}_s \xrightarrow{p} ATT_s \text{ and } \widehat{ATT} \xrightarrow{p} ATT \text{ as } N \to \infty.$$

Proof:

$$\begin{aligned}\widehat{ATT_t} &= \frac{1}{\sum_i D_{it}} \sum_{D_{it}=1} Y_{it} - \mathbf{X}_{it}'\hat{\beta} - \hat{\mu} - \hat{\alpha}_i - \hat{\xi}_t \\ &= \frac{1}{\sum_i D_{it}} \sum_{D_{it}=1} \left\{ \mathbf{X}_{it}'(\beta - \hat{\beta}) + (\mu - \hat{\mu}) + (\alpha_i - \hat{\alpha}_i) + (\xi_t - \hat{\xi}_t) + \delta_{it} \right\}\end{aligned}$$

Using lemma 1, we know that

$$\begin{aligned}\mathbb{E}[\widehat{ATT_t}] &= \frac{1}{\sum_i D_{it}} \sum_{i,D_{it}=1} \left\{ E[\mathbf{X}_{it}'(\beta - \hat{\beta})] + E[\mu - \hat{\mu}] + E[\alpha_i - \hat{\alpha}_i] + E[\xi_t - \hat{\xi}_t] + \delta_{it} \right\} \\ &= \frac{1}{\sum_i D_{it}} \sum_{i,D_{it}=1} \delta_{it} \\ &= ATT_t\end{aligned}$$

Therefore, unbiasedness holds. For consistency, we know from the proof of lemma 2 that:

$$\begin{aligned}\lim_{N,T\to\infty} \widehat{ATT_t} &= \lim_{N,T\to\infty} \frac{1}{\sum_i D_{it}} \sum_{i,D_{it}=1} Y_{it} - \mathbf{X}_{it}'\hat{\beta} - \hat{\mu} - \hat{\alpha}_i - \hat{\xi}_t \\ &= \frac{1}{\sum_i D_{it}} \sum_{i,D_{it}=1} (\delta_{it} + \alpha_i - \hat{\alpha}_i) + \bar{X}_{it}'(\beta - \hat{\beta}) + (\mu - \hat{\mu}) + (\xi_t - \hat{\xi}_t)\end{aligned}$$

Lemma 2 indicates that as $N_{co} \to \infty$, $\hat{\mu}$, $\hat{\beta}$ and $\hat{\xi}_t$ converge to μ, β, and ξ_t, respectively. The only thing to be shown is $\frac{1}{\sum_i D_{it}} \sum_{i,D_{it}=1} (\alpha_i - \hat{\alpha}_i) = 0$. This is true since $E[\alpha_i - \hat{\alpha}_i] = 0$ and $\text{Var}[\alpha_i - \hat{\alpha}_i]$ is bounded by the regularity conditions. Therefore $\lim_{N,T\to\infty} \widehat{ATT_t} = \frac{1}{\sum_i D_{it}} \sum_{i,D_{it}=1} \delta_{it} = ATT_t$, consistency holds. ■

Proposition 2 (Unbiasedness and Consistency of IFEct) *: Under model specification (A2), as well as regularity conditions,*

$$\mathbb{E}[\widehat{ATT}_s] = ATT_s \text{ and } \mathbb{E}[\widehat{ATT}] = ATT;$$
$$\widehat{ATT}_s \xrightarrow{p} ATT_s \text{ and } \widehat{ATT} \xrightarrow{p} ATT \text{ as } N, T \to \infty.$$

Proof: From lemma 3, we know that estimates for β, μ, α_i, ξ_t, λ_i, and f_t are unbiased and consistent as $N, T \to \infty$. Hence, $\widehat{ATT_t}$ and $\widehat{ATT}$ are also unbiased and consistent, following the same logic in the proof of Proposition 1. ■

A.2.2. FEct as a Weighting Estimator

Proposition 3 (FEct as a weighting estimator) *: Under model specification (A1), and when there is no covariate,*

$$\widehat{ATT}_t = \frac{1}{\sum_i D_{it}} \sum_{i,D_{it}=1} [\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}]$$

Where $\hat{Y}_{it}(0) = \mathbf{W}\mathbf{Y}_{D_{it}=0}$ *is a weighted average of untreated observations.*

Proof: When there is no covariate,

$$\hat{\mu} = \bar{Y}_{D_{it}=0} = \frac{1}{N_{co}} \iota_{N_{co}}^{'} \mathbf{Y}_{D_{it}=0}$$

$$\hat{\alpha}_i + \hat{\xi}_t = \nu_{it}' \begin{pmatrix} \hat{\alpha} \\ \hat{\xi} \end{pmatrix} = \nu_{it}' \Phi_{11}^{-1} \Delta' \tilde{Y}_{D_{it}=0} = [\nu_{it}' \Phi_{11}^{-1} \Delta' (I - \frac{1}{N_{co}} \iota_{N_{co}} \iota_{N_{co}}^{'})] \mathbf{Y}_{D_{it}=0}$$

Therefore,

$$\begin{aligned}
\widehat{ATT_t} &= \frac{1}{\sum_i D_{it}} \sum_{i,D_{it}=1} Y_{it} - \hat{\mu} - \hat{\alpha}_i - \hat{\xi}_t \\
&= \frac{1}{\sum_i D_{it}} \sum_{i,D_{it}=1} Y_{it} - [\nu_{it}' \Phi_{11}^{-1} \Delta' (I - \frac{1}{N_{co}} \iota_{N_{co}} \iota_{N_{co}}^{'}) + \frac{1}{N_{co}} \iota_{N_{co}}^{'}] \mathbf{Y}_{D_{it}=0} \\
&= \frac{1}{\sum_i D_{it}} \sum_{i,D_{it}=1} (Y_{it} - \mathbf{W}' \mathbf{Y}_{D_{it}=0}) \\
&= \frac{1}{\sum_i D_{it}} \sum_{i,D_{it}=1} [\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}]
\end{aligned}$$

.

where $\mathbf{W}' = \nu_{it}' \Phi_{11}^{-1} \Delta' (I - \frac{1}{N_{co}} \iota_{N_{co}} \iota_{N_{co}}^{'}) + \frac{1}{N_{co}} \iota_{N_{co}}^{'}$. It is easy to see that:

$$[\Delta' (I - \frac{1}{N_{co}} \iota_{N_{co}} \iota_{N_{co}}^{'}] \mathbf{Y}_{D_{it}=0} = \begin{pmatrix} T_1(\bar{Y}_{1.} - \bar{Y}_{D_{it}=0}) \\ \vdots \\ T_N(\bar{Y}_{N.} - \bar{Y}_{D_{it}=0}) \\ N_1(\bar{Y}_{.1} - \bar{Y}_{D_{it}=0}) \\ \vdots \\ N_T(\bar{Y}_{.T} - \bar{Y}_{D_{it}=0}) \end{pmatrix}.$$

In addition, Φ_{11}^{-1} is approximately the precision matrix for all the fixed effects dummy variables. Therefore, an observation Y_{it}'s weight is larger when there are more untreated observations for either unit i or period t.

For a generalized DiD design, we can calculate the weight of each untreated observation. Now the first N_0 units belong to the control group, and the rest N_1 units are in the treated group. The treatment turns on after T_0 periods and lasts for T_1 periods. We can derive the following expression:

$$\begin{pmatrix} \hat{\alpha}_i \\ \hat{\xi}_t \end{pmatrix} = \begin{pmatrix} \frac{1}{T} \sum_{t=1}^{T} \tilde{Y}_{it} + \frac{1}{NT - N_1 T_1} \frac{N_1}{N_0} \sum_{t=1}^{T_0} \sum_{i=1}^{N} \tilde{Y}_{it} - \frac{1}{NT - N_1 T_1} \frac{T_1 N_1}{T N_0} \sum_{i=1}^{N_0} \sum_{t=1}^{T} \tilde{Y}_{it} \\ \frac{1}{N} \sum_{i=1}^{N} \tilde{Y}_{it} + \frac{1}{NT - N_1 T_1} \frac{T_1}{T_0} \sum_{i=1}^{N_0} \sum_{t=1}^{T} \tilde{Y}_{it} - \frac{1}{NT - N_1 T_1} \frac{T_1 N_1}{T_0 N} \sum_{t=1}^{T_0} \sum_{i=1}^{N} \tilde{Y}_{it} \end{pmatrix}$$

$\widehat{Y_{it}(0)} = \hat{\mu} + \hat{\alpha}_i + \hat{\xi}_t = \sum_{t=1}^{T}\sum_{i=1}^{N} W_{it}\tilde{Y}_{it}$. It is thus clear that FEct uses all the observations with $D_{it} = 0$ to construct the counterfactual. ■

A.3. Monte Carlo Evidence

In this section, we report results of three sets of Monte Carlo exercises to demonstrate (1) the finite sample properties of the proposed inferential methods; (2) the differences between the IFEct and MC estimators, and (3) the main advantages of the equivalence test over the F test. Before doing so, we first describe the data generating processes (DGP) of the simulated sample.

A.3.1. Describing the DGP of the Simulated Example

We describe the DGP of the simulated example as follows. Treatment assignment follows staggered adoption (Athey and Imbens 2018): each unit is assigned a number of pretreatment periods $T_{0i} \in \{20, 23, 26, 29, 32, 35\}$ such that $D_{it} = 0$ if $1 \leq t \leq T_{0i}$ and $D_{it} = 1$ if $t > T_{0i}$. Treatment assignment is determined by a latent variable $tr_i^* = \lambda_{i1} + \lambda_{i2} + \alpha_i + \omega_i$, in which $\omega_i \sim N(0, 1)$ are i.i.d. white noises. A one-to-one mapping from the percentile of tr_i^* to T_{0i} exists. Specifically, $T_{0i} = 35$ (controls) if $pct(tr_i^*) \leq 50$, $T_{0i} = 32$ if $pct(tr_i^*) \in (50, 60]$, $T_{0i} = 29$ if $pct(tr_i^*) \in (60, 70]$, $T_{0i} = 26$ if $pct(tr_i^*) \in (70, 80]$, $T_{0i} = 23$ if $pct(tr_i^*) \in (80, 90]$, $T_{0i} = 20$ if $pct(tr_i^*) \in (90, 100]$. This means that units that have low values of tr_i^* are more likely to be assigned to the control group ($T_{0i} = 35$) and units that have high values of tr_i^* are more likely to receive the treatment early on. It is obvious that selection on the factor loadings and unit fixed effects will lead to biases in the causal estimates if the model does not account for them.

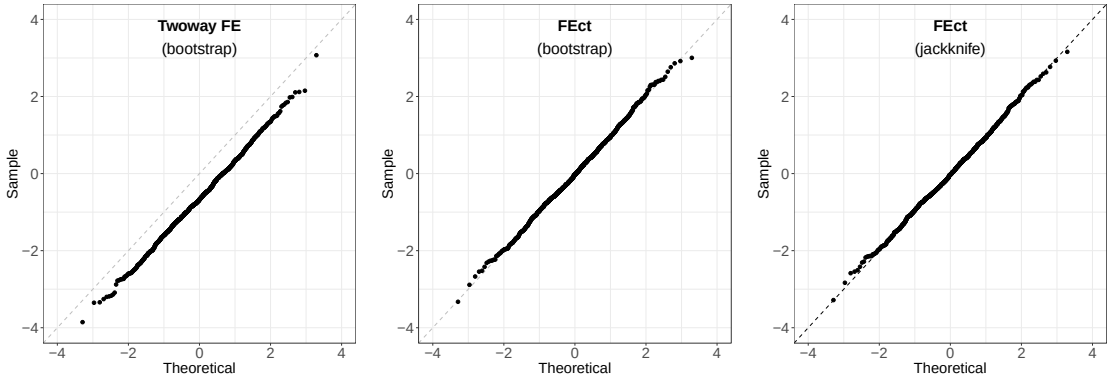
The individual treatment effect for unit i at time t is generated by $\delta_{it,t>T_{0i}} = 0.2(t - T_{0i}) + e_{it}$, in which e_{it} is i.i.d. $N(0, 1)$. This means the expected value of the treatment effect gradually increases as a unit takes up the treatment, e.g. from 0.2 in the first period after receiving the treatment to 2.0 in the tenth period. The factors are assumed to be two-dimensional: $f_t = (f_{1t}, f_{2t})$; so are the factor loadings: $\lambda_i = (\lambda_{1i}, \lambda_{2i})$. f_{1t} is a drift process with a deterministic trend: $f_{1t} = a_t + 0.1t + 3$, in which $a_t = 0.5a_{t-1} + \nu_t$ and $\nu_t \overset{i.i.d}{\sim} N(0, 1)$. f_{2t} is an i.i.d $N(0, 1)$ white noise. Both λ_{i1} and λ_{i2} are i.i.d $N(0, 1)$. Two covariates $X_{1,it}$ and $X_{2,it}$ are included in the model. They are both i.i.d. $N(0, 1)$. Unit fixed effects $\alpha_i \sim N(0, 1)$. Time fixed effects ξ_t also follows a stochastic draft as f_{1t}. The error term ε_{it} is also i.i.d. $N(0, 1)$. Note that this DGP satisfies Assumptions 1-3.

A.3.2. Quantifying Uncertainties

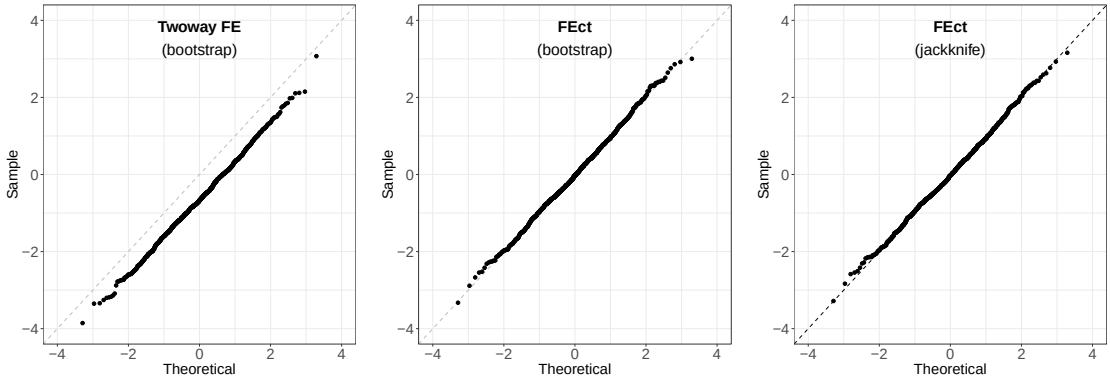
We study the finite sample properties of the bootstrap and jackknife variance estimators based on DGPs similar to what is described above. We simulate samples with $T = 20$ and $N = 30, 50, 100$ and a staggered adoption treatment assignment mechanism. We assume that no time-varying confounders exist while the treatment effects are heterogeneous, hence, FEct is consistent for the ATT while the two-way fixed effects model is not.

We present additional evidence that the bootstrap and jackknife methods provide good approximation for the sampling distribution even when the sample size is relatively small, i.e., $N = 30$ and $N = 50$.

FIGURE A3. QQ PLOTS FOR BOOTSTRAPPED AND JACKKNIFE STANDARD ERRORS



(a) $N = 30$



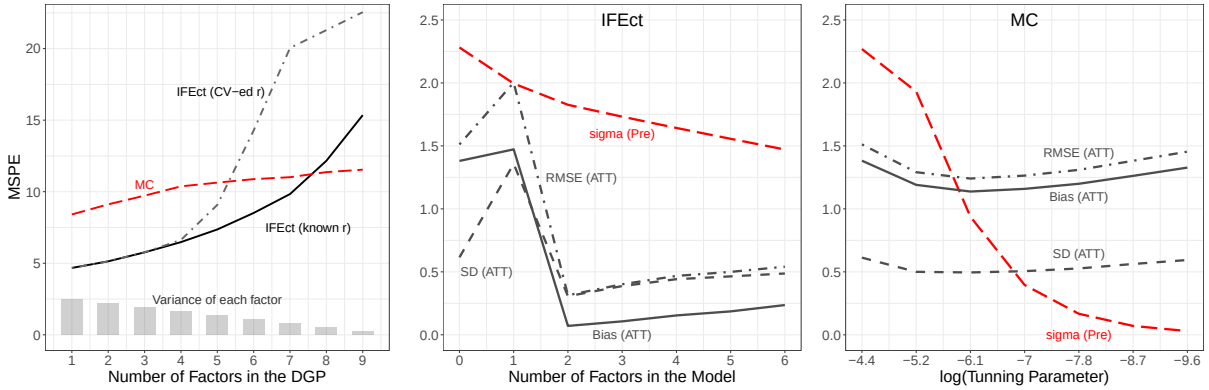
(b) $N = 50$

Note: The above figures show the standard Gaussian QQ plot of the standardize errors $(\widehat{ATT} - ATT)/(\widehat{\text{Var}}(\widehat{ATT}))^{1/2}$ for the following combination of estimators and inferential methods: (1) two-way fixed effects with block bootstrapped standard errors; (2) FEct with block bootstrapped standard errors; and (3) FEct with jackknife standard errors, each aggregated from 1,000 simulations using samples with dimensions $T = 20$ and $N = 30$ or $N = 50$. The 45-degree indicates the benchmark: consistent point estimates with perfectly calibrated Gaussian standard errors.

A.3.3. IFEct versus MC

We compare the performance of the IFEct and MC estimators using DGPs similar to that of the simulated example: $Y_{it} = \delta_{it} D_{it} + 5 + \frac{1}{\sqrt{r}} \sum_{m=1}^{r} \lambda_{im} \cdot f_{mt} + \alpha_i + \xi_t + \varepsilon_{it}$. We simulate samples of 200 units and 30 time periods, and all treated units receive the treatment at period 21 ($T_0 = 20$). Following Li (2018), we vary the number of factors r from 1 to 9 and adjust a scaling parameter $\frac{1}{\sqrt{r}}$ such that the total contribution of all factors (and their loadings) to the variance of Y remains constant. Our intuition is that IFEct (i.e., hard impute) performs better than MC (i.e., soft impute) when only a small number of factors are present and each of them exhibits relatively strong signals while MC outperforms IFEct when a large set of weak factors exist. In other words, MC should handle sparsely distributed factors better than parametric models like IFEct.

FIGURE A4. MONTE CARLO EXERCISES: IFECT VS. MC



(a) Comparing Estimators (b) Two-factor DGP: IFEct (c) Two-factor DGP: MC

Note: The above figures show the results from two Monte Carlo exercises that compare IFEct with MC. Figure (a) compares the mean squared prediction errors (MSPEs) for treated counterfactuals using the IFEct and MC estimators with different DGPs in which the total variance of all factors are kept constant. Figures (b) and (c) compare the two estimators in terms of mean squared error (MSE) for the ATT, biases and standard deviations (SD), as well as pretreatment root residual sums of squares (sigma (Pre)) using different tuning parameters; the DGP is an IFE model with two factors.

The results are shown in Figure A4(a), which depicts the MSPE of treated counterfactuals, i.e., $\frac{1}{\#\mathbf{1}\{(i,t)|D_{it}=1\}} \sum_{D_{it}=1} [Y(0) - \hat{Y}_{it}(0)]^2$, from 500 simulations using these two methods. The black solid line and gray dashed line represent the MSPE of IFEct with the correct number of factors (r) and with cross-validated r's, respectively, while the red dot-dashed line marks the MSPEs of the MC estimator with a crossed validated tuning parameter λ. The result shows that MC gradually catches up with, and eventually beats, IFEct (with correctly specified r) as the number of factor grows and each factor produces weaker signals. It also suggests that, when factors become weaker, it is more difficult for the cross-validation scheme to pick them up, resulting in worse predictive performance, while the MC estimator is robust to a large number of factors because factors and loadings are not directly estimated.

In Figure A4(b), we fix the DGP with two factors and compare the performances of IFEct and MC with different tuning parameters. The black solid lines, gray long-dashed lines, gray

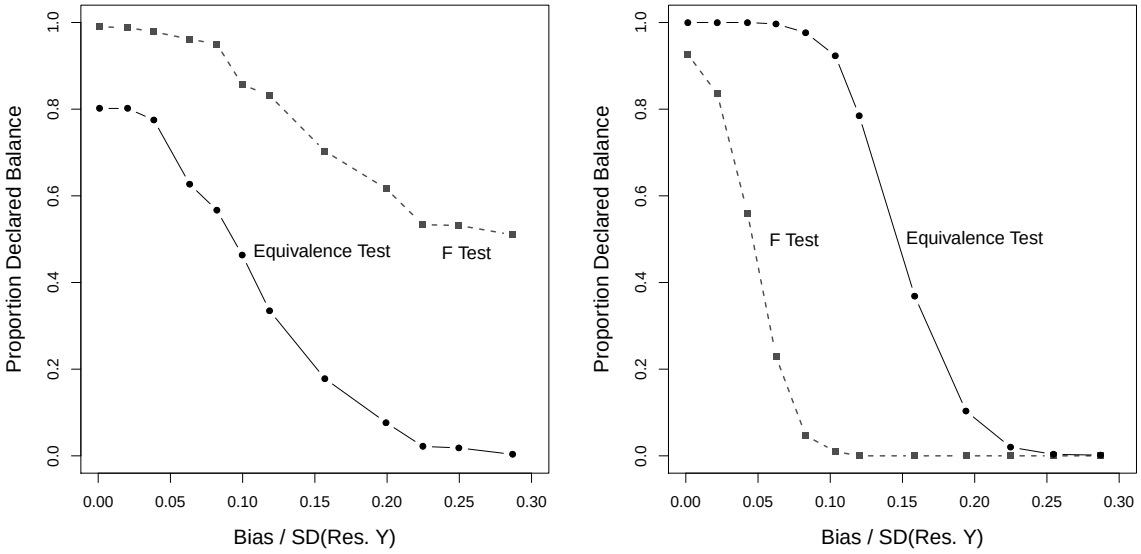
dot-dashed lines, and red long-dashed lines represent (1) the biases of the ATT estimates, i.e, $\mathbb{E}(\widehat{ATT} - ATT)$, (2) their standard deviations (SD), (3) their MSEs and (4) standard deviation of residuals in the pretreatment periods—a measure of pretreatment model fitness—respectively. We observe several patterns. First, as expected, with the most favorable tuning parameters, IFEct outperforms MC in terms of both bias and variance because an MC model is mis-specified. Second, as more factors are included in the IFEct model or the tuning parameter becomes smaller with MC, both models start to over-fit: the model fits pretreatment data better and better while MSE for the ATT deteriorates. These findings strongly suggest that the pretreatment model fitness is poor indicator of model performance and a high level of model fitness in the pretreatment periods does not necessarily lead to more precise estimates of the ATT.

A.3.4. F Test versus the Equivalence test.

As explained in the paper, we prefer the equivalence test to the F test for testing no pre-trend for two reasons: (1) the former is more conservative than the latter in the presence of a large confounder when the sample size is small; and (2) when the sample size is relatively large, the former can tolerate confounders that only result in a small amount of bias in the causal estimates while the latter cannot. To illustrate these, we simulate data using the following DGP similar to that in the previous section but with only one factor: $Y_{it} = \delta_{it} D_{it} + 5 + k \cdot \lambda_i f_t + \alpha_i + \xi_t + \varepsilon_{it}$, in which we vary k to adjust the influence of a potential confounder $U_{it} = \lambda_i f_t$, which is correlated with D_{it}. For each k, we run 600 simulations. In each simulation, we first generate a sample of $N = 100$ units (50 treated and 50 controls) of 40 periods. We estimate a FEct model without taking into account the time-varying confounder. We then expand the sample size such that $N = 300$ and re-do the analysis.

In Figure A5(a), we plot the proportion of times the equivalence test (solid line) or the F test (dashed line) backs the no-time-varying-confounder assumption against the normalized bias induced by the confounder when $N = 100$. It shows that it is highly likely that an F test cannot reject the null of zero residual average due to lack of power, even when the biases are large. In contrast, the probability that the equivalence test rejects inequivalence (hence, declaring equivalence) drops quickly as the bias increases. In other words, the equivalence test is more powerful in detecting imbalances than the conventional F test. Figure A5(b) shows that, when the sample size is relatively large ($N = 300$), the non-rejection rate of the F test declines quickly as the influence of the confounder grows. In comparison, the equivalence test rejects inequivalence (hence, declaring equivalence) when a inconsequential confounder is at present; as the confounder becomes more influential (e.g. causing a bias of 0.15 standard deviation of the residualized Y in the ATT), it starts to sound the alarm. These patterns are similar to what Hartman and Hidalgo (2018) report in a cross-sectional setting.

FIGURE A5. MONTE CARLO EXERCISES: F VS. EQUIVALENCE TESTS

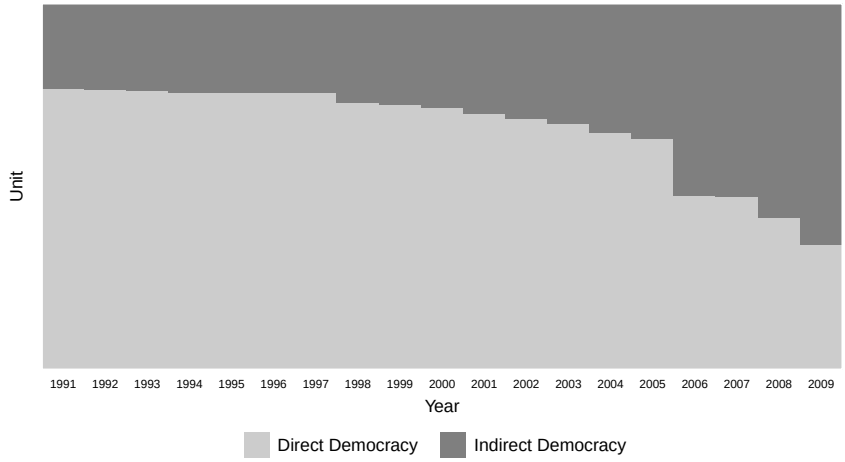


(a) $N = 100$ (b) $N = 300$

Note: The above figures show the results from Monte Carlo exercises that compare the F test and the equivalence test when an unobserved confounder exists. In plot (a), $N = 100$; in plot (b), $N = 300$. Each dot is based on results from 600 simulations.

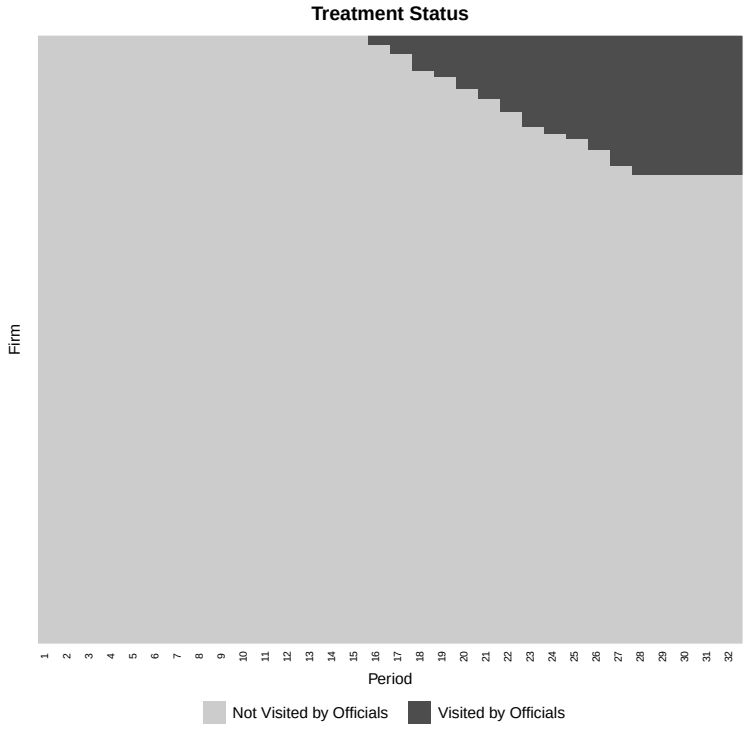
A.4. Additional Information on Empirical Examples

FIGURE A6. TREATMENT STATUS: HAINMUELLER AND HANGARTNER (2015)



Note: The above figure plots the treatment status for the first 50 units using data from Hainmueller and Hangartner (2015).

FIGURE A7. TREATMENT ASSIGNMENT: OFFICIALS' VISIT TO FIRMS AND FIRMS' ACCESS TO LOANS



Note: The above figure plots the treatment status (officials' visits to firms) using data obtained by the authors.