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

Licheng Liu[*] (Tsinghua) Ye Wang[†] (NYU) Yiqing Xu[‡] (Stanford)

13th July 2019

Abstract

We introduce a simple framework of *counterfactual estimators* that directly impute counterfactuals for treated observations in a time-series cross-sectional setting with a dichotomous treatment. Examples include (1) the fixed-effect counterfactual estimator, (2) the interactive fixed-effect counterfactual estimator, and (3) the matrix completion estimator, which differ in the underlying model of predicting treated counterfactuals. They provide more reliable causal estimates than conventional two-way fixed effects models when the treatment effect is heterogeneous or there exist unobserved time-varying confounders. Moreover, we propose two diagnostic tests, an equivalence test and a placebo test, accompanied by visualization tools, to assist researchers to gauge the validity of the identifying assumptions. We illustrate these methods with three empirical examples from political economy and develop an open source package, `fect`, in `R` to facilitate implementation.

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

[*]Department of Finance, Tsinghua University. Email: liulch.16@sem.tsinghua.edu.cn.
[†]Department of Political Science, New York University. Email: yw1576@nyu.edu.
[‡]Department of Political Science, Stanford University. Email: yiqingxu@stanford.edu.

1. Introduction

The linear two-way fixed effects model is one of the most commonly used statistical routines in the social sciences for establishing causal relationships with observational data. With time-series cross-sectional (TSCS) data or panel data, researchers often choose a two-way fixed-effects approach as their chief identification strategy because a potentially large set of unobserved unit- and time-invariant confounders are being controlled for. Two crucial assumptions underlie this approach. The first one is linearity, which assumes 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 is reduced to the assumption of *constant treatment effect*. The second assumption is the *absence of time-varying confounders*. It means that there exists no unobserved, time-changing factors that are correlated with the treatment and the potential outcomes at the same time. When there exist only two periods and two treatment groups, this assumption implies "parallel trends," i.e., the average non-treated potential outcomes of the treated and control groups follow parallel paths. As Angrist and Pischke (2009) famously put in their textbook, two-way fixed effects models produce interpretable causal estimate only in "parallel worlds." Failures of these two assumptions lead to biases in the estimates for researchers' causal quantity of interest.[1]

In this paper, we introduce a group of estimators that relax these two assumptions under a simple, unified framework when the treatment is dichotomous. These estimators take observations under the treatment condition as missing data and directly estimate their counterfactuals. We thus call them *counterfactual estimators*. They include (1) the fixed-effect counterfactual (FEct) estimator, of which difference-in-differences (DiD) is a special case, (2) the interactive fixed-effect counterfactual (IFEct) estimator (Gobillon and Magnac 2016;

[1]The second assumption is implied by strict exogeneity. A recent paper by Imai and Kim (2018) clarifies the assumptions needed for the causal identification of the average treatment effect using linear two-way fixed effects models: (a) no unobserved time-varying confounders, (b) past outcomes do not directly affect current treatment, (c) past treatments do not directly affect current outcome (or what they call *no carryover effect*), and (d) linearity. (a) and (b) are implied by strict exogeneity while (c) and (d) are implied by the functional form assumption. See more discussion below.

Xu 2017) and (3) the matrix completion (MC) estimator (Athey et al. 2018; Kidziński and Hastie 2018). They differ from each other in the way of constructing treated counterfactuals. Because treated observations are not used in the modeling stage, these estimators allow the treatment effects to be arbitrarily heterogeneous. As a result, the constant treatment effect assumption is relaxed.

To address violations of the second assumption, both the IFEct and MC estimators attempt to condition on unobserved, time-varying confounders using a latent factor approach. Mathematically, both estimators seek to construct a lower-rank approximation of the outcome data matrix using information of non-treated observations only. They differ in the way of regularizing the latent factor model. Simulations and results from the empirical examples suggest that they can provide more reliable causal effects than conventional methods when the identifying assumptions appear to fail.

Moreover, this paper aims to provide practical guidance to social scientists when they analyze TSCS data with a dichotomous treatment. An increasingly popular practice among social scientists to evaluate 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 in a two-way fixed effects model. The idea behind this test is as follows. Assuming that no time-varying confounder exists, the residuals from the two-way fixed effects model using only non-treated data should be uncorrelated with the number of time periods leading to the onset of the treatment. If this is true, visually, we shall not expect to see the averages of residuals for the treated units exhibit a trend leading towards the beginning of the treatment, known as a "pre-trend." If no strong "pre-trend" is observed, researcher will have more confidence in the identifying assumptions of the two-way fixed effects approach. However, this method has several limitations. First, it relies on the parametric assumption that the treatment effect

is the same for all units in a given time period. Second, researchers often rely on heuristic rules when conducting such a test, e.g. by either eye-balling whether a strong pre-trend exist or checking whether there exist a pre-treatment coefficient that is statistically different from zero using a t-test, but there is no guarantee that these practices can always offer correct guidance. For example, the t-test strategy may suffer from a multiple-testing issue: as the number of pre-treatment periods grows, it is more likely to find at least one period in which the null hypotheses of no difference is rejected due to pure randomness of the data. Third, a time-varying confounder does not have to present itself as a trending factor. For example, a zero-mean seasonal shock that differentially affects the treatment and control groups will also contribute to biases in the causal estimates.

Taking advantage of the counterfactual estimators, we improve the practice of estimating 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 additional to visual checks, we develop two formal diagnostic tests for the presence of time-varying confounders.

The first one is an equivalence test, which directly tests whether the error term is orthogonal to the timing of the treatment in the pre-treatment periods. To overcome the multiple-testing issue, one straight-forward solution would be to develop a variant of the Wald test (or F-test) to jointly test whether averages of the residuals from the treatment units are close to zero for all pre-treatment periods. However, as Hartman and Hidalgo (2018) demonstrate in a cross-sectional setting, a test for inequivalence, such as a t-test or an F-test, has two potential problems when being used to test equivalence. First, non-rejection does not imply equivalence because the test can be under-powered due to insufficient data. Second, and perhaps more importantly for our setting, when data are abundant, such a test will almost always lead to rejection in the face of a small confounder whose influence on the casual estimates is neglectable. To address these challenges, Hartman and Hidalgo (2018) recommend to flip the null and alternative hypotheses. In other words, they argue

that researchers should ask whether they have enough evidence to reject the hypothesis of inequivalence; as a result, the more data researchers collect, the more likely they are able to accept the identifying assumption as valid when it is indeed the case (or sufficiently close). We follow their advice and develop an equivalence test for the pre-trends in a TSCS setting. We provide Monte Carlo evidence that the equivalence test performs better than a Wald test in terms of reducing the chances of declaring inequivalence in the face of non-consequential confounders. The main limitation of this test is that the choice of the equivalence bound may require user discretion.

We complement the equivalence test with a placebo test, in which we hide a few periods of observations for units that later receive the treatment. We then use the same counterfactual estimators to estimate the "treatment effects" for those periods, which are presumably zero under the identifying assumptions. If the estimated average treatment effects in the placebo periods are statistically different from zero, researchers should be concerned about the validity of the assumptions. The placebo test has the merits of being intuitive and robust to model mis-specification, however, it does not make use of information in all pre-treatment periods and can be under-powered.

The contribution of this paper is two-folded. First, it introduces three novel estimators that recently emerge from the literature under the same analytical framework. These estimators relax the constant treatment effect assumption and two of them (i.e., the IFEct and MC estimators) can deal with time-varying confounders that can be decomposed into time-varying factors and unit-specific factor loadings. Although these estimators already exist—specifically, the FEct estimator is a special case of the generalized synthetic control (gsynth) method when the number of factors is set to zero (Xu 2017), the IFEct estimator is proposed by Gobillon and Magnac (2016) and Xu (2017), and the MC estimator is first introduced by Athey et al. (2018) to the social sciences—it is important to compare them with one another in the same setting for both theoretical and practical reasons. Therefore, our second main contribution is to a set of develop visualization and diagnostic tools to

allow to assist researchers choose the most suitable estimator for their applications. Even if researchers in the end opt for the FEct estimator because of its simplicity, the diagnostic tools we provide will make it much more convenient and transparent to evaluate the key identifying assumption, i.e., the absence of time-varying confounders.

This paper is closely related to an emerging literature on causal inference with TSCS data. Compared with conventional two-way fixed effect models, the counterfactual estimators relax the constant treatment effect assumption with only minor efficiency loss. Compared with the weighted fixed effect (wFE) estimator proposed by Imai and Kim (2018), the counterfactual estimators are often more efficient and easier to accommodate covariates (with an additional linearity assumption) and time trends. Because the counterfactual estimators preserve the temporal structure of the longitudinal data, a placebo test can be more easily constructed, as previously discussed. Compared with the generalized synthetic control method (Xu 2017), the counterfactual estimators can accommodate more complex TSCS designs such as staggering adoption (Athey and Imbens 2018) and treatment reversal (with an additional assumption that the treatment effect does not carry over to future periods). Compared with TSCS methods based on matching and reweighting (e.g. Imai, Kim and Wang 2018; Hazlett and Xu 2018; Strezhnev 2018), the counterfactual estimators can not only accommodate more complex designs, but are also more efficient because it makes use of all available data.

These advantages do not come at no cost. 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). As the authors correctly point out, these are the cost researchers have to pay if they believe unit-level time-invariant heterogeneities are important confounding factors. However, the strict exogeneity assumption underlying the IFEct and MC estimators are presumably less stringent than that behind a conventional fixed effects model because

they condition on latent factors that can potentially capture a large set of time-varying confounders. Overall, we believe that the counterfactual estimators strike a good balance among model simplicity and robustness, and the reliability of the identifying assumptions.

The rest of the paper is organized as follows. Section 2 discusses three counterfactual estimators under a simple, unified framework and briefly explains the estimation strategies. We then introduce the diagnostic tools, including the dynamic treatment effects plot and two tests, in Section 3. Section 4 provides Monte Carlo evidence for the estimators and tests. In Section 5, we apply these methods to three empirical examples in political economy. The last section concludes.

2. Counterfactual Estimators

In this section, we introduce three counterfactual estimators. Before doing so, we first set up the basic analytical framework, define the main causal quantity of interest, discuss the required identification assumptions, and explain the drawbacks of conventional two-way fixed effects models.

2.1. Setting

The workhorse model we use throughout this paper is a variant of the linear (interactive) fixed effects model with a dichotomous treatment, in which the fixed effects can be both additive and interactive. For notational convenience, suppose that we have a balanced panel of N units and T periods in which both N and T are large. All the estimators can easily be extended to accommodate unbalanced panels. We first make the following functional form assumption:

Assumption 1 (Two-way fixed effects) *For any* $i = 1, 2, \cdots, N$ *and* $t = 1, 2, \cdots, T$,

$$Y_{it} = \delta_{it} D_{it} + X_{it}'\beta + \alpha_i + \xi_t + \varepsilon_{it}, \tag{1}$$

in which Y_{it} is the outcome for unit i at time t; D_{it} is treatment indicator that equals 1 if unit i is treated at time t and equals 0 otherwise; δ_{it} is the treatment effect on unit i at time t, X_{it} is a $(p \times 1)$ vector of exogenous covariates; β is a $(p \times 1)$ vector of unknown parameters; α_i and ξ_t are *additive* unit and time fixed effects, respectively; and ε_{it} represents unobserved idiosyncratic shocks for unit i at time t and has zero mean. Note that Equation (1) is different from a conventional two-way fixed effects model in that we allow the treatment effect to be heterogeneous both across units and over time. In that regard, the two-way FE model is a special case of Equation (1) when δ_{it} is constant.[2] It is worth noting that we still assume the relationship between covariates X and the outcome Y to be linear. This model can be extended to incorporate interactive fixed effects (Bai 2009; Xu 2017):

Assumption 1a (Interactive fixed effects) *For any* $i = 1, 2, \cdots, N$ *and* $t = 1, 2, \cdots, T$,

$$Y_{it} = \delta_{it} D_{it} + X_{it}' \beta + \lambda_i^{'} f_t + \alpha_i + \xi_t + \varepsilon_{it} \tag{2}$$

in which $f_t = [f_{1t}, \ldots, f_{rt}]'$ is an $(r \times 1)$ vector of unobserved common factors and $\lambda_i = [\lambda_{1r}, \ldots, \lambda_{ir}]'$ is an $(r \times 1)$ vector of unknown factor loadings. For identification, we need to impose two constrains: $\mathbf{\Lambda}'\mathbf{\Lambda} = diagnal$ and $\mathbf{F}'\mathbf{F}/T = \mathbf{I}_r$ in which $\mathbf{\Lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_N]'$ and $\mathbf{F} = [f_1, f_2, \ldots, f_T]'$. Intuitively, factors can be understood as time-varying trends that affect each unit differently and factor loadings capture their heterogeneous impacts caused by each unit's various unobserved characteristics. For example, the 2008 financial crisis hit the US economy but it had heterogeneous impacts on different cities depending on a city's demographics, real estate market, industrial structure, etc. Because $\lambda_i^{'} f_t$ captures such differential trends, the no-time-varying-confounder assumption will be relaxed to a certain extent.[3] When the number of factors $r = 0$, Equation (2) is reduced to Equation (1).

Using notations from the causal inference literature, we define the potential outcomes of

[2] In order to probe the validity of the no time-varying confounder assumption, scholars sometime assume a more flexible "dynamic effects" model: $Y_{it} = \sum_{s=T_{0i}-p}^{T_{0i}+q} \delta_s D_{it} + X_{it}' \beta + \alpha_i + \xi_t + \varepsilon_{it}$. Still, the treatment effects are assumed to be constant across units in a specific time period.

[3] More details can be found in Xu (2017).

Y_{it} as $Y_{it}(D_{it})$. Equation (2) therefore can be rewritten as:

$$\begin{cases} Y_{it}(0) = X_{it}'\beta + \lambda_i' f_t + \alpha_i + \xi_t + \varepsilon_{it}; \\ Y_{it}(1) = Y_{it}(0) + \delta_{it}. \end{cases}$$

As clarified by Imai and Kim (2018), Assumption 1a (or 1) implies the follow *causal* assumptions: (a) *no carryover effect*, i.e., $Y_{it}(D_{i1}, D_{i2}, \cdots, D_{it}) = Y_{it}(D_{it})$, and (b) *no lagged dependent variables (LDVs)*, i.e. past outcomes do not directly affect current outcome.[4]

Similar to the no-time-varying-confounder assumption required by a conventional two-way fixed effects model, we need to impose the assumption of *strict exogeneity* for identification:

Assumption 2 (Strict exogeneity for two-way fixed effects) *For any* $i, j = 1, 2, \cdots, N$ *and* $t, s = 1, 2, \cdots, T$,

$$\varepsilon_{it} \perp\!\!\!\perp D_{js}, X_{js}, \alpha_j, \xi_s$$

Assumption 2 implies $\varepsilon_{it} \perp\!\!\!\perp D_{js} | X_{js}, \alpha_j, \xi_s, \ \forall i, j, t, s$. Therefore,

$$\begin{aligned} &\mathbb{E}[Y_{it'}(0) - Y_{it}(0) | D_{it'} = 1, D_{it} = 0, X_{it}, X_{it'}, \xi_{t'}, \xi_t, \alpha_i] \\ = &\mathbb{E}[Y_{it'}(0) - Y_{it}(0) | D_{it'} = D_{it} = 0, X_{it}, X_{it'}, \xi_{t'}, \xi_t, \alpha_i], \quad \forall i, t', t \end{aligned}$$

which is commonly know as the parallel/common trends assumption in DID settings.

Assumption 2a (Strict exogeneity for interactive fixed effects) *For any* $i, j = 1, 2, \cdots, N$ *and* $t = 1, 2, \cdots, T$,

$$\varepsilon_{it} \perp\!\!\!\perp D_{js}, X_{js}, \alpha_j, \xi_s, \lambda_j, f_t$$

Assumption 2a is weaker than Assumption 2 as the former allows potential confounders to vary over both time and units as long as they can be approximated by the product of two low-rank matrices $\mathbf{\Lambda}$ and $\mathbf{F}$). However, both assumptions rule out the possibility that past

[4]The *no carryover effect* assumption can be dropped in the generalized difference-in-differences case where a unit remains treated after first exposed to the treatment. The model can be extended to include LDVs; see more details below.

outcomes directly affect current treatment. In the rest of the paper, we proceed with the premise that Assumptions 1 and 2 (or 1a and 2a) are satisfied.[5]

Estimands. The primary causal quantity of interest is the average treatment effect on the treated (ATT):

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

in which $\mathcal{T} \coloneqq \{i \mid \exists t, t' \text{ s.t. } D_{it} = 0, D_{it'} = 1\}$. More precisely, the treatment group $\mathcal{T}$ is defined as units who treatment status has changed during the observed time window. For units that have never been exposed to the treatment condition, it is difficult to compute their treated potential outcomes without strong structural assumptions.[6] Consequently, the estimated effect may not be representative of the average treatment effect on the population of interest, especially when only a few units experience the change in treatment status. Often times, we are also interested in the average treatment effect on the treated in a period that is s (s being a positive integer) 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 that we introduce later in the paper, we define $ATT_s = 0$ for any $s \leq 0$.

2.2. Estimators

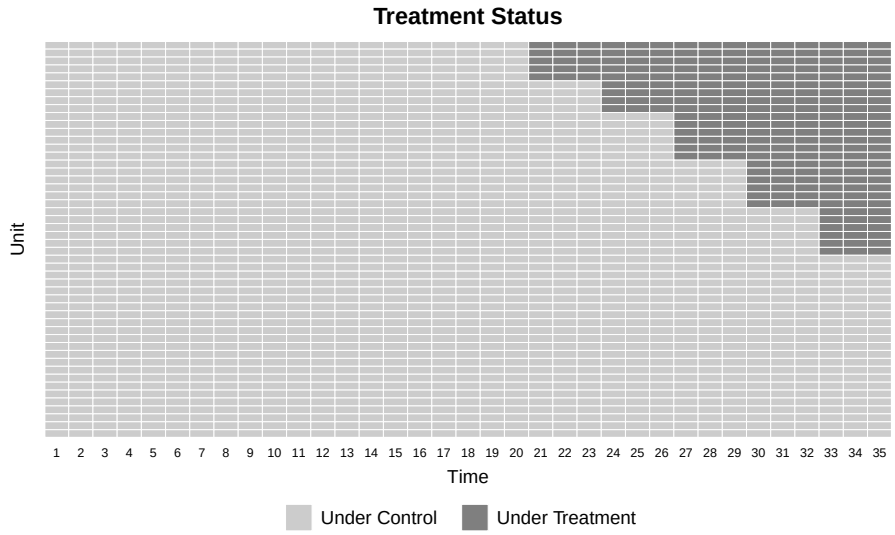
We now introduce three counterfactual estimators for the ATT (or ATT_s). They are conceptually similar because all of them construct a counterfactual for each observation under the treatment condition, $\hat{Y}_{it,D_{it}=1}(0)$. To be more concrete, consider a simulated panel dataset of 200 units and 35 time periods. It follows a DGP of an interactive fixed effect model with

[5] A directed acyclic graph (DAG) for the counterfactual estimators are depicted in the Supplementary Materials.

[6] This quantity of interest is also called the Average Treatment Effect on the Changed (ATC) in the 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.

a "staggered adoption" treatment assignment mechanism. We will discuss this DGP with full details in the next section. Figure 1 plot the treatment status of the first 40 units of this sample, with observations under the treatment and control conditions painted in dark and light gray, respectively. In the first step, the counterfactual estimators treat data under

FIGURE 1. DATA STRUCTURE OF A SIMULATED EXAMPLE



Note: The above figure shows the treatment status of the first 40 units of a simulated panel dataset of 200 units and 35 time periods. Dark and light gray cells represent observations under the treatment and control conditions, receptively.

the treatment condition as missing and use data under the control condition (light gray cells in Figure 1) to build a model for non-treated potential outcome $Y_{it}(0)$. They then predict $Y_{it}(0)$ for the observations under the treatment conditions (dark gray cells). ATT (or ATT_s) is obtained by taking an average of the differences $Y_{it}(1)$ and $\hat{Y}_{it}(0)$ for the treated observations. The following three estimators, in the order of computational complexity, differ in the modeling assumption based on which of $\hat{Y}_{it}(0)$ is estimated.

a) Fixed-effect counterfactual estimator. We start by introducing the *two-way fixed-effect counterfactual (FEct)* estimator. It has the advantage of sharing the same assumptions with conventional two-way fixed effects models, i.e., Assumptions 1 and 2, while relaxes the constant treatment effect assumption (with two-way fixed effects, $\delta_{it} = \delta$ for all i and t). The estimating strategy takes three steps. First, we estimate a two-way fixed effect model

using these non-treated observations only. Second, we predict treated counterfactuals using coefficients estimated from the first step. Finally, we obtain the ATT and ATT_s estimates. The algorithm is briefly summarized as follows:

Step 1. Estimate a two-way fixed effect model using non-treated observations only:

$$Y_{it}(0) = X_{it}'\beta + \mu + \alpha_i + \xi_t + \varepsilon_{it},\ \forall i, t, D_{it} = 0$$
$$\textstyle\sum_i \alpha_i = 0,\ \sum_t \xi_t = 0$$

obtaining $\hat{\mu}$, $\hat{\beta}$, $\hat{\alpha}_i$ and $\hat{\xi}_t$. Two linear constraints over the fixed effects are imposed to achieve identification.

Step 2. Estimate treated counterfactuals, obtaining:

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

Step 3. Obtain the ATT and ATT_s:

$$\hat{\delta}_{it} = Y_{it} - \hat{Y}_{it}(0)$$
$$\widehat{ATT} = \frac{1}{\sum_{\forall i,t} D_{it}} \sum_{D_{it}=1} \hat{\delta}_{it}, \qquad \widehat{ATT}_s = \frac{1}{|\mathcal{S}|} \sum_{(i,t)\in\mathcal{S}} \hat{\delta}_{it}, \quad s = 1, 2, \cdots$$
$$\text{in which } \mathcal{S} = \{(i,t) | D_{i,t-s} = 0, D_{i,t-s+1} = D_{i,t-s+2} = \cdots = D_{it} = 1\}$$

$|\mathcal{S}|$ represents the number of elements in set $\mathcal{S}$. Compared with conventional two-way FE models, *FEct* produces unbiased and consistent estimates for the ATT and ATT_t under Assumptions 1 and 2—bearing in mind that Assumption 2 implies the absence of unobserved time-varying confounders. It is robust when the treatment effect δ_{it} is heterogeneous. Note that δ_{it}'s are not individually identifiable while their averages are.

Proposition 1 (Unbiasedness and Consistency of FEct) *: Under Assumptions 1 and 2 as well as some regularity conditions,*

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

(See Supplementary Materials for the proof.)

When there exists no covariates, *FEct* can also be written as a matching estimator, that is, each treated observation is matched with its predicted counterfactual $\hat{Y}_{it}(0)$, which is a weighted sum of observations under the control status. Specifically, we have the following proposition:

Proposition 2 (FEct as a matching estimator) *: Under Assumptions 1 and 2 and when there exists 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 $\hat{Y}_{it}(0) = \mathbf{W}\mathbf{Y}_{D_{it}=0}$ *is a weighted average of all the non-treated observations.*

The specific form of weighting matrix $\mathbf{W}$, as well as the proof, is given in the Supplementary Materials.

Remark 1: DiD is a special case of FEct. It is easy to see that when no covariates exist and there are only two periods and two groups, one of which receives a treatment in the second period, the FEct estimator is the DiD estimator.

Remark 2: FEct and weighted Fixed Effects (wFE). Chernozhukov et al. (2013) show that the conventional two-way fixed effects model, or so called *within* estimator, is biased and inconsistent for the ATT when the treatment effect is heterogeneous.[7] To address this issue, Imai and Kim (2018) propose a weighted fixed effect (wFE) estimator that matches each treated (control) observation with an average of the control (treated) observation within the same unit while controlling for time effects either parametrically or by matching on control (treated) observations observed at the same time period. It is easy to see that FEct and wFE are mathematically equivalent when only unit fixed effects (but not time fixed effects) exist. As we will see in the next section, FEct also makes it easier to conduct additional diagnostic tests for the assumption of no time-varying confounders.

[7] Aronow and Samii (2016) establish similar results for cross-sectional data. It is well known that OLS weights each observation based on the treatment's conditional variance on additional covariates (Angrist and Pischke 2009). In other words, regression coefficients normally do not reveal the ATE, but gives a conditional-variance-weighted ATE.

Remark 3: Including lagged dependent variables. Although Assumption 2 rules out lagged dependent variables, in many applications, the bias induced by including lagged dependent variables with fixed effects is small when T is large while the improvement in precision and reduction in biases can be substantial (Beck and Katz 2011; Plmper and Troeger 2019). FEct can easily incorporate lagged dependent variables just as in fixed effects models.

b) Interactive fixed-effect counterfactual (IFEct) estimator. FEct will leads to biased estimates when unobserved time-varying confounders exist. A couple of authors have proposed to use factor-augment models to address this issue when the confounders can be decomposed into time-specific factor interacted with unit-specific factor loadings (Bai 2009; Gobillon and Magnac 2016; Xu 2017). Different from FEct, the IFEct estimator estimates a factor-augmented model instead of a two-way fixed effect model using observation under the control condition. A sketch of the algorithm is as follows:

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 non-treated ($D_{it} = 0$):

Step 2. Update $\hat{\beta}^{(h+1)}$, $\hat{\mu}^{(h)}$, $\hat{\alpha}_i^{(h)}$, $\hat{\xi}_t^{(h)}$, $\hat{\lambda}_i^{(h)}$, and $\hat{f}_t^{(h)}$ using an Expectation-Maximization (EM) algorithm, with treated counterfactuals being the missing values.

Step 3. Estimate treated counterfactual, obtaining:

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

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

When the model is correctly specified, the algorithm will produces consistent estimates for the ATT and ATT_t. The exact algorithm is in Supplementary Materials A.2.1. Again, we use a block bootstrap procedure to obtain the uncertainty estimates.

Proposition 3 (Consistency of IFEct) *: Under Assumptions 1a and 2a as well as some*

regularity conditions,

$$\widehat{ATT} \xrightarrow{p} ATT \ \ as \ N, T \to \infty.$$

(See Supplementary Materials for the proof.)

Remark 1: choosing the number of factors r. In order to choose r, we repeat Step 2 on a subset of non-treated observations until $\hat{\beta}$ converges. The optimal r is then chosen based on model performance measured by mean-squared prediction error (MSPE) using a k-fold cross-validation scheme. To preserve temporal correlations in the data, the test set consists of a number of triplets of three consecutive non-treated observations from one unit in the treatment group.[8] When $r = 0$, IFEct reduces to FEct.

Remark 2: related literature. The IFEct estimator is first proposed by Gobillon and Magnac (2016) in a DID setting where the treatment take places at the same time for a subset of units. It is also closely related to the generalized synthetic control method proposed by Xu (2017), in which factors are estimated using only the control group data, i.e., units that are never exposed to the treatment, to allow speedy cross-validation, achieve fast convergence, and prevent estimated latent factors from changing due to inclusion of additional treated units. 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) during the observed time window. Hence, the generalized synthetic control method estimator can be seen as a special case of IFEct when the treatment does not switch back.

[8] A recent study has shown that the cross-validation approach tends to overestimate the true r in small samples (Li 2018). The author introduces a new PC criterion that attempts to balance the model fitness and the impact of a small number of treated units on the test statistic:

$$PC(r) = \frac{1}{\sum \mathbf{1}\{D_{it}=0\}} \sum_{D_{it}=0} (Y_{it} - \hat{\lambda}_i' \hat{f}_t)^2 + r\hat{\sigma}c \left(\frac{N+T}{\sum \mathbf{1}\{D_{it}=0\}} \right) ln \left(\frac{N+T}{\sum \mathbf{1}\{D_{it}=0\}} \right),$$

where $\hat{\sigma}$ is an estimate of the variance of idiosyncratic error and c is a penalty factor decided by N and T. $c \to 1$ as N and T approaches to infinity. Therefore, the PC criterion penalizes large r more severely when the sample size is small. In our simulation exercises, the MSPE criterion seems to perform better than the PC criterion in most cases.

c) 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 the FEct and IFEct estimators, 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, it also builds upon an underlying interactive fixed effects structure. It assumes that the non-treated potential outcome matrix $\mathbf{Y(0)}_{(N\times T)} = [Y_{it}(0)]_{i=1,2,...,N,t=1,2,...,T}$ can be approximated by $\mathbf{L}_{(N\times T)} = [L_{it}]_{i=1,2,...,N,t=1,2,...,T}$ (we omit covariates and additive fixed effects for simplicity):

$$\mathbf{Y(0)} = \mathbf{L} + \varepsilon,\ \mathbb{E}[\varepsilon|\mathbf{L}] = \mathbf{0},$$

in which ε 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: $\mathbf{L} = \mathbf{\Lambda F}$. Different from IFEct, however, instead of estimating factors $\mathbf{F}$ and factor loadings $\mathbf{\Lambda}$ separately, the MC estimator seek to directly estimate $\mathbf{L}$ by solving the following minimization problem:

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

in which $\mathcal{O} = \{(i,t)|D_{it} = 0\}$ and $|\mathcal{O}|$ is the number of elements in $\mathcal{O}$. $\|\mathbf{L}\|$ is the chosen matrix norm of $\mathbf{L}$ and θ is a tunning 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 below. First, define $P_{\mathcal{O}}(\mathbf{A})$ and $P^{\perp}_{\mathcal{O}}(\mathbf{A})$ for any matrix $\mathbf{A}$:

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

Conduct Singular Value Decomposition (SVD) on matrix $\mathbf{A}$ and obtain $\mathbf{A} = \mathbf{S\Sigma R^T}$. The matrix shrinkage operator is defined as $\text{shrink}_{\theta}(\mathbf{A}) = \mathbf{S\tilde{\Sigma}R^T}$, where $\tilde{\mathbf{\Sigma}}$ equals to $\mathbf{\Sigma}$ with the i-th singular value $\sigma_i(A)$ replaced by either (1) $\max(\sigma_i(A) - \theta, 0)$ or (2) $\sigma_i(A)\mathbf{1}\{\sigma_i(A) \geq \theta\}$.

The former form is called soft impute and the latter, hard impute. Below is an illustration adapted from Athey et al. (2018), in which hard compute (left) selects two factors:

Hard Impute ("best subset"/IFEct)

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

Soft Impute

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

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

A sketch of the iterative algorithm is describe as follows:

Step 0. Given a tunning 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.

Remark 1: Implementation. In practice, we use k-fold cross-validation to select the θ that minimizes the MSPE. The test set is constructed in the same way as in IFEct. We also allow covariates and additive fixed effects and use the partialing-out method in the iterative process. The uncertainty estimates are obtain through a block bootstrap procedure.

Remark 2: Relationship with IFEct (and FEct). In essence, hard impute is equivalent to IFEct as both algorithms penalize on the number of factors.[9] Therefore, in the rest of the paper we refer to the soft impute method as MC. When θ is larger than biggest eigenvalue of the error matrix after the influences of covariates and additive fixed effects are eliminated, MC is reduced to FEct.

[9]MC with hard impute and IFEct differ only in initial values, for example, Mazumder, Hastie and Tibshirani (2010) recommend to use soft impute estimates as the initial value for hard impute. The difference in final results is often negligible.

Compared to FEct or conventional two-way fixed effects models, both IFEct and MC use lower-rank matrix approximation to make better predictions for treated counterfactuals. They differ in the way of regularization, that is, IFEct chooses the number of factors while MC relies on a tunning parameter θ. Whether IFEct or MC performs better depends on the context. We provide Monte Carlo evidence in Section 4 and show that when when the factors are strong and sparse, IFEct out-performs MC. In practice, researchers may choose between the two models either based on (1) how they behave under diagnostic tests introduced in the next section and/or (2) their relative predictive power (e.g., as measured by MSPE).

3. Diagnostics

In this section, we introduce a set of diagnostic tools to assist researchers to probe the validity of Assumption 2 (or 2a). We first introduce a plot for dynamic treatment effects based on the counterfactual estimators. We then propose three statistical tests for the presence of time-varying confounders.

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 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 eyeballing whether the coefficients in the pre-treatment periods (when $s \leq 0$) exhibit a trend or are statistically significant from zero.[10]

We improve the dynamic treatment effect plot by taking advantage of the previously

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

discussed counterfactual estimators. Instead of plotting of the interaction terms, we plot the averages of 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 $t, i \in \mathcal{T}$. Therefore,

$$\begin{aligned} \widehat{ATT}_s &= \frac{1}{|\mathcal{S}|} \sum_{(i,t)\in\mathcal{S}} \hat{\delta}_{it}, \\ \text{and } \mathcal{S} &= \{(i,t) | D_{i,t-s} = 0, D_{i,t-s+1} = D_{i,t-s+2} = \cdots = D_{it} = 1\} \\ s &= m, m+1, \cdots, 0, 1, 2, \cdots n \end{aligned} \tag{3}$$

in which $m < 0$ and $n \geq 1$. When Assumption 2 (or 2a) is correct, it is easy to see that average pre-treatment residuals converge to zero, i.e., $\widehat{ATT}_{s\leq 0} \xrightarrow{p} 0$. Therefore, we should expect pre-treatment estimates to be bouncing around 0, hence, no strong "pre-trend." This method has two primary advantages over the traditional method. First, it relaxes the constant treatment effect assumption—even though the conventional dynamic treatment effect plot allow the treatment effects to be different over time (relative to the onset of the treatment), it assumes they are same for each treated unit in a given time period. Second, because a unit's non-treated average have already been subtracted from $\hat{\delta}_{it}$, it is no longer necessary for researcher to arbitrarily choose a base category; to put differently, the base category is set at a unit's non-treated average after time effects are partialed out.

We illustrate the dynamic treatment effects plot using a simulated panel dataset of 200 units and 35 time periods based on the following outcome model:

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

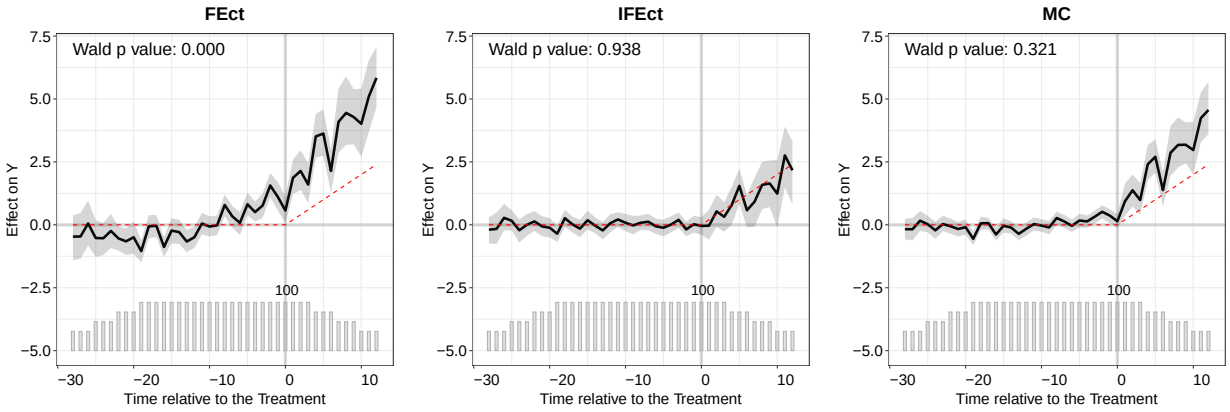
This simulated dataset has a generalized DiD structure with staggered adoption: once a unit adopts the treatment, it remains treated in the remainder of all treatment periods. Following Athey and Imbens (2018), each unit is assigned a number of pre-treatment periods $T_{0i} \in \{20, 23, 26, 29, 32, 35\}$ such that $D_{it} = 0$ if $1 \leq t \leq T_{0i}$ and $D_{it} = 1$ if $t > T_{0i}$. The 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 they are not accounted for.

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 effect 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 1a and 2a.

Figure 2 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 are depicted with red dashed lines. From the left panel of Figure 2, 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 pre-treatment periods are significantly different from 0; and (2) there are sizable positive bias in the ATT estimates in the post-treatment periods. We see a similar patten in the post-treatment periods from the right panel where the MC estimator is applied, though with smaller biases. However, when using the IFEct estimator, the ATT estimates in

FIGURE 2. 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 histogram 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.

both pre- and post-treatment 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 is able to pick the correct number of factors.

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 statistics tests that formally test the implications of this assumption.

3.2. Two Statistical Tests

To test the presence of potential time-varying confounders, we propose two statistical tests. A natural approach is to jointly test a set of null hypotheses that the average of residuals for any pre-treatment period is zero, i.e., $ATT_s = 0$ for any $s \leq 0$. We can construct the following variant of a Wald statistic (F statistic):

$$F = \frac{\sum_{i \in \mathcal{T}} \sum_{s=m}^{0} \left(\hat{e}_{is}^2 - (\hat{e}_{is} - \widehat{ATT}_s)^2 \right) / (1-m)}{\sum_{i \in \mathcal{T}} \sum_{t=1}^{T_0} (\hat{e}_{is} - \widehat{ATT}_s)^2 / (|\mathcal{O}_{\mathcal{T}}| - m + 1)}$$

in which $\mathcal{O}_{\mathcal{T}} = \{(i,t) \mid D_{it} = 0, i \in \mathcal{T}\}$ and $(1-m)$ is the total number of pre-treatment

periods ($m < 0$).[11]

However, as pointed out by Hartman and Hidalgo (2018), there are two potential problems with such a test if our goal is to provide evidence for zero residual means in the pre-treatment periods—a type of equivalence. The first issue is the lack of power, that is, when the number of observations is limited, failing the reject the null of joint zero means does not mean equivalence holds. "The absence of evidence is not the evidence of absence." This is arguably less of a concern for us because we need both T and N for the model to converge in the first place, at least in the case of IFEct and MC. Second, when the sample size is large, a small confounder (or a few outliers) which only contributes to a neglectable amount of bias in the causal estimates will almost always cause rejection of the null of joint zero means. This is especially problematic for TSCS data analysis because the chances that there exist some non-decomposable but non-influential time-varying confounders are high.

An equivalence test. To address these concerns, we proposed a variant of the equivalence test proposed by Hartman and Hidalgo (2018) in the cross-sectional setting. The null hypothesis is reversed:

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

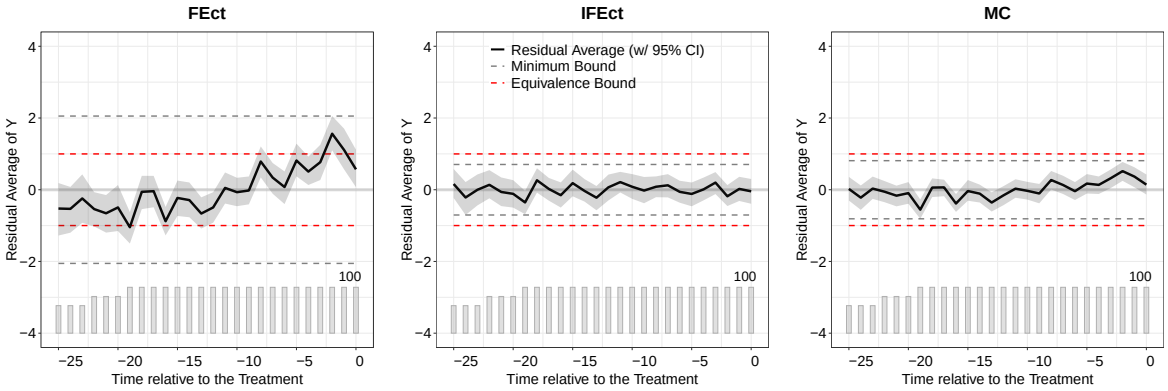
in which $-\theta_2 < 0 < \theta_1$ are pre-specified parameters. Rejection of the null hypothesis implies the opposite holds with a high probability, i.e., $-\theta_2 \leq ATT_s \leq \theta_1$ for any $s \leq 0$. In other words, if we collect sufficient data and show that the pre-treatment residual 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. $[-\theta_2, \theta_1]$ are therefore called the *equivalence bound*. The logic is the opposite to what behinds a Wald test. For a Wald test, as N grows, any slight deviation of ATT_s from zero will lead to statistically significance at a conventional threshold. With an equivalence test, however, a greater N makes it more likely to reject the null of non-equivalence and, at the same time, justify the identification assumption when it

[11] The details of the test is provided in the Supplementary Materials.

is valid. We will illustrate this point via simulations in the next section.

We use the two one-side 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 pre-treatment periods generate significant results.[12] Following the suggestion in 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 non-treated outcome.[13] Given this choice, each TOST fails (i.e. equivalence holds) when the bootstrapped one-side confidence interval of ATT_s falls within $[0.6\hat{\sigma}_\varepsilon, 0.6\hat{\sigma}_\varepsilon]$, the equivalence bound. In addition, we also calculate the minimum bound, the smallest symmetric bound below which we cannot reject the null under the equivalence test. In other words, the minimal bound is determined by largest absolute value of the range of confidence intervals of $\widehat{ATT}_{s,s\leq 0}$ in the pre-treatment periods. A rule of thumb is that when the minimum bound is within the range of the equivalence bound, the test is considered passed.

FIGURE 3. EQUIVALENCE TESTS FOR THE SIMULATED EXAMPLE



Note: The above figures show the results of the equivalence tests based on three different estimators: FEct, IFEct, and MC. The histogram 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 mark the equivalence bound while the gray dashed line mark the minimum bound.

Figure 3 shows the results of the equivalence based on the three estimators: FEct, IFEct, and MC, using the simulated dataset. With FEct, the trend leading towards the onset of

[12] This is clearly a conservative standard as ATT_s should be correlated with each other thus the probability to commit Type I error is actually smaller than the threshold (say, 0.05) we select.

[13] Specifically, we run a two-way fixed effect model with time-varying covariates using non-treated data only and calculate the standard deviation of the residuals.

the treatment goes beyond the equivalence bound and results in a wide minimum bound. Therefore, we cannot reject the null of the pre-treatment residual averages are not zero—in other words, we cannot say that equivalence holds with a high probability. However, both IFEct and MC pass the test. The confidence intervals of pre-treatment residual averages are within the equivalence bounds and the minimum bounds are narrow.

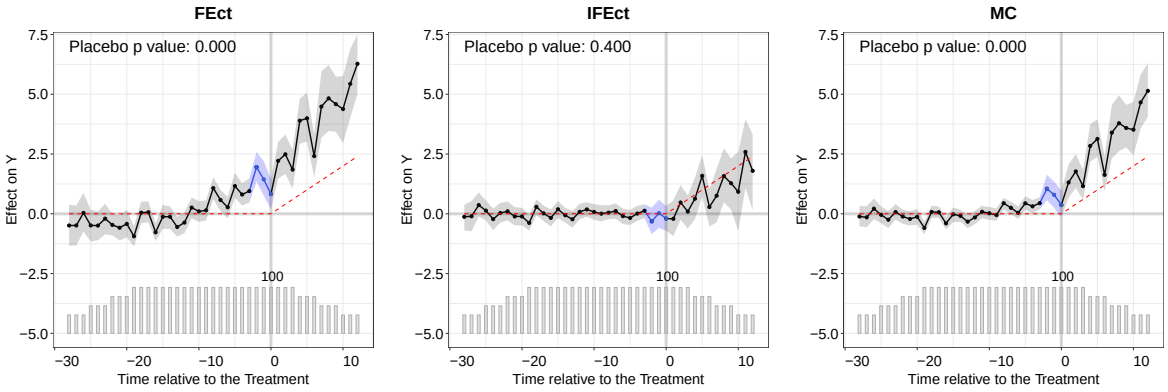
The main shortcoming of the equivalence test is that the choice of the equivalence is arbitrary. Because we usually do not know the DGP, it is difficult to establish the relationship between the amount of imbalance in the pre-treatment periods and bias in the ATT estimates. $0.36\sigma_\varepsilon$ may also be too lenient when the effect size is small relative to the variance of the outcome. Another approach suggested in the literature is to benchmark the minimum bound against a reasonable guess of the magnitude of the effect of interest *a prior* based on previous studies (e.g., Wiens 2001). 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 bound in a *post-hoc* fashion can lead to problematic results (Campbell and Gustafson 2018). It is recommended that researchers pre-register a plausible effect size and use it to set the equivalence bound 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 the model (IFEct or MC) fits the non-treated data better and better, the variance of the residuals becomes smaller and smaller, which makes it easier to pass the equivalence test. To guard against such risks, next, we introduce an (out-of-sample) placebo test.

A placebo test. The idea of 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 pre-treatment periods. If

no time-varying confounder assumption holds, this ATT estimate should be statistically indistinguishable from zero and an insignificant result will be in favor of the assumption's validity. In practice, S should be not set too large because the larger S is, the fewer pre-treatment period will remain to be used in estimation. If both S and N_{tr} are too small, however, the test may be under-powered. In this and following examples, we set $S = 3$. We find that if there exists a time-varying confounder that strongly biases the ATT estimates, it will be captured by the placebo test. An important property of the proposed placebo test is that it is robust to model mis-specification and immune from over-fitting because it relies on out-of-sample predictions of $Y(0)$ during the placebo periods.

FIGURE 4. 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 histogram 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 pre-treatment periods ($s = -2, -1, 0$) serving as the placebo are painted in blue. The p-value of the placebo test is shown at the top-left corner of each figure.

Figure 4 shows the results from the placebo tests based on the three counterfactual estimators: FEct, IFEct, and MC. We see that both FEct (without including any latent factors) and MC (approximating the outcome matrix using soft-impute) fail the placebo test while IFEct passes the test. Two patterns are worth-noting. First, the confidence intervals of the ATE estimates in the post-treatment periods are slightly wider than those in Figure 4. This is because fewer data points of the treated units are being used to estimate the latent factor model and they contribute to the precision of the ATT estimates. Second, although the MC method fits the pre-treatment periods well, it fails the test due to model-misspecification,

which leads to biases in the causal estimates. This finding confirms that achieving better model fitness in the pre-treatment periods does not guarantee reduced biases; in fact, model mis-specification and over-fitting can make things much worse. Hence, an equivalence test may be too easy to pass for IFEct and MC as pre-treatment data are used to fit the model. A placebo test complements the equivalence test and safeguards against such risks.

Neither of the proposed tests is perfect. The equivalence test may suffer from over-fitting and model-misspecification. On the other hand, the the placebo test may have a power issue. It also ignores unobservable factors that appear only periodically. For example, if a shock occurs every 5 periods, our model may pass the placebo test with $S = 3$ but still return a biased estimate. Therefore, we recommend researchers use the two tests together and justify the identifying assumption when it is supported by both.

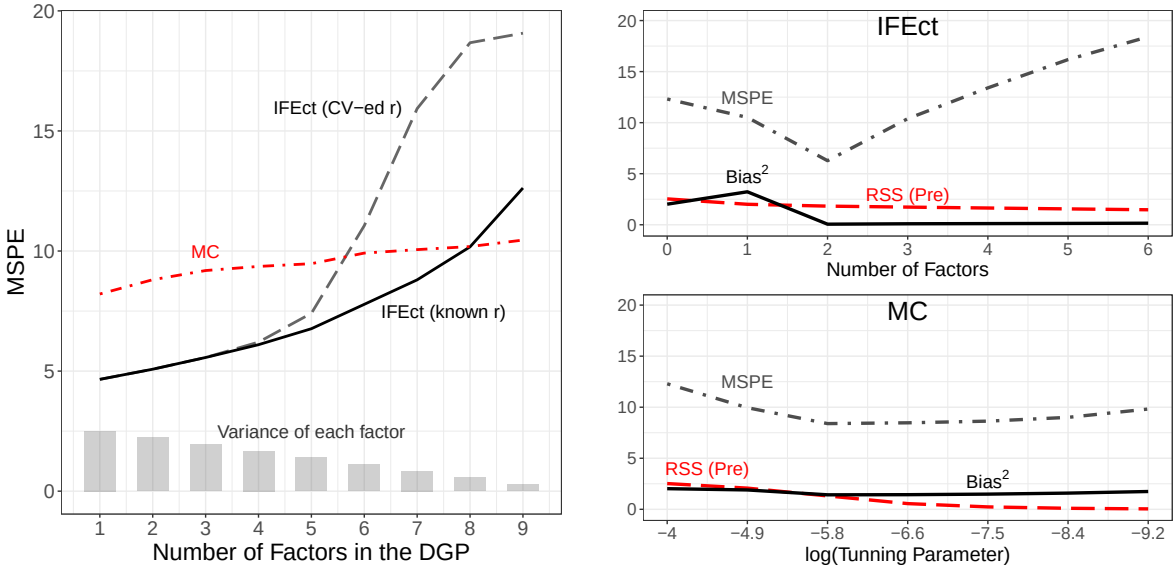
4. Monte Carlo Evidence

In this section, we seek to answer two questions using Monte Carlo exercises. First, we want to understand under what circumstances the MC estimator outperforms IFEct and vice versa. Second, we want to know what the main advantages are of using an equivalence test instead of a more intuitive Wald test in a TSCS setting.

IFEct versus MC. We first compare the performance of the IFEct and MC estimators using DGPs similar to what is specified in the previous section: $Y_{it} = \delta_{it} D_{it} + 5 + \frac{1}{\sqrt{r}} \sum_{m=1}^{r} \lambda_{im} \cdot f_{mt} + \alpha_i + \xi_t + \varepsilon_{it}$. 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 out-performs 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 5. MONTE CARLO EXERCISES: IFECT VS. MC



(a) Comparing Estimators (b) DGP with Two Factors

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. Figure (b) compares the biases (squared), MSPEs for treated counterfactuals, and residual sum of squares (RSS) in the pre-treatment periods of IFEct and MC using different tuning parameters when the DGP is fixed with two factors.

The results are shown in Figure 5(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 1000 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 turning parameter λ. The result shows that MC gradually catch 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 more sparsely distributed, 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 they are not directly estimated.

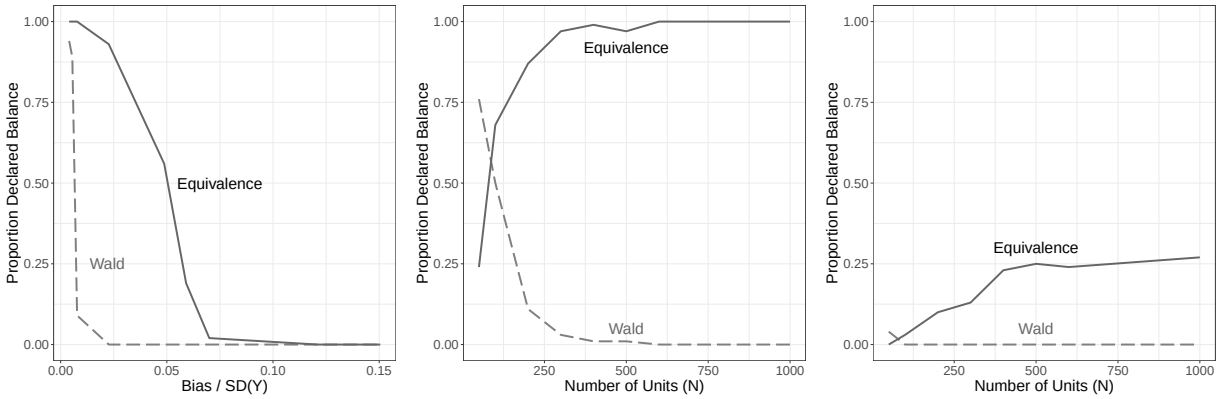
In Figure 5(b), we fix the DGP with two factors and compare the performances of IFEct

and MC with different tuning parameters. The black solid lines, gray dot-dashed lines, and red long-dashed lines represent (1) the squared biases of the estimated ATT, i.e, $[\mathbb{E}(\widehat{ATT} - ATT)]^2$, (2) MSPE for treated counterfactuals, and (3) residual sum of squares (RSS) in the pre-treatment periods, respectively. We observe several patterns. First, as expected, with the most favorable tuning parameters, IFEct performs better than MC in terms of both MSPE and bias because the DGP follows an IFE model exactly. 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 RSS keeps diminishing while MSPE keeps increasing. Third, the MSPE of IFEct increases faster than that of MC as model complexity passes the optimal level while biases stay low for both models. These findings strongly suggest that the RSS is poor indicator of model performance and a high level of model fitness in the pre-treatment periods does not necessarily lead to more precise estimates of the ATT.

Wald test versus the equivalence test. As explained in the previous section, we prefer the equivalence test to a conventional Wald test mainly because the former can tolerate potential confounders that only result in a small amount of bias in the causal estimates. To illustrate this point, 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 1000 simulations. In each simulation, we generate a sample of 200 units (100 treated and 100 controls) of 40 periods. The probability of receiving the treatment is positively correlated with factor loadings. In Figure 6(a), we plot the proportion of times the equivalence test (black solid line) or the Wald test (gray dashed line) backs the no-time-varying-confounder assumption against the normalized bias induced by the confounder when no factors are directly being controlled for. We see that, with a relatively small sample size, as long as a confounder exists, the Wald test almost always suggests the failure of the assumption—which is not technically wrong—while the

equivalence test exhibit a relatively higher level of tolerance of biases. The probability of rejecting inequivalence (hence, declaring equivalence) declines smoothly as the bias increases, which is similar what Hartman and Hidalgo (2018) report in a cross-sectional setting. For example, the identification assumption will still be considered valid with a 80% probability if the confounder leads to a bias of 2.5% standard deviation of Y in the ATT. This is important because for most observational TSCS data, potential confounders exist; the question is to what extent they will significantly change the causal estimates of interest. Figure 6(b) and

FIGURE 6. MONTE CARLO EXERCISES: WALD VS. EQUIVALENCE



(a) Varying Bias (b) Small Bias Case (c) Large Bias Case

Note: The above figures show the results from Monte Carlo exercises that compare the Wald test and the equivalence test when an unobserved confounder exists. Figure (a) compares the performance of the Wald test and the equivalence in the presence a time-varying confounder when we vary the amount of bias in the ATT induced by the confounder. With a relatively small sample size ($T = 30$, $N = 200$), the Wald test has a much smaller level of tolerance of bias than the equivalence tests. Figures (b) and (c) show the performances of both test in two scenarios as N increases: when the bias is very small ($0.01\sigma_Y$) and when the bias is relatively large ($0.07\sigma_Y$).

(c) show how the performances of the two tests change as the number of units N increases under two circumstances: when the bias induced by the confounder is neglectable ($0.01\sigma_Y$) and when the bias is relative large ($0.07\sigma_Y$). The results suggest that in both scenarios, the Wald test quickly rejects the null (hence, declaring inequivalence) as the sample size grows while the equivalence test behaves differently: when the bias is small, the probability of declaring equivalence quickly grows with sample size and stays high; when the bias is relatively large, the probability of declaring equivalence firstly increases with the sample

size, then stays at a low level. The equivalence test has more desirable properties because it tolerates small biases while being capable of detecting large ones.

5. Empirical Examples

Finally, we apply the counterfactual estimators, as well as the proposed diagnostic tests, to three empirical example in political economy. For any empirical study, we recommend researchers to start with the simplest estimator, FEct. If the results from FEct pass both the equivalence and placebo tests as well as a visual inspection, there is little need for more complex methods (except for efficiency gains under some rare circumstances). If, however, one of the tests fails, which suggests the failure of Assumption 2, researcher can turn to IFEct or MC (or both) and conduct diagnostic tests again. In all three applications, we set $S = 3$ in the placebo testes. All uncertainty estimates are obtained based on block bootstraps at the unit level for 1,000 times.

Hainmueller and Hangartner (2015). Our first example comes from Hainmueller and Hangartner (2015), who study whether minority immigrants obtain a higher level of naturalization rate under indirect democracy than under direct democracy in Swiss municipalities using a two-way fixed effects design. The outcome variable is minorities' naturalization rate in municipality i during year t. The treatment is a dummy variable indicating whether the decision of naturalization is made by elected municipality councils rather than citizens in popular referendums. The dataset includes 1,211 Swiss municipalities over 19 years, from 1991 to 2009. The authors report that naturalization rate increase by over 1% on average after a municipality shifts the decision making power from popular referendums to elected officials. The result is replicated in column 1 of Panel (A) in Figure 1.

We then apply the FEct estimator and obtain an estimate of 1.767 with a standard error 0 0.192 (column 2), even larger than the original estimate. Plots for the dynamic treatment effects and results of the two diagnostic tests are shown in Figure 7. We see that (1) the

Table 1. Results from Three Empirical Examples

Panel A: Hainmueller and Hangartner (2015)

Outcome: Naturalization Rate (%)	Two-way	FEct	IFEct (FEct)	MC (FEct)
	(1)	(2)	(3)	(4)
Councilor (vs. Referendums)				
Coefficient	1.339	1.767	1.767	1.767
Standard Error	(0.125)	(0.199)	(0.199)	(0.199)
95% Confidence Interval	[1.093, 1.584]	[1.389, 2.176]	[1.389, 2.176]	[1.389, 2.176]
Unit & Period FEs	Yes	Yes	Yes	Yes
r (IFEct) / λ_L (MC)	N/A	N/A	0	$> \sigma_1$
Equivalence Test	N/A	Pass	Pass	Pass
Placebo Test	N/A	Pass	Pass	Pass
MSPE	N/A	21.7	21.7	21.7

Panel B: Xu (2017)

Outcome: Turnout (%)	Two-way	FEct	IFEct	MC
	(1)	(2)	(3)	(4)
Election Day Registration				
Coefficient	0.778	1.425	3.998	2.970
Standard Error	(3.177)	(3.267)	(1.924)	(2.063)
95% Confidence Interval	[-5.459, 7.004]	[-5.525, 7.432]	[-0.472, 7.269]	[-1.289, 6.786]
Unit & Period FEs	Yes	Yes	Yes	Yes
r (IFEct) / λ_L (MC)	N/A	N/A	2	$0.07\sigma_1$
Equivalence Test	N/A	Fail	Pass	Pass
Placebo Test	N/A	Fail	Pass	Pass
MSPE	N/A	27.2	17.7	15.6

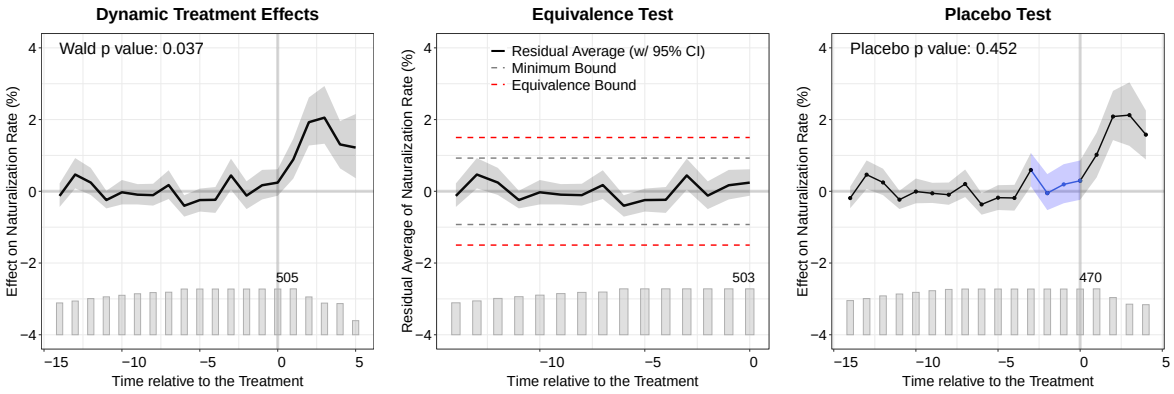
Panel C: Acemoglu et al. (2019)

Outcome: GDP (log)	Two-way	FEct	IFEct	MC
	(1)	(2)	(3)	(4)
Democracy				
Coefficient	-10.112	1.215	2.330	2.084
Standard Error	(4.315)	(6.517)	(5.457)	(5.515)
95% Confidence Interval	[-18.570, -1.655]	[-11.803,13.329]	[-8.611, 13.252]	[-9.727, 12.032]
Unit & Period FEs	Yes	Yes	Yes	Yes
r (IFEct) / λ_L (MC)	N/A	N/A	4	$0.07\sigma_1$
Equivalence Test	N/A	Fail	Pass	Pass
Placebo Test	N/A	Pass	Pass	Pass
MSPE	N/A	0.056	0.013	0.012

Notes: The uncertainty estimates are based on block bootstraps at the unit level (municipality, country dyad, and state, respectively) for 1,000 times.

residual averages in the pre-treatment periods are almost flat and around zero and the effect shows up right after the treatment begins; (2) the FEct estimates pass both the equivalence test—the minimum bound is way narrower than the equivalence bound—and (3) in the placebo test, we cannot reject the null of zero effect in the three pre-treatment periods at any reasonable level of statistical significance. All above evidence points to a valid design with the exception of the Wald test, which reports a p-value below 5%. As discussed earlier, this may be driven by noises in the data that are inconsequential to the causal estimates. We also apply both IFEct and MC estimators to this example. It turns out that the cross-

FIGURE 7. RESULTS FROM HAINMUELLER AND HANGARTNER (2015)



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 figure on the left presents the estimated dynamic treatment effects using FEct. The figure in the middle shows the result of the equivalence with red and gray dashed lines representing equivalence bound and minimum bound, respectively. The right figure shows the results from a placebo test using the "treatment" in three pre-treatment periods as a placebo. The histogram at the bottom of each figure illustrates the number of treated units at a given time period relative to the onset of the treatment.

validation schemes find 0 factors, in the case of IFEct, and the largest tunning parameter θ, in the case of MC, which also implies maximum regularization on the eigenvalues or 0 factors. Hence, both methods are reduced to FEct and give the exact same estimates.

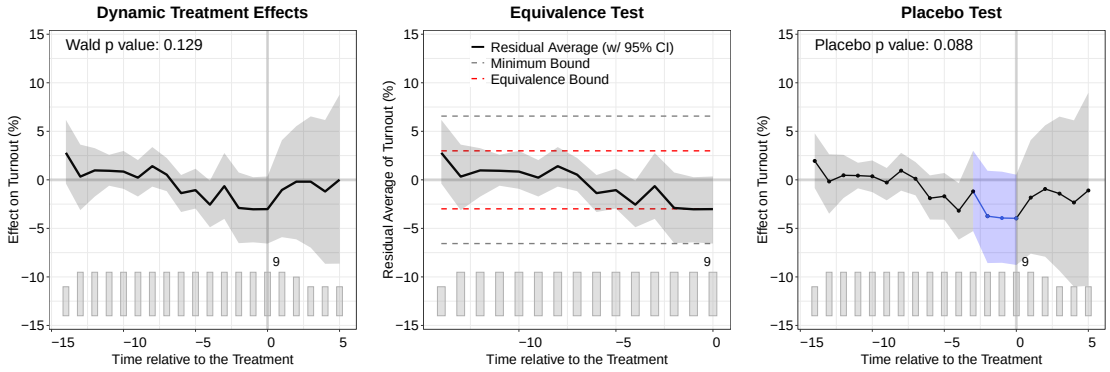
In short, results from FEct are substantively similar to those from conventional two-way fixed effects models. However, counterfactual estimators like FEct allow us to check the validity of the no-time-varying-confounder assumption in a more convenient and transparent fashion. We can see that in this example, the effect of indirect democracy on naturalization

rates rises gradually after the institutional change, which may bear important implications for policy makers.

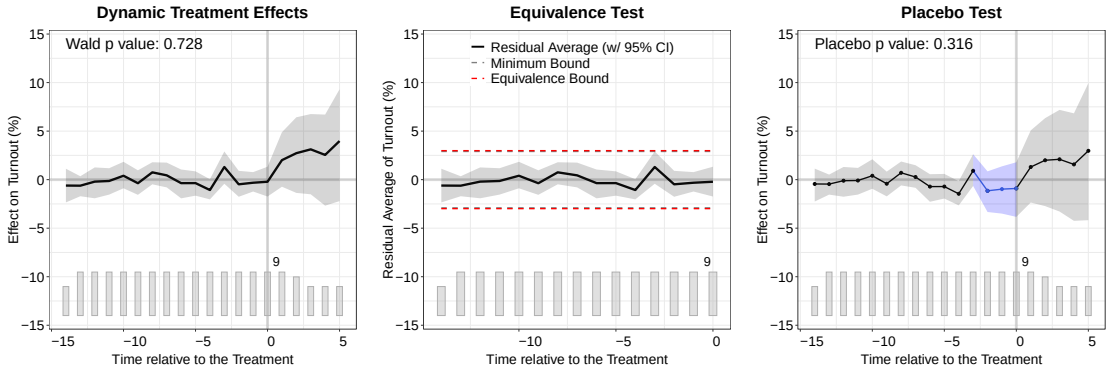
Xu (2017). The second example adapts from Xu (2017), in which the author illustrates the generalized synthetic control method by investigating the effect of Election Day Registration (EDR) on voter turnout in the United States. The unit of analysis is state by year. The dataset covers 24 presidential elections (1920–2012) for 47 states, 9 of which adopted EDR before 2012. As shown in columns 1 and 2 in Panel B of Table 1, both two-way fixed effects and FEct produce a null result—a small ATT estimate with a relatively large standard error, while IFEct and MC give an estimate of 3-4 percentage points increase brought by the EDR (smaller than the gysnth estimate of 4.9 percentage points). Given the relatively small sample size, the coefficients are not very precisely estimated, especially with MC. A visual check of Figure 8 shows that there exists a strong downward "pre-trend" leading to EDR, which may confound the relationship between EDR and turnout. Both IFEct and MC pass the equivalence test and the placebo. The example suggests that presence of time-varying confounder may also lead to significant downward bias and applying a factor-augmented approach can be fruitful in a TSCS setting even with relatively small datasets.

Acemoglu et al. (2019). The last example concerns the effect of democracy on economic growth using data from Acemoglu et al. (2019). The unit of analysis is country by year. The treatment is a dichotomous measures of democracy; and the outcome is the logarithm of GDP *per capita*. The dataset covers 184 countries in total, spanning from 1960 to 2010. The authors report a large and statistically significant using both dynamic panel models and panel inverse propensity score reweighting. We will mainly compare our finding with the result from the latter approach. Columns (1) in Panel C of Table 1 shows that a two-way fixed model give an estimate of -10.112 with a standard error of 4.315, a incredibly large negative effect. When we apply the FEct estimator, the estimated effect dramatically shrinks to 1.215 (column (2)), which indicates that the rigid modeling assumptions behind

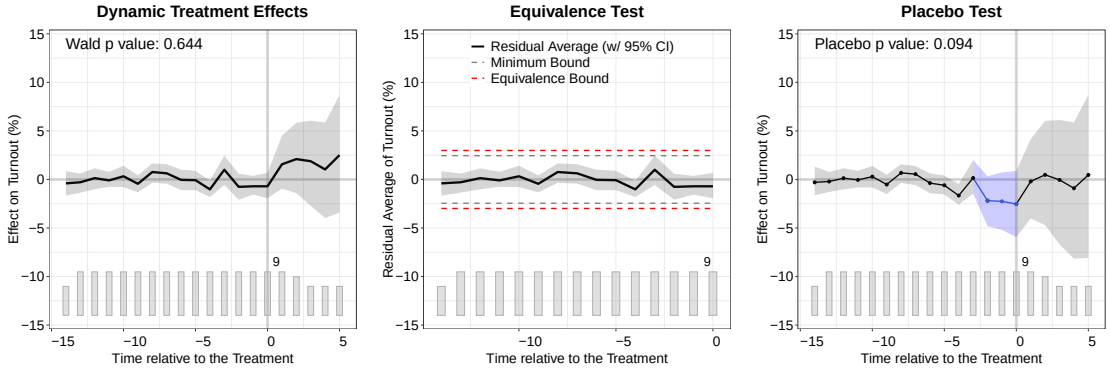
FIGURE 8. RESULTS FROM XU (2017)



(a) FEct



(b) IFEct

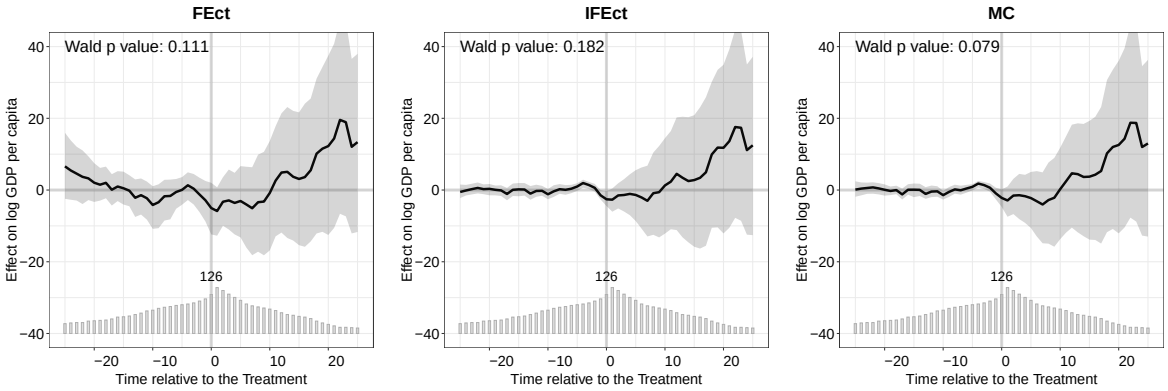


(c) MC

Note: The above figures show the results from applying the counterfactual estimators to data from Xu (2017), who investigates the effect of election day registration on voter turnout. Panels (a), (b) and (c) show the results using the FEct, IFEct, and MC estimators, respectively. The histogram 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 two-way fixed effects model potentially lead to significant bias. Figure 9(a) suggests that, using FEct, there exists a strong downward "pre-trend" leading towards transition to democracy, which may be due to the fact that many countries are "selected into" democracy when experiencing economic difficulties. As shown in the Supplementary Materials, FEct fails the equivalence test.

FIGURE 9. RESULTS FROM ACEMOGLU ET AL. (2019)



Note: The above figures show the results from applying the counterfactual estimators to data from Acemoglu et al. (2019), who investigate the effect of democracy on (log) GDP *per capita*. Figure (a), (b) and (c) show the results using the FEct, IFEct, and MC estimators, respectively. The histogram at the bottom of each figure illustrates the number of treated units at a given time period relative to the onset of the treatment.

To address this problem, we apply both the IFEct and MC estimators. Figures 9(b) and (c) show that, with both methods, the "pre-trend" mostly disappear. Both estimators give similar estimates for the effect of democracy on growth, which is substantively similar to what the authors obtain using an inverse propensity score reweighting approach, however, our bootstrapped standard errors are considerably larger.[14] It is worth noting that the standard errors estimated from IFEct and MC are smaller than that from FEct, suggesting that the loss of degree of freedom with more complex models is not the main reason behind the large uncertainties. As the histograms at the bottom of the figures suggest, the large point estimate of the long-run effect (e.g., 20 years from a democratic transition) is mainly driven by a very small number of (a dozen) countries.

[14]See Supplementary Materials for a comparison, as well as results for the equivalence tests and placebo tests.

6. Conclusion

This paper attempts improve practice in making causal inference with observational TSCS data. We acknowledge that two-way fixed effect models, one of the most commonly used methods to analyze TSCS data in the social sciences, requires demanding assumptions to have a causal interpretation for its estimates. However, instead of suggesting a complete abolishment of such models, we follow the literature's incremental approach and seek to make marginal but meaningful improvement upon existing fixed effects methods.

We focus on cases in which the treatment is dichotomous and introduce a group of counterfactual estimators for the average treatment effect on treated observations. They include the fixed effect counterfactual (FEct) estimator, the interactive fixed effect (IFEct) estimator and the matrix completion (MC) estimator, which rely on different functional form assumptions. IFEct and MC can be understood as generalizations of FEct and they both reduce to FEct when the regularization is sufficiently large. These estimators directly impute treated counterfactuals and then take averages of the differences between the observed outcome data and imputed counterfactuals. By doing so, these counterfactual estimators are consistent without making the homogeneous treatment effect assumption. It is important to note that these estimators are not this paper's invention. Our main contribution is to unify these estimators in a simple framework, which allows researchers to evaluate each model's respective assumptions, compare their performances, and make an informed decision of which one is the most suitable for their applications.

To do so, we propose a set of diagnostic tools to help researchers gauge the validity of the no-time-varying-confounder assumption. We improve an existing practice of estimating and plotting dynamic treatment effects and develop two statistical tests—an equivalence test and a placebo test—based on the new method. These tests generate intuitive visuals and are easy to interpret. They can also assist researcher to decide which model produces the most reliable causal estimates. Our recommendation is to use the simplest model (i.e., FEct

$\succ$ IFEct $\succeq$ MC) that passes both tests.

We illustrate the counterfactual estimators and diagnostic tests using three empirical examples from political economy. Using data from Hainmueller and Hangartner (2015), we find that FEct is appropriate and it generates substantively similar results to two-way fixed effects. In the other two examples (Xu 2017 and Acemoglu et al. 2019), however, the assumption of no time-varying confounders with FEct clearly fails. In Xu (2017), the results from IFEct and MC are qualitatively different from those using FEct or two-way fixed effects; in Acemoglu et al. (2019), all three methods point to a much smaller in magnitude, but more credible, estimate than that from two-way fixed effects. Theses examples suggest a sequential approach of apply the three estimators (first FEct, then IFE or MC) in empirical analysis. Both the estimators and tests can be easily implemented with an open source package, `fect`, in `R`.

Our methods have limitations. First, both IFEct and MC require both a large N and a large T; so do the diagnostic tests when they rest upon IFEct and MC estimators. When T is short compared with the underlying number of factors, the incidental parameter problem will lead to bias in the causal estimates, especially with IFEct. Second, the equivalence test requires users to pre-specify a bound, which may leave room for *post hoc* model justification. Last but not the least, it is worth re-iterating that strict exogeneity is the cost we have to pay when we control for time-invariant heterogeneity and common shocks, which may be unreasonable assumption. For example, the usual caveat of Nickell bias applies when past outcomes are taken as strictly exogenous covariates (Nickell 1981). However, it is our belief that the framework of counterfactual approach is general and can be extended to support a wide range of models for TSCS analysis.

References

Acemoglu, Daron, Suresh Naidu, Pascual Restrepo and James A. Robinson. 2019. "Democracy Does Cause Growth." *Journal of Political Economy* 121(1):47–100.

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

Aronow, Peter M and Cyrus Samii. 2016. "Does regression produce representative estimates of causal effects?" *American Journal of Political Science* 60(1):250–267.

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.

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.

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.

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

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.

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 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.

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.".

Mazumder, Rahul, Trevor Hastie and Robert Tibshirani. 2010. "Spectral regularization algorithms for learning large incomplete matrices." *Journal of machine learning research* 11(Aug):2287–2322.

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.

Plmper, Thomas and Vera E. Troeger. 2019. "Not so Harmless After All: The Fixed-Effects Model." *Political Analysis* 27(1):21–45.

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.

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 Materials (Not For Publication)

Table of Contents

A.1. A Directed Acyclic Graph (DAG)

A.2. Algorithms

A.2.1. The IFEct Algorithm

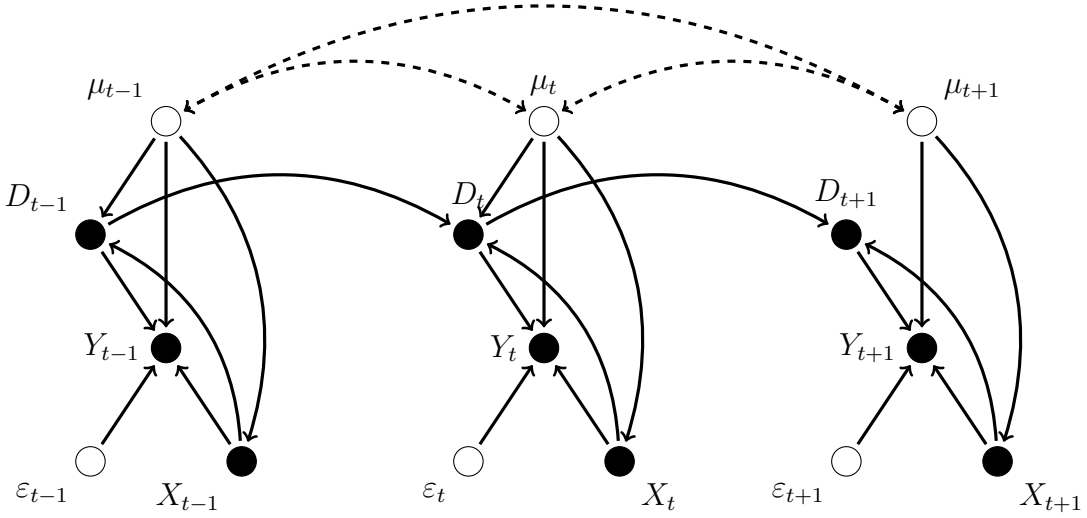
A.2.2. Algorithm for the Wald Test

A.3. Proofs

A.5. Additional Information on Empirical Examples

A.1. A Directed Acyclic Graph (DAG)

FIGURE A1. A DAG ILLUSTRATION



Note: Unit indicies are dropped for simplicity. Vector μ_t represents unobserved time-invariant and decomposable time-varying (for IFEct and MC) confounders.

A.2. Algorithms

A.2.1. 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 non-treated ($D_{it}=0$):

Step 2a. Update $\hat{\beta}^{(h+1)}$ using the non-treated 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 fe model to initialize μ, α_i and ξ_t):

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

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

Step 2b. For all i, t, define

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

For all non-treated 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:

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

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

A.2.2. Algorithm for the Wald Test

We use a wild bootstrap procedure to obtain an empirical distribution of the statistic under the Null hypothesis that $ATT_s = 0, \forall s \leq 0$. Compared with suing a closed-form asymptotic sampling distribution, the wild bootstrap has several advantage: (1) it has better finite sample properties and results in smaller Type-I error; (2) it preserves the temporal correlation in the residuals; and (3) it corporates uncertainties from the model-fitting stage, which is especially important when we apply IFEct or MC estimators. Note that this test does not impose any additional parametric assumptions.

For notational convenience only, let's consider a DGP in which N_{tr} units receives a treatment at period $T_0 + 1$ and it persists. The same algorithm can be applied to more complicated scenarios in which a treatment starts at different time periods or switches on and off. The procedure is as follows:

Step 1. Fit a model using observations under the control condition ($D_{it} = 0$) with a tunning parameter (i.e., r or θ) and obtain the residuals for each observation $\hat{e}_{it}$.

Step 2. For treated units ($i \in \mathcal{T}$), estimate the "ATT" for each pre-treatment period by averaging the residuals at period t:

$$\widehat{ATT}_t = \sum_{i \in \mathcal{T}} \hat{e}_{it}/N_{tr}, \quad t \leq T_0$$

and obtain an F statistic:

$$F^{obs} = \frac{\sum_{i \in \mathcal{T}} \sum_{t=1}^{T_0} \left(\hat{e}_{it}^2 - (\hat{e}_{it} - \widehat{ATT}_t)^2 \right) / T_0}{\sum_{i \in \mathcal{T}} \sum_{t=1}^{T_0} (\hat{e}_{it} - \widehat{ATT}_t)^2 / (N_{tr} \times T_0 - T_0)}$$

Step 3. To construct the h^{th} bootstrapped sample, randomly assign unit i the weight $w_i^{(h)} = 1$ with probability 0.5 and $w_i^{(h)} = -1$ with probability 0.5, and generate new pseudo-residuals $\tilde{e}_{it}^{(h)} = \hat{e}_{it} \times w_i^{(h)}$ and the corresponding new outcomes: $\tilde{y}_{it}^{(h)} = \hat{Y}_{it}(0) + \tilde{e}_{it}^{(h)}$

Step 4. Using the same method in Steps 1 and 2, this time, with the bootstrapped sample, obtaining a new F statistic: $F^{(h)}$.

Step 5. Repeat Steps 3 and 4 for B times and obtain an empirical distribution of the F statistic under H0: $F^{(1)}, F^{(2)}, \ldots, F^{(B)}$.

Step 6. Calculate the p value: $p = \sum_{h=1}^{B} \mathbb{1}[F^{(h)} > F^{obs}]/B$.

A.2.3. Permutation test

A.3. Proofs

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_{tr}, respectively. Under FEct, parameters $\hat{\beta}$, $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\xi}_t$ are estimated using the following equations:

$$Y_{it} = X_{it}'\beta + \mu + \alpha_i + \xi_t + \varepsilon_{it},\ D_{it} = 0, \tag{A1}$$
$$\sum_{D_{it}=0} \alpha_i = 0, \sum_{D_{it}=0} \xi_t = 0,$$

The dataset used for estimating the parameters consists an unbalanced panel since we do not use data with $D_{it} = 1$. 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 units observed in period t as N_t, then $N_t \leq N$ and $\sum_{t=1}^{T} N_t = n_{co}$, the number of observations in the dataset. Similarly, denote the number of periods in which unit i is observed 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 $\Delta_1 * \Delta_2' = \mathbf{D}$.

Under IFEct, parameters $\hat{\beta}$, $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\xi}_t$, $\hat{\lambda}_i$, $\hat{F}_t$, are estimated via the following equations:

$$Y_{it} = X_{it}'\beta + \lambda_i' f_t + \alpha_i + \xi_t + \varepsilon_{it},\ D_{it} = 0$$
$$\sum_{D_{it}=0} \alpha_i = 0, \sum_{D_{it}=0} \xi_t = 0, \mathbf{\Lambda}'\mathbf{\Lambda} = diagnal, \mathbf{F}'\mathbf{F}/T = \mathbf{I}_r$$

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 $Q_{\mathbf{A}}$.

Some regularity conditions are necessary for proving the ATT estimator's consistency. 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 $E\left[\mathbf{x_{i,t}}\mathbf{x_{i,t}}'\right]$ 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 Assumptions (1), (2) and regularity conditions, all the following limits 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:

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

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

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

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.

(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 Assumptions (1), (2) 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 constrains 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,

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

$$= \begin{pmatrix}\tilde{\mathbf{X}}'\tilde{\mathbf{X}} & \tilde{\mathbf{X}}'\Delta_2 \\ \Delta_2'\tilde{\mathbf{X}} & \Delta_2'\Delta_2\end{pmatrix} - \begin{pmatrix}\tilde{\mathbf{X}}'\Delta_1 \\ \Delta_2'\Delta_1\end{pmatrix} diag\{\frac{1}{T_i}\}(\tilde{\mathbf{X}}\Delta_1', \Delta_2\Delta_1')$$

Using Lemma 1, we know that as $N_{co} \to \infty$, each term in the expression above will converge to a fixed matrix.[A1] Using the Slutsky 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.[A2] ■

Lemma 3 *Under Assumptions (1a), (2) 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 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 Assumptions (1) and (2) as well as regularity conditions,*

$$\mathbb{E}[\widehat{ATT}_s] = ATT_s \text{ and } \mathbb{E}[\widehat{ATT}] = ATT;$$

$$\widehat{ATT}_s \xrightarrow{p} ATT_s \text{ and } \widehat{ATT} \xrightarrow{p} ATT \text{ as } N \to \infty.$$

[A1] As $N_{co} \to \infty$, N also goes to infinity. And replacing $\mathbf{X}$ with $\tilde{\mathbf{X}}$ won't change the basic results.

[A2] 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.

Proof:

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

Using lemma 1, we know that

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

Therefore, unbiasedness holds.

For consistency, we know from the proof of lemma 2 that:

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

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 $Var[\alpha_i - \hat{\alpha}_i]$ 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 (FEct as a matching estimator) *: Under Assumptions (1) and (2), 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 all the non-treated observations.*

Proof: When there is no covariate,

$$\hat{\mu} = \bar{Y}_{D_{it}=0} = \frac{1}{n}\iota_n' Y_{D_{it}=0}$$
$$\hat{\alpha}_i + \hat{\xi}_t = \nu_{it}'(\hat{\alpha}, \hat{\xi})' = \nu_{it}'\Phi_{11}^{-1}\Delta'\tilde{Y}_{D_{it}=0} = [\nu_{it}'\Phi_{11}^{-1}\Delta'(I - \frac{1}{n}\iota_n')]Y_{D_{it}=0}$$

Therefore,

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

.

For generalized DID, we can further calculate the weights of each observation in $\widehat{Y_{it}(0)}$. 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. In this scenario, equation (8) in the proof of lemma 1 becomes:

$$\begin{pmatrix} \Delta'\Delta, & \Pi' \\ \Pi, & 0 \end{pmatrix} \begin{pmatrix} \hat{\gamma} \\ \hat{\lambda} \end{pmatrix} = \begin{pmatrix} \Delta'\tilde{Y} \\ 0 \end{pmatrix}$$

where $\Delta'\Delta = \begin{pmatrix} diag\{T_i\} & \mathbf{D}' \\ \mathbf{D} & diag\{N_t\} \end{pmatrix}$, and $\Pi = \begin{pmatrix} T & \cdots & T & T_0 & \cdots & T_0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & N & \cdots & N & N_0 & \cdots & N_0 \end{pmatrix}$

These linear equations are easy to solve via elementary transformation as they are highly symmetric. We skip the details here. The intuition is to subtract from each row the row below it in order to simplify the equations. Finally, we get the following the solution for fixed effects:

$$\begin{pmatrix} \hat{\alpha}_i \\ \hat{\xi}_t \end{pmatrix} = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T}\tilde{Y}_{it} + \frac{1}{NT-N_1T_1}\frac{N_1}{N_0}\sum_{t=1}^{T_0}\sum_{i=1}^{N}\tilde{Y}_{it} - \frac{1}{NT-N_1T_1}\frac{T_1N_1}{TN_0}\sum_{i=1}^{N_0}\sum_{t=1}^{T}\tilde{Y}_{it} \\ \frac{1}{N}\sum_{i=1}^{N}\tilde{Y}_{it} + \frac{1}{NT-N_1T_1}\frac{T_1}{T_0}\sum_{i=1}^{N_0}\sum_{t=1}^{T}\tilde{Y}_{it} - \frac{1}{NT-N_1T_1}\frac{T_1N_1}{T_0N}\sum_{t=1}^{T_0}\sum_{i=1}^{N}\tilde{Y}_{it} \end{pmatrix}$$

for $t = 1, 2, ..., T_0$ and $i = 1, 2, ..., N_0$.

And

$$\begin{pmatrix} \hat{\alpha}_i \\ \hat{\xi}_t \end{pmatrix} = \begin{pmatrix} \frac{1}{T_0}\sum_{t=1}^{T_0}\tilde{Y}_{it} - \frac{1}{NT-N_1T_1}\frac{T}{T_0}\sum_{t=1}^{T_0}\sum_{i=1}^{N}\tilde{Y}_{it} + \frac{1}{NT-N_1T_1}\frac{T_1}{T_0}\sum_{i=1}^{N_0}\sum_{t=1}^{T}\tilde{Y}_{it} \\ \frac{1}{N_0}\sum_{i=1}^{N}\tilde{Y}_{it} - \frac{1}{NT-N_1T_1}\frac{N}{N_0}\sum_{i=1}^{N_0}\sum_{t=1}^{T}\tilde{Y}_{it} + \frac{1}{NT-N_1T_1}\frac{N_1}{N_0}\sum_{t=1}^{T_0}\sum_{i=1}^{N}\tilde{Y}_{it} \end{pmatrix}$$

for $t = T_0+1, T_0+2, ..., T$ and $i = N_0+1, N_0+2, N$. So $\widehat{Y_{it}(0)} = \hat{\mu} + \hat{\alpha}_i + \hat{\xi}_t = \sum_{t=1}^{T}\sum_{i=1}^{N} W_{it}\tilde{Y}_{it}$. It is thus clear that FEct uses all the observations with $D_{it} = 0$ to construct the counterfactual.

When we use the full panel, the last two terms in each of the four expressions become zero and we get the conventional within-estimator. ■

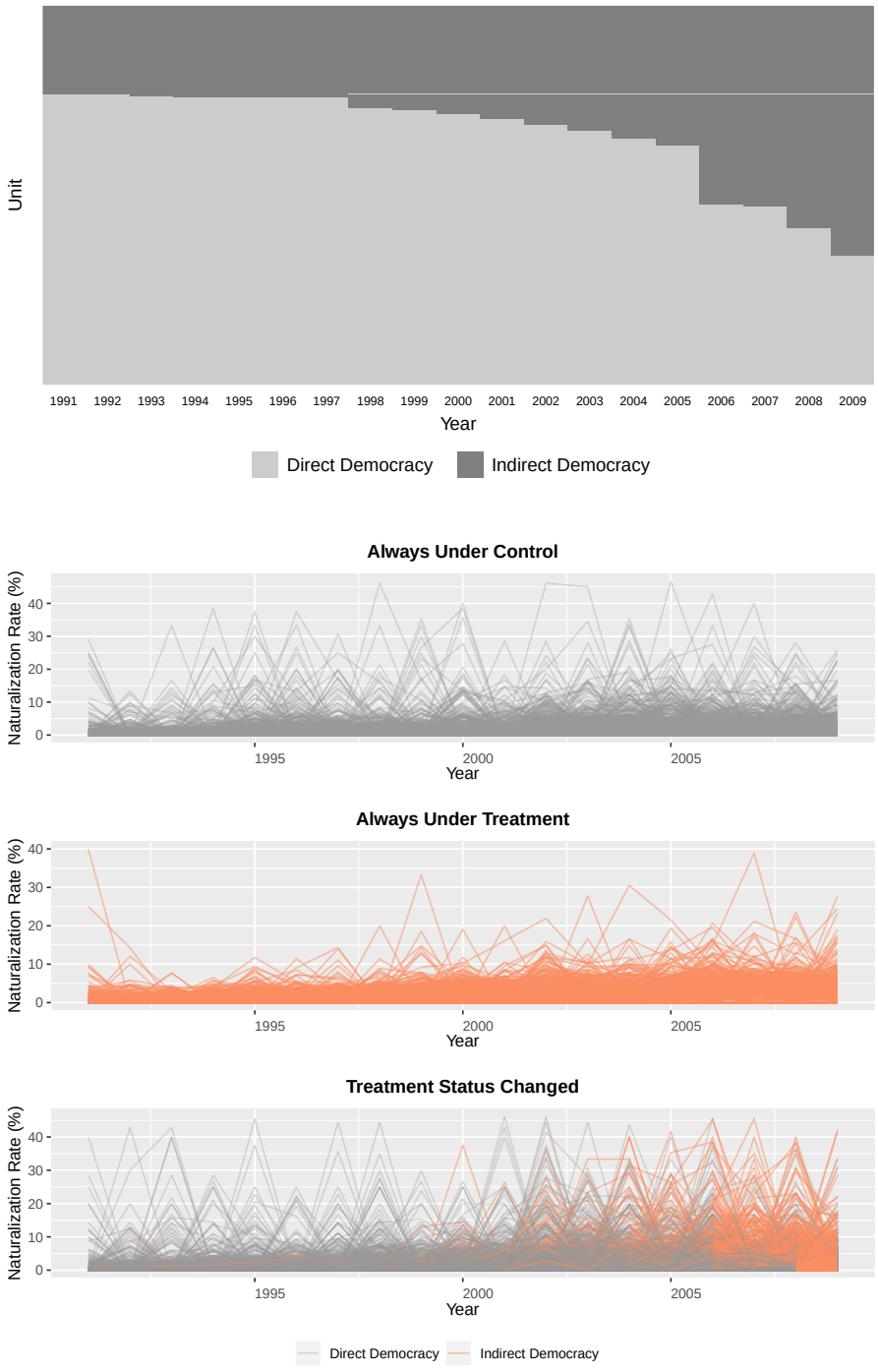
Proposition 3 (Unbiasedness and Consistency of IFEct) *: Under Assumptions (1a) and (2) as well as regularity conditions,*

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

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. ■

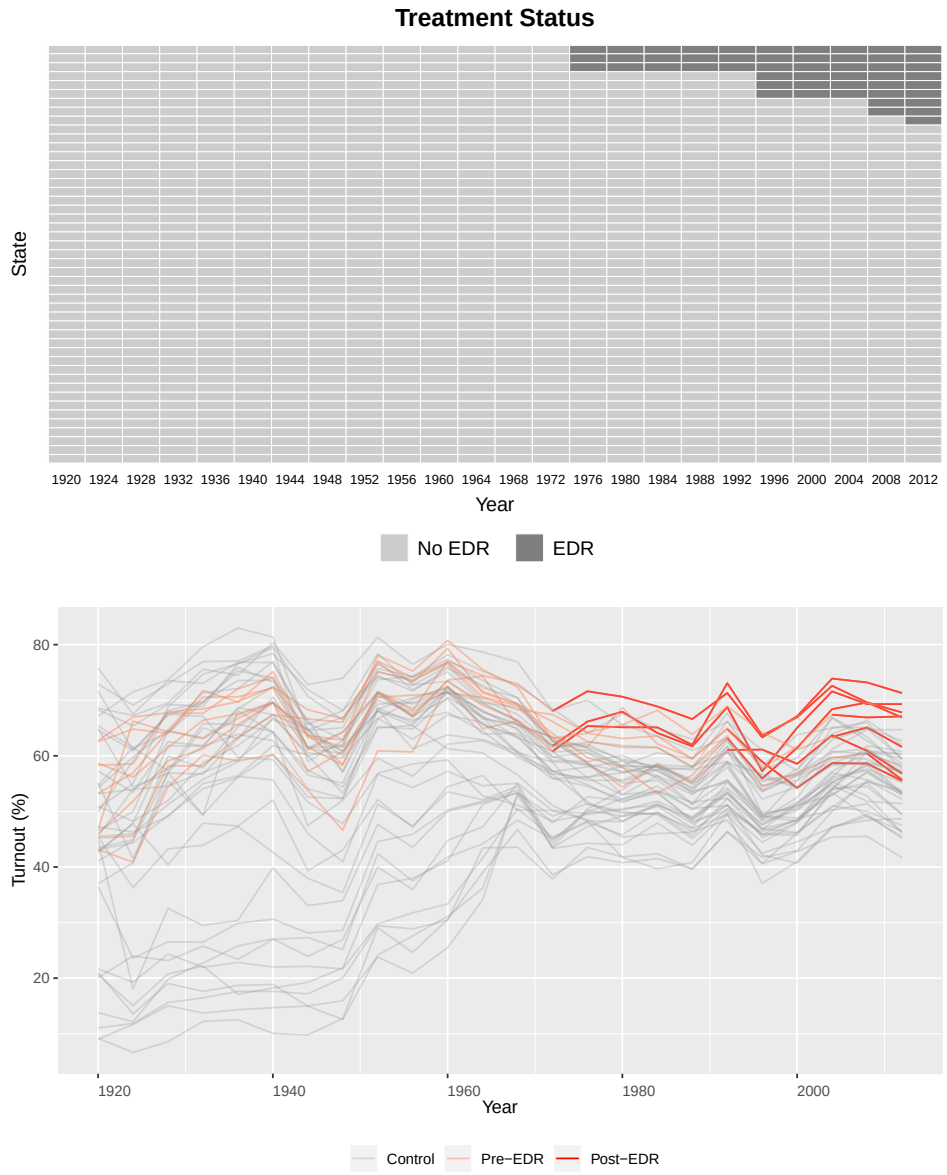
A.4. Additional Information on Empirical Examples

FIGURE A2. PLOT OF RAW DATA: HAINMUELLER AND HANGARTNER (2015)



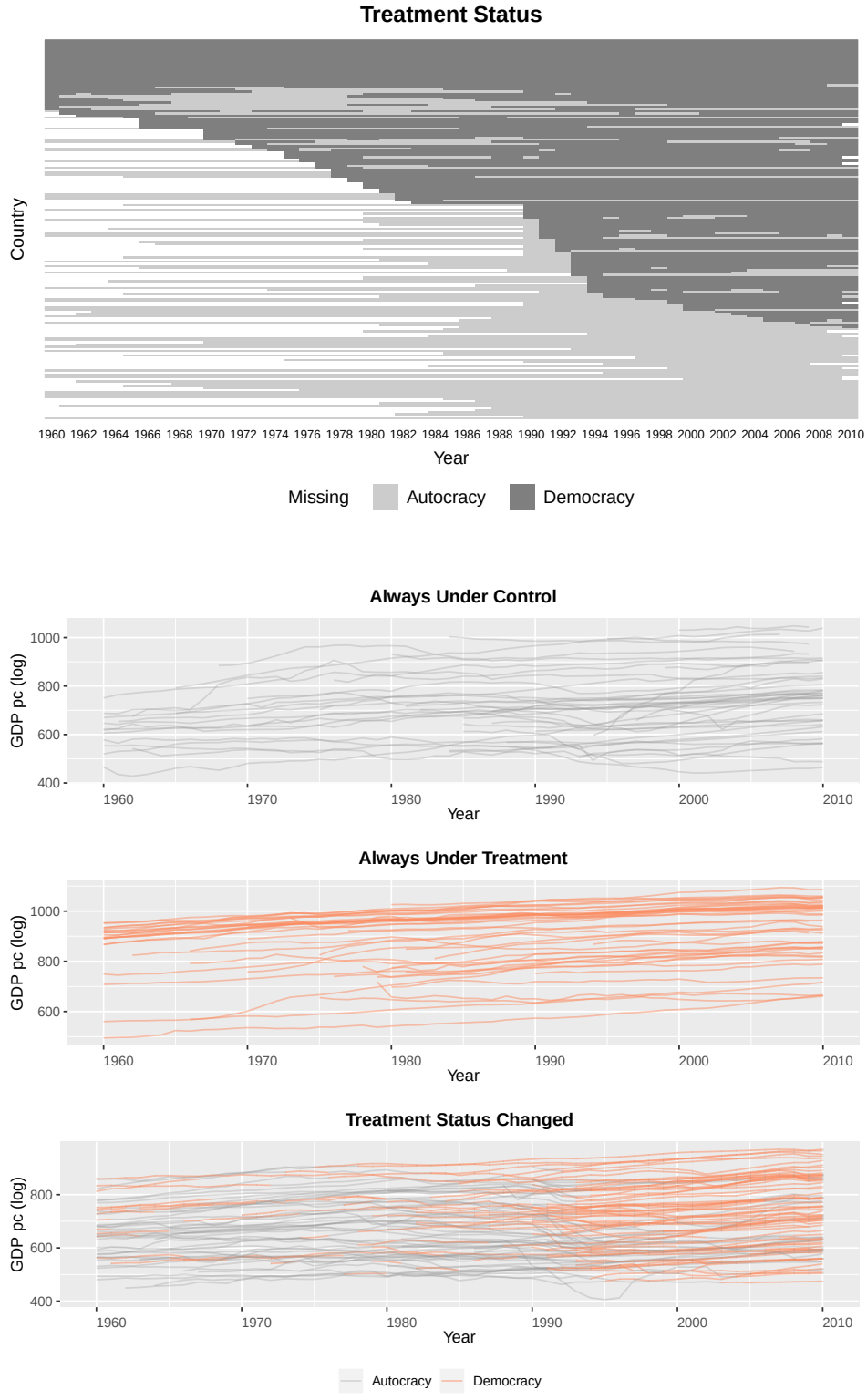
Note: The above figures plot the treatment status (for the first 50 units, top) and outcome variable (bottom) using data from Hainmueller and Hangartner (2015).

FIGURE A3. PLOT OF RAW DATA: XU (2017)



Note: The above figures plot the treatment status (top) and outcome variable (bottom) using data from Xu (2017).

FIGURE A4. PLOT OF RAW DATA: ACEMOGLU (2019)



Note: The above figures plot the treatment status (top) and outcome variable (bottom) using data from Acemoglu et al. (2019).

FIGURE A5. ORIGINAL FINDING: ACEMOGLU (2019)

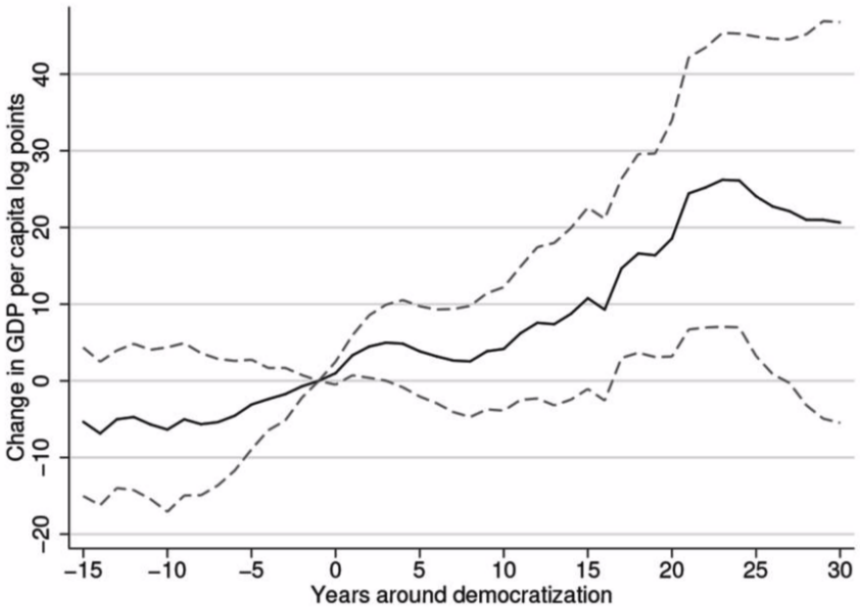
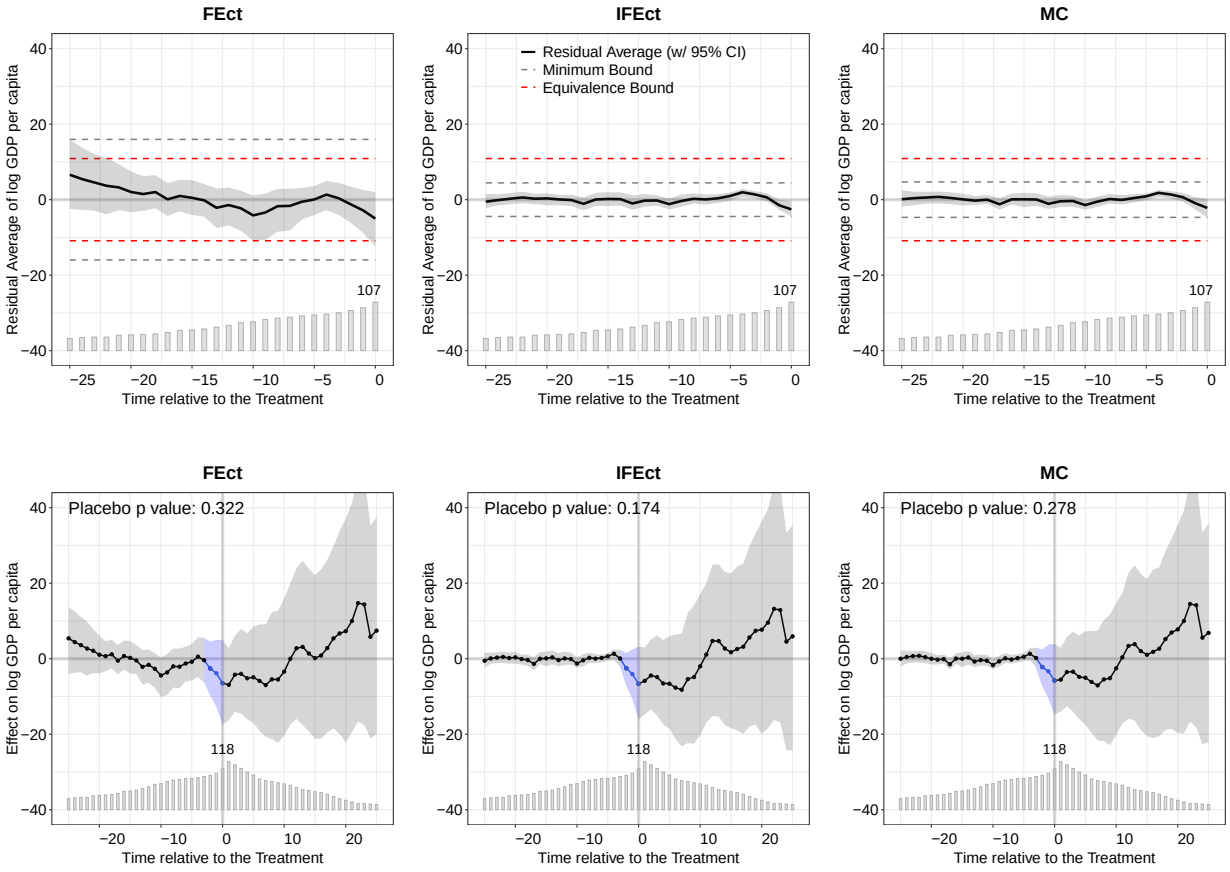


FIG. 4.—Semiparametric estimates of the over-time effects of democracy on the log of GDP, obtained with inverse-propensity-score reweighting. This figure plots semiparametric estimates of the effect of democratizations on GDP per capita in log points. The solid line plots the estimated average effect on GDP per capita on countries that democratized (in log points), with a 95 percent confidence interval in dashed lines. Time (in years) relative to the year of democratization runs on the horizontal axis. The estimates are obtained by assuming and estimating a probit model for democratizations based on GDP lags, which we use to estimate the propensity score and reweight the data. Section IV explains our approach in full detail.

Note: The above figure adapts from Figure 4 in Acemoglu et al. (2019).

Figure A6. Additional Results: Acemoglu (2019)



Note: The above figures show the results from the equivalence test (top) and the placebo test (bottom) for the effect of democracy on growth using data from Acemoglu et al. (2019).