Forecasted Treatment Effects

Irene Botosaru, Raffaella Giacomini, and Martin Weidner

August 31, 2023

WP 2023-32

https://doi.org/10.21033/wp-2023-32

FEDERAL RESERVE BANK of CHICAGO

Forecasted Treatment Effects*

Irene Botosaru†, Raffaella Giacomini‡, Martin Weidner§

August 31, 2023

Abstract

We consider estimation and inference about the effects of a policy in the absence of a control group. We obtain unbiased estimators of individual (heterogeneous) treatment effects and a consistent and asymptotically normal estimator of the average treatment effects, based on forecasting counterfactuals using a short time series of pre-treatment data. We show that the focus should be on forecast unbiasedness rather than accuracy. Correct specification of the forecasting model is not necessary to obtain unbiased estimates of the individual treatment effects. Instead, simple basis function (e.g., polynomial time trends) regressions deliver unbiasedness under a broad class of data-generating processes for the individual counterfactuals. Basing the forecasts on a model can introduce misspecification bias and does not necessarily improve performance even under correct specification. Consistency and asymptotic normality of the Forecasted Average Treatment effects (FAT) estimator attains under an additional assumption that rules out common and unforecastable shocks occurring between the treatment date and the date at which the effect is calculated.

JEL classification: C32, C53

Keywords: Polynomial regressions; Forecast unbiasedness; Counterfactuals; Misspecification; Heterogeneous treatment effects

*We thank Chris Muris, Krishna Pendakur and seminar and conference participants at several venues for helpful comments. The views expressed here are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve system. Irene Botosaru gratefully acknowledges financial support from the Social Sciences and Humanities Research Council of Canada IG 435-2021-0778 and from the Canada Research Chairs Program.

†McMasters University, Department of Economics. Email: botosari@mcmaster.ca

‡University College London, Department of Economics/Cemmap and Federal Reserve Bank of Chicago. Email: r.giacomini@ucl.ac.uk

§Oxford University, Department of Economics. Email: martin.weidner@economics.ox.ac.uk

1 Introduction

Evaluating policies, or treatments, is a central task in economics. Such evaluations usually require the construction of *counterfactuals*, i.e. outcomes that the treated would have gotten had they not received the treatment. The convention in the treatment effects literature is to construct counterfactuals for the treated using observed data on the *control*, i.e. individuals who do not receive the treatment and who are observed at every time period as the treated individuals. This strategy is used for evaluating policies with partial participation at every point in time.[1] When there is universal participation,[2], researchers typically rely on structural econometrics for policy evaluation.[3]

In this paper, we introduce a simple framework for the evaluation of policies with universal participation. Our approach can also be used as a robustness check when there is a control group, but its validity is uncertain or the assumptions that existing methods impose may be too strong to justify in the empirical application considered.[4] While our approach does not rely on structural models, it is unique in its focus on understanding what assumptions on the data-generating processes

[1]For early work on estimation of the average treatment effect on the treated by imputing or estimating the counterfactual for individuals in the treatment group by using observations from the control group, see, e.g., Ashenfelter and Card (1985); Heckman et al. (1997, 1998).

[2]A partial list of universal policies includes: child care programs targeted at all children (e.g., Baker et al. (2008); Cornelissen et al. (2018)), mandatory job search programs targeted at all unemployed youths in the labor market (e.g., Blundell et al. (2004)), environmental policies targeted at reducing air pollution and congestion (e.g., Gallego et al. (2013), Chen and Whalley (2012)), country-level trade agreements such as NAFTA (Trefler (2004)), fiscal policies (Oscar and Taylor (2016)), state-level texting bans (Abouk and Adams (2013)), and, of course, global pandemics such as COVID-19.

[3]For example, Heckman and Vytlacil (2005) discuss how functional form restrictions and support conditions can be used to substitute for the lack of control individuals.

[4]For example, estimators for the average treatment effect under unconfoundedness require that researchers choose both which observed covariates to condition on to guarantee that unconfoundedness holds (and include them in the propensity score) and the functional form of the propensity score. Both are fraught with specification uncertainty, see, e.g., Hirano and Imbens (2001); Kitagawa and Muris (2016). Likewise, difference-in-differences-type estimators require some sort of parallel-paths assumption. Although there is work that allows for a variety of robustness and sensitivity analyses, researchers must choose how to construct the control group and the weights placed on different time periods, c.f. Roth et al. (2023).

for the counterfactuals are needed to consistently estimate treatment effects. The treatment effects literature in the presence of a control group does not usually discuss the data-generating processes, opting instead for high-level assumptions such as unconfoundedness or parallel paths. Recent literature has illustrated how the parallel paths assumption implicitly imposes restrictions on individual dynamic choice (e.g., Ghanem et al. (2022); Marx et al. (2023)), and our approach can be seen as reflecting a possibly growing interest in the literature on making the assumptions on the data-generating process explicit.

Our parameter of interest is the average treatment effect on the treated (ATT).[5] Our baseline method uses individual time series of pre-treatment outcomes to forecast individual counterfactuals. The cross-sectional average of the individual differences between the observed post-treatment outcome at a particular time and the forecasted counterfactual is taken as an estimate of the ATT at that time period. We call our estimator the Forecasted Average Treatment effect (FAT). We show that FAT is a consistent and asymptotically normal estimator of the ATT, even when the number of time periods used to forecast the counterfactual is small. This is true under two high-level assumptions: 1) the forecasts are unbiased on average; and 2) the average differences between the observed post-treated outcomes and the forecasted counterfactuals satisfy a central limit theorem. We then proceed to characterize the class of data generating processes for the individual counterfactuals and the forecast methods that satisfy these requirements.

The focus on an individual-level forecast analysis has some advantages, e.g. it allows for heterogeneous treatment effects, unbalanced panels and staggered treatment timing. On the other hand, achieving robustness to the absence of a valid control group naturally comes at a cost. For example, a key implication of the central limit theorem assumption is that we will not be able to control for common shocks that affect all treated individuals between the time of the treatment and the time at which we compute the effects, unless these shocks are forecastable using pre-treatment data. A key insight of the paper is that these costs do not however include the strong and

[5] In the baseline case without a control group, the ATT is the same as the average treatment effect.

unrealistic assumption that forecasts are based on correctly specified models for the individual counterfactuals. All we need, in fact, is the ability to produce forecasts of the counterfactuals that are *unbiased on average* across individuals.

Forecast unbiasedness is typically not the main concern in time series and panel forecasting, which focuses instead on accuracy (e.g., achieving variance reduction at the cost of some bias). Our first contribution is to show that forecasts based on basis function regressions on pre-treatment outcomes (e.g., polynomial time trends) are unbiased estimators of the individual treatment effects, and that the FAT based on these forecasts is a consistent and asymptotically normal estimator of the ATT. This is true for a large class of data generating processes that express the counterfactuals as the sum of, potentially, two individual-specific unobserved components: a mean-stationary process and a random walk. This flexible class encompasses, e.g., linear panel models with fixed effects and lagged outcomes with heterogeneous coefficients or unit roots, and allows for stationarity or stochastic trends. If the data-generating process for sure includes a deterministic trend, the additional requirement is that the type of basis function used to produce the forecast is correctly specified and the number of basis functions is larger than the true one. This result shows that it is not necessary to model the stochastic component of the counterfactuals to obtain unbiased forecasts. One should instead focus on correctly specifying the deterministic trend component (up to its order), if one believes that this component is present.

Basis function regressions are easy to implement and allow for short time series of pre-treatment data. The length of the estimation window and the order of the basis function are tuning parameters. Our practical recommendation is to choose short estimation windows to guard against possible structural instability in pre-treatment data and to report results for a small range of values for the number of basis functions. For example, for polynomial regressions of order q and assuming the same basis functions are used across individuals, one can report FAT based on, say, $q = 0, ..., 3$ using an estimation window of length $q + 1$. If feasible, plotting the individual time series of pre-treatment data can guide the choice of polynomial order. For example, if this series shows trending behaviour, one can then decide if this is plausibly due to a stochastic trend (in which case any choice of q is valid) or a

polynomial time trend (in which case q should be larger than the true order). If the panel is balanced and one uses the same basis functions across individuals, the time series of pre-treatment outcomes averaged across individuals can guide the choice of basis functions, as in this case the FAT is simply obtained by a basis function regression on averaged outcomes.

The baseline case considers forecasts that are only based on pre-treatment outcomes and do not require specifying a model. Covariates could also be incorporated in the estimation of the FAT, by specifying a model and using it to forecast the counterfactuals (we call this the model-based FAT). This approach however requires stronger assumptions, including: correct specification of the model; availability of a consistent estimator for the coefficients if these coefficients are homogeneous; symmetry of the error term if the covariates are lagged outcomes with heterogeneous coefficients. Under these additional assumptions the model-based FAT is biased but consistent and asymptotically normal. Our simulations indicate that the model-based FAT is however sensitive to misspecification bias, and has comparable performance to the FAT obtained by basis function regression under correct specification. This suggests that the basis-function approach possesses desirable robustness properties.

Suppose that a control group is available, but that it is not valid from the perspective of existing approaches such as, e.g., difference in differences (DiD) estimators. Such a control group could be used to relax the assumption that there are no shocks that affect all individuals between the treatment and the time at which the FAT is computed. In practice, this is achieved by computing the FAT for both the treated and the control and then taking the difference of the two (we call this the DFAT estimator). For example, this approach allows one to eliminate the effect of an unpredictable common shocks that has the same average effect on the treated and control groups, while permitting the data-generating process for the counterfactuals to contain fully heterogeneous time trends. The heterogeneous time trends would instead violate the parallel paths assumption required by DiD.

2 Related literature

Our baseline case considers the absence of a control group and provides a consistent estimator of ATT based on forecasting counterfactuals. There is a small literature that discusses forecasting counterfactuals using Bayesian methods. For example, Brodersen et al. (2015) proposes Bayesian estimation of a time series model based on trend extrapolation, seasonal effects and covariates. Varian (2014) argues in favor of such models, stating that "a good predictive model can be better than a randomly-chosen control group, which is usually thought to be the gold standard." Besides assuming correct specification of the model, the Bayesian approach typically results in biased forecasts. In contrast, our approach does not require correct specification of the model and delivers unbiasedness.

This paper forecasts individual counterfactuals using panel data. There is a long literature on forecasting with panel data, e.g., Baltagi (2013), including some recent Bayesian approaches for panel data with a short time dimension (Liu et al. (2020)). The main difference with this literature is their focus on forecast *accuracy*, whereas an insight of this paper is that one should focus on forecast *unbiasedness* if the goal is forecasting counterfactuals as an ingredient for estimation of the ATT. The literature typically also assumes correct specification of the model, whereas we show that this is not necessary for consistent ATT estimation.

Forecast unbiasedness has been studied in the context of time series models. Fuller and Hasza (1980) show that a correctly specified AR(1) model gives unbiased forecasts; Dufour (1984) extends the result to possibly misspecified AR(p) models and Cryer et al. (1990) to possibly misspecified ARIMA(p, d, q) models. Key assumptions in the above literature are a symmetry restriction on the model residuals and stationarity. A contribution of this paper is to show that we can obtain unbiased forecasts for a much more general class including stationary and nonstationary data-generating processes. Our result accommodates misspecified models and does not maintain symmetry or stationarity.

With a short panel, Mavroeidis et al. (2015) consider an AR(1) model with correlated random effects and heterogeneous autoregressive parameter. The authors

propose an MLE-based method that could be used to produce unbiased forecasts, however both a stationarity and a symmetry assumption on the data generating process are maintained. To the best of our knowledge, the issue of unbiased forecasts with a short panel and parameter heterogeneity has not been considered for non-stationary processes.

So-called "event studies" in finance (e.g., Brown and Warner (1985); MacKinlay (1997)) are used to assess the impact on stock returns of firm-specific events. The methodology used in this literature is a special case of the approach considered here, since it is based on forecasting stock returns after the event (i.e., counterfactuals) using either the sample mean of returns before the event (i.e., a basis function regression of order zero) or a regression model such as the CAPM estimated over pre-event data. This literature leverages specific characteristics of financial data that are not generally extendable to other types of data, such as the availability of long time series and consensus about the plausible data-generating process for the counterfactuals (e.g., a mean-zero serially uncorrelated process for daily stock returns). In contrast, this paper addresses the uncertainty about the data-generating process (in particular in terms of the presence of deterministic and/or stochastic trends) and the short time dimension that characterize the treatment effects literature in other fields. An implication of our findings for this literature is that correct specification is not necessary to produce unbiased forecasts of counterfactuals, implying that the sample mean is valid under weaker assumptions than those in, e.g., MacKinlay (1997). Another implication is that the existing approach can also be applied to stock prices, since the sample mean is also valid even when the counterfactuals have a unit root.

Interrupted time series analysis (ITS) is a quasi-experimental design sometimes used in, e.g., health economics and criminology to evaluate the impact of a policy implemented at the population level. ITS does not require a control group, and uses time series data on a single treated unit from before the policy to construct the unit's counterfactual outcome trend, that is then compared to the unit's post-policy outcome trend, see e.g. Bernal et al. (2017), Baiker and Svoronos (2019), Miratrix (2022). ITS is concerned with an aggregate time series, it models only a linear trend, and it implicitly makes strong assumptions about the data-generating process for the

counterfactual.

A prime application of our method uses polynomial regressions to forecast counterfactuals. Polynomial regressions are also used in the literature on regression discontinuity (RD), e.g., Cattaneo and Titiunik (2022), Gelman and Imbens (2019). Event studies that use time as a score variable in RD designs, or "RDD in time," are often applied in environmental economics to evaluate policies with universal participation (see, e.g., Hausman and Rapson (2018); Gillingham et al. (2020); Tu et al. (2020); Li et al. (2020); Greenstone et al. (2022)) as well as in other fields (e.g., Kuhn and Shen (2023) and Aguilar et al. (2021)). RDD in time relies on high frequency data around the treatment time and uses polynomial regressions before and after the treatment time. In contrast, we consider only polynomial regressions before the treatment time as a way to forecast counterfactuals, which we then compare with actual outcomes. As discussed in Hausman and Rapson (2018), the RD in time approach is problematic since it can mix short and long-run effects of the treatment, see also Cattaneo and Titiunik (2022).

The synthetic control method has been increasingly used to evaluate the effect of interventions implemented at an aggregate level (such as a cities, regions, or countries) on an aggregate outcome, see Abadie (2021) for a recent review. In the conventional setting for synthetic controls, there is only one unit that is treated. However, there are many untreated units in the donor pool from which pseudo-controls can be chosen, i.e. these are untreated units selected such that the weighted average of their past outcomes "resembles" the trajectory of past outcomes of the treated unit. The counterfactual outcome for the treated unit is then constructed as a weighted average of the post-treatment outcomes of the selected pseudo-control units. In comparison, in our baseline setting, all individuals in the population are treated and there are no control units. The counterfactual outcome for each treated unit is a weighted average of the unit's *own* past outcomes. The properties of our estimator rest on averaging across many treated units, an advantage of which is standard inference. Other differences with the synthetic control framework are that we describe the class of data generating processes that obtains a consistent estimator of the ATT, and that our results apply even when the number of pre-treatment time periods is small.

We leverage the cross-sectional dimension for this result, which is not possible in the conventional synthetic control setting.

Our analysis allows for heterogeneous treatment effects. It is well known that in this case Ordinary Least Squares or Two Way Fixed Effects estimators in linear panel data models are generally inconsistent for the average treatment effect, see, e.g., Wooldridge (2005); Chernozhukov et al. (2013); Imai and Kim (2019); Sloczyński (2020); de Chaisemartin and D'Haultfoeuille (2020); Goodman-Bacon (2021). Our approach delivers a consistent estimator of the average treatment effect (on the treated) in the absence of a control group. All proposed solutions instead assume the existence of a control group in every period, see, e.g., Cengiz et al. (2019); Callaway and Sant'Anna (2021); Sun and Abraham (2020); Goodman-Bacon (2021); Baker et al. (2022); Borusyak et al. (2021); Liu et al. (2023); Chan and Kwok (2022); Roth et al. (2023). In addition, in order to avoid the incidental parameter problem when estimating panel models with fixed effects, the maintained assumptions in this literature are the absence of lagged outcomes when the panel is short and the homogeneity of the regression coefficients that enter potential outcomes, e.g., Angrist and Pischke (2009). Our class of data-generating processes instead allows for lagged dependent outcomes and fully heterogeneous parameters.

Least-squares estimates of treatment effects are often interpreted as consistent estimates of certain weighted average treatment effects, to allow for the possibility of model misspecification, see e.g. Theorem 1 in Chernozhukov et al. (2013). The researcher generally has no control over the weights, which can become negative in panel data models where fixed effects are included in the estimation. In addition, the weighted average interpretation does not address the incidental parameter problem, which occurs even for correctly specified models with homogeneous treatment effects. By contrast, our results show that, if treatment effects are estimated via averages over unbiased forecasts of counterfactuals, the correct unweighted treatment effect is consistently estimated, as long as the average is over a sufficiently large cross-section of observations. Our method thus signals a shift in focus: rather than trying to obtain consistent estimates of the model's parameters, we focus on unbiased forecasts of counterfactuals. In this way, our method avoids both the incidental parameter

problem and the negative weighting issue mentioned above.

The idea of imputing missing values in the outcome matrix, i.e. the counterfactual outcomes, has recently been used by, e.g., Athey et al. (2021); Bai and Ng (2021); Fernández-Val et al. (2021), where the goal is to control flexibly for complicated interactions between individual specific and time specific heterogeneity in panel data models using low-rank matrix approximations. Using language from that literature, our framework has a thin matrix of outcomes since the cross-sectional dimension is much larger than the time dimension. In the thin matrix case, missing potential outcomes in the last period are imputed using control units with similar lagged outcomes. The main assumption that allows this is unconfoundedness. Since we do not observe cross-sectional control units, our outcome matrix does not contain outcomes for control units and we do not appeal to unconfoundedness.

Conceptually, when there exists a control group, our solution resembles difference-in-differences or, more generally, an "event-study design analysis" as defined by, e.g., Borusyak et al. (2021). Although the estimated outcome equations may look similar, there is an important distinction between these methods and ours. For example, the extension of FAT to the case of a control group (DFAT) uses control groups to correct for the effect of a common shock, while the other methods use control groups to correct for selection into treatment (under different assumptions). Additionally, FAT allows for heterogeneous time trends as well as for heterogeneous effects of lagged pre-treatment outcomes. In contrast, there is no straightforward way to control for pre-treatment lagged outcomes in the specifications of, e.g., Sun and Abraham (2020); Callaway and Sant'Anna (2021); Borusyak et al. (2021).

3 Baseline case: no control group, no covariates

In this section, we introduce the parameter of interest and our proposed estimator. We first show that our estimator is consistent and asymptotically normal under a high-level unbiasedness assumption for the forecasts of the counterfactuals. We then derive sufficient conditions on the class of data-generating processes and the forecast methods that satisfy the unbiasedeness assumption.

3.1 Parameter of interest and estimator

Consider a treatment or a policy that is implemented at a time τ. In this section we consider the case where the treatment affects all individuals in the population at the same time,[6] that is, the treatment indicator of individual i at time t is given by

$$d_{it} := \ 1(t > \tau) \text{ for all } i = 1, \ldots, n.$$

We adopt the potential outcomes framework with each individual i having two potential outcomes at each time t: $y_{it}(1)$ if the individual is exposed to the treatment (or treated) and $y_{it}(0)$ if the individual is not exposed to the treatment (or control). Under the stable unit treatment value assumption (SUTVA), the observed outcome of individual i at t is given by:

$$y_{it} = (1 - d_{it})\, y_{it}(0) + d_{it}\, y_{it}(1).$$

In our baseline setting where all individuals are treated after time τ, the observed individual outcome y_{it} is

$$y_{it} = \begin{cases} y_{it}(0) & \text{for } t \leq \tau, \\ y_{it}(1) & \text{for } t > \tau. \end{cases} \tag{1}$$

We follow the literature on heterogeneous treatment effects in defining the average treatment effect on the treated (ATT)[7] $h \geq 1$ periods after τ as:

$$\text{ATT}_h := \frac{1}{n} \sum_i \mathbb{E}\left[y_{i\tau+h}(1) - y_{i\tau+h}\,(0)\right] \tag{2}$$

$$= \frac{1}{n} \sum_i \mathbb{E}\left[y_{i\tau+h} - y_{i\tau+h}\,(0)\right], \tag{3}$$

[6] In Section 3.3 we consider the more general case of staggered adoption with the treatment timing heterogeneous across individuals.

[7] Since all individuals are treated in our baseline setting, the ATT equals the average treatment effect (ATE).

where we used that $y_{i\tau+h}(1) = y_{i\tau+h}$ for $h \geq 1$. Note that with identically and independently distributed data, the right hand side of (2) reduces to the conventional $\mathbb{E}\left[y_{i\tau+h} - y_{i\tau+h}(0)\right]$.

The challenge in identifying and estimating ATT_h is that the counterfactual $y_{i\tau+h}(0)$ is not observed for $h \geq 1$. If a control group were available, then the conventional approach would be to impose sufficient assumptions that identify the parameter of interest from the observed post-treatment outcomes of the control group. In the absence of a control group, we exploit pre-treatment individual time series to obtain a forecast for the individual counterfactual $y_{i\tau+h}(0)$. We denote this forecast by $\widehat{y}_{i\tau+h}(0)$.

We call our proposed estimator for ATT_h the forecasted average treatment effect estimator, defined as:

$$\widehat{\mathrm{FAT}}_h := \frac{1}{n}\sum_{i=1}^{n}\left[y_{i\tau+h} - \widehat{y}_{i\tau+h}(0)\right], \tag{4}$$

where $\widehat{y}_{i\tau+h}(0)$ is a measurable function of past outcomes $\{y_{it}\}_{t\leq\tau}$.[8] Assumption 1 below guarantees that $\mathbb{E}(\widehat{\mathrm{FAT}}_h) = \mathrm{ATT}_h$. We explain how to obtain the individual-level forecast $\widehat{y}_{i\tau+h}(0)$ in the sections that follow below. For now, we note that $\widehat{y}_{i\tau+h}(0)$ uses individual-specific pre-treatment outcomes, and since it is individual-specific, our estimator naturally accommodates unbalanced panels and heterogeneous treatment effects.

We make the following high-level assumptions.

Assumption 1 (Average unbiasedness)**.** *The forecast for time $\tau + h$, $h \geq 1$, is unbiased on average, in the sense that:*

$$\frac{1}{n}\sum_{i}\mathbb{E}\left(\widehat{y}_{i\tau+h}(0) - y_{i\tau+h}(0)\right) = 0. \tag{5}$$

Let $u_{i\tau+h} := y_{i\tau+h} - \widehat{y}_{i\tau+h}(0)$ be the forecasted individual treatment effect, i.e. the

[8] In the baseline case, the individual forecast depends only on the past outcomes of the treated, in particular, there are no covariates in the information set.

individual-specific difference between the observed post-treatment outcome (that is, the outcome with the treatment) and the forecasted counterfactual at $\tau + h,\ h \geq 1$.

Assumption 2 (CLT)**.** *Let* $\{u_{i\tau+h}\}$ *be a sequence of random variables that satisfies a CLT in the sense that*

$$\frac{\frac{1}{\sqrt{n}} \sum_i \left(u_{i\tau+h} - \mathbb{E}u_{i\tau+h}\right)}{\bar{\sigma}_n} \Rightarrow \mathcal{N}\left(0,1\right), \tag{6}$$

where $\bar{\sigma}_n^2 := \mathrm{Var}(\frac{1}{\sqrt{n}} \sum_i u_{i\tau+h}) < \infty$.

For example, when $\{u_{i\tau+h}\}$ is a sequence of independent but not identically distributed random variable, Theorem 5.11 in White (2001) gives an asymptotic normality result.

Note that Assumption 2 allows for weak cross-sectional dependence. It excludes shocks that affect all individuals after the treatment and that are unforecastable. In principle, the assumption allows for the possibility that some common shocks could be captured by the method used to forecast the counterfactuals. We will discuss in Section 5 how this assumption can be further weakened in the presence of a control group.

The next result shows that, under Assumptions 1 and 2, our estimator in (4) is consistent and asymptotically normal.

Lemma 1 (Consistency and asymptotic normality)**.** *For each* $i = 1, \ldots, n$*, let the forecast* $\widehat{y}_{i\tau+h}(0)$*,* $h \geq 1$*, be a function of* $\{y_{it}\}_{t \leq \tau}$*. Let Assumptions 1 and 2 hold. Then* $\widehat{\mathrm{FAT}}_h$ *satisfies:*

$$\frac{\sqrt{n}\left(\widehat{\mathrm{FAT}}_h - \mathrm{ATT}_h\right)}{\bar{\sigma}_n} \Rightarrow \mathcal{N}\left(0,1\right).$$

Proof. We have

$$\begin{aligned}\widehat{\mathrm{FAT}}_h - \mathrm{ATT}_h &= \frac{1}{n}\sum_{i=1}^{n}\left(y_{i\tau+h} - \widehat{y}_{i\tau+h}(0) - \mathbb{E}\left[y_{i\tau+h} - y_{i\tau+h}(0)\right]\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(y_{i\tau+h} - \widehat{y}_{i\tau+h}(0) - \mathbb{E}\left[y_{i\tau+h} - \widehat{y}_{i\tau+h}(0)\right]\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(u_{i\tau+h} - \mathbb{E}u_{i\tau+h}\right),\end{aligned}$$

where we used Assumption 1 to obtain the second equality above. Since our assumptions guarantee that $(u_{i\tau+h} - \mathbb{E}u_{i\tau+h})$ has zero mean and satisfies a CLT, we obtain the desired result. □

In the remainder of the paper, we provide low-level sufficient assumptions, including a full description of the class of data generating processes for $y_{it}(0)$ and the forecast methods which satisfy Assumption 1.

3.2 Unbiased forecasts of counterfactuals

In this section we characterize classes of data generating processes (DGPs) for the counterfactuals as well as forecasting methods that satisfy the unbiasedness assumption, Assumption 1. Note that the need to discuss the DGP for counterfactuals arises because of the lack of a control group, which means that we must rely on forecasting counterfactuals from pre-treatment observations. Note that one key difference with the forecasting literature is that forecasting typically requires knowledge of the DGP. In contrast, we can allow for a broad class of DGPs because we only require the ability to produce forecasts that are unbiased on average.

3.2.1 Stationary or stochastic trends DGPs

In this section we consider a class of DGPs such that Assumption 1 is satisfied generally, namely by any forecast that can be written as a weighted average of pre-treatment outcomes with weights summing to 1. The class of DGPs expresses the

counterfactual as the sum of potentially two unobserved stochastic components. This includes a variety of processes, such as stationary and non-stationary (unit root) ARMA processes with individual-specific parameters.

Assumption 3 (Stationary or stochastic trends DGPs)**.** *For $j = 1, 2$, let $I_j \in \{0, 1\}$ be a non-random binary indicator. The potential outcome $y_{it}(0)$ follows the process:*

$$y_{it}(0) = I_1 y_{it}^{(1)}(0) + I_2 y_{it}^{(2)}(0), \tag{7}$$

where $y_{it}^{(1)}(0)$ is an unobserved mean-stationary process and $y_{it}^{(2)}(0) = y_{it-1}^{(2)}(0) + u_{it}(0)$ is an unobserved random walk process with innovations satisfying $\mathbb{E}u_{it}(0) = 0$, for all $t \geq 2$.

Remark 1. Note that only one of the unobserved components in Assumption 3 is required to be present. This means that we accommodate stationarity as well as non-stationarity due to a stochastic trend. The user does not need to take a stand on whether the outcomes may be stationary or have a stochastic trend, as our method is robust to both.

Remark 2. When both components in Assumption 3 are present, the assumption is equivalent to the classical trend-cycle decomposition of macroeconomic time series with stochastic trends (e.g., Nelson and Plosser (1982); Watson (1986).

Remark 3. This class of DGPs is a plausible assumption for applications where either: 1) the time series of pre-treatment outcomes does not display a trend; 2) there is a trend in pre-treatment outcomes that is plausibly stochastic, rather than deterministic; 3) there is only one pre-treatment observation so a deterministic trend could never be modelled anyway. Applications where there is a trend in pre-treatment outcomes that is more plausibly deterministic than stochastic will be better suited for the class of DGPs that we consider in the next section.

A key insight of this paper is that one does not need a correctly specified model (beyond satisfying Assumption 3) to obtain unbiased forecasts of the counterfactuals. In fact, as the next result shows, any forecast expressed as a weighted average of pre-

treatment observations (with weights summing to one) satisfies the unbiasedness condition.

Theorem 1 (Unbiasedness for stationary or stochastic trends DGP)**.** *Let Assumption 3 hold. Denote by* $\mathcal{T}_i = \{\tau - R_i + 1, \ldots, \tau\}$ *the set of* R_i *time periods directly preceding the treatment date. Consider a weighted average of the pre-treatment outcomes:*

$$\widehat{y}_{i\tau+h}(0) = \sum_{t \in \mathcal{T}_i} w_{it} y_{it}, \tag{8}$$

where w_{it} *are non-random weights such that* $\sum_{t \in \mathcal{T}_i} w_{it} = 1$*. Then,*

$$\mathbb{E}\left[\widehat{y}_{i\tau+h}(0) - y_{i\tau+h}(0)\right] = 0. \tag{9}$$

Proof. It is sufficient to show that for each component $y_{it}^{(r)}(0)$, $r \in \{1, 2\}$,

$$\mathbb{E}\left[\sum_{t \in \mathcal{T}_i} w_{it}\, y_{it}^{(r)} - y_{i\tau+h}^{(r)}(0)\right] = 0. \tag{10}$$

For both the mean stationary component ($r = 1$) and the random walk component ($r = 2$) we have $\mathbb{E}\left(y_{it}^{(r)} - y_{i\tau+h}^{(r)}(0)\right) = 0$. Multiplying this equation by w_{it}, summing over $t \in \mathcal{T}_i$, and using the fact that the non-random weights w_{it} sum to 1, we obtain (10) for $r = 1$ and $r = 2$. □

Remark 4. Note that Theorem 1 shows how to obtain unbiased estimates of the individual (possibly heterogeneous)treatment effects. This is a stronger result than the requirement in Assumption 1 that the unbiasedness only holds on average. This means that under the assumptions of Theorem 1 one can not only obtain consistent and asymptotically normal estimates of the average effects, but also unbiased estimates of the *individual* effects.

There are many ways to obtain forecasts that are weighted averages of pre-treatment data - the sample mean being the most obvious example. In this paper

we focus on a general class of forecasts obtained via basis function regressions, such as polynomial time trends regressions. This class has the sample mean as a special case.

Definition 1 (Forecasts via basis function regressions)**.** *Consider a sequence of linearly independent functions* $\{b_k(t)\}_{k=0}^{q_i}$*,* $q_i \in \{0, 1, 2, \ldots, \tau - 1\}$*, on the interval* $\mathcal{T}_i$=$\{\tau - R_i + 1, \ldots, \tau\}$ *with* $R_i \in \{q_i + 1, \ldots, \tau\}$*, and such that* $b_0(t) = 1$ *for all* t*. For example, polynomial time trends set* $b_k(t) = t^k$*, with* q_i *the order of the polynomial. For each individual* i*, we forecast the counterfactual via individual-specific regressions of pre-treatment outcomes on the basis functions* $\{b_k(t)\}_{k=0}^{q_i}$*:*

$$\widehat{y}_{i\tau+h}^{(q_i,R_i)} := \sum_{k=0}^{q_i} \widehat{c}_{ik}^{(q_i,R_i)} b_k(\tau + h), \tag{11}$$

$$\widehat{c}_i^{(q_i,R_i)} := \underset{c \in \mathbb{R}^{q_i+1}}{\operatorname{argmin}} \sum_{t \in \mathcal{T}_i} \left(y_{it} - \sum_{k=0}^{q_i} c_k \, b_k \, (t) \right)^2, \tag{12}$$

where $c_i = (c_{i0}, \ldots, c_{iq_i})$ *is a* $q_i + 1$ *vector of individual-specific coefficients.*[9]

The definition above makes it clear that for any type of basis function the choice $q_i = 0$ yields the sample mean of pre-treatment outcomes as the forecast. The following example illustrates how the weighted average representation can arise quite naturally.

Example 1. *Consider* $b_k(t) = t^k$*. Set* $R_i = q_i + 1$ *and* $h = 1$*. In this case,* $\widehat{y}_{i\tau+1}^{(q_i)}(0) := \widehat{y}_{i\tau+1}^{(q_i,q_i+1)}(0)$ *can be defined iteratively as:*

$$\widehat{y}_{i\tau+1}^{(q_i)}(0) = \begin{cases} y_{i\tau} & \text{for } q_i = 0, \\ \widehat{y}_{i\tau+1}^{(q_i-1)}(0) - \left[\widehat{y}_{i\tau}^{(q_i-1)}(0) - y_{i\tau} \right] & \text{for } q_i > 0. \end{cases} \tag{13}$$

[9]Note that when $q_i = \tau - 1$, all pre-treatment outcomes are used in constructing the forecast, i.e. $R_i = \tau$. However, fewer observations can be used. We discuss the choice of the tuning parameters q_i and R_i in Section 3.2.3.

This iteration is quite intuitive: the q_i-forecast $\widehat{y}_{i\tau+1}^{(q_i)}(0)$ is formed by subtracting the lagged forecast error $\widehat{y}_{i\tau}^{(q_i-1)}(0) - y_{i\tau}$ from the $q_i - 1$ forecast $\widehat{y}_{i\tau+1}^{(q_i-1)}(0)$. Explicit formulas for this case are given by

$$\begin{aligned}
&\widehat{y}_{i\tau+1}^{(0)}(0) = y_{i\tau}, \\
&\widehat{y}_{i\tau+1}^{(1)}(0) = 2y_{i\tau} - y_{i\tau-1}, \\
&\widehat{y}_{i\tau+1}^{(2)}(0) = 3y_{i\tau} - 3y_{i\tau-1} + y_{i\tau-2}, \\
&\widehat{y}_{i\tau+1}^{(q_i)}(0) = \sum_{t=\tau-q_i}^{\tau} w_{it}^{(q_i)}\, y_{it}, \qquad\qquad w_{it}^{(q_i)} = (-1)^{(\tau-t)} \binom{q_i+1}{\tau-t+1},
\end{aligned}$$

where $\binom{a}{b} = \frac{a!}{b!(a-b)!}$ is the binomial coefficient.[10] *In all cases, the weights sum to 1.*

We now show that forecasts obtained by basis function regressions satisfy the weighted average requirement of Theorem 1.

Lemma 2. *For known basis function $b_k(t)$, $k = 0, 1, \ldots, q_i + 1$, $q_i = 0, 1, \ldots, \tau_i - 1$ that are linearly independent on $\mathcal{T}_i$ with $b_0(t) = 1$, the forecast in (11) can be written as a weighted average of past outcomes with weights that sum to 1.*

Proof. Let $R_i = q_i + 1$ and $c_s \equiv \tau_i - R_i + s$, $s = 1, 2, \ldots, R_i$. Define the $R_i \times (q_i + 1)$ alternant matrix X_i and the $1 \times (q_i + 1)$ vector H_i as, respectively,

$$X_i \equiv \begin{bmatrix} 1 & b_1(c_1) & \ldots & b_{q_i}(c_1) \\ 1 & b_1(c_2) & \ldots & b_{q_i}(c_2) \\ 1 & b_1(c_3) & \ldots & b_{q_i}(c_3) \\ \ldots & \ldots & \ldots & \ldots \\ 1 & b_1(c_{\tau_i}) & \ldots & b_{q_i}(c_{\tau_i}) \end{bmatrix}, H_i \equiv \begin{bmatrix} 1 & b_1(\tau+h) & \ldots & b_{q_i+1}(\tau+h) \end{bmatrix}. \tag{14}$$

The OLS coefficients from regressing $y_i = (y_{i\tau_i - R_i + 1}, \ldots, y_{i\tau_i})$ on $b_k(t)$ are given by

$$\hat{\alpha}^{(q_i, R_i)} = (X_i' X_i)^{-1} X_i' y_i,$$

[10] Laderman and Laderman (1982) derive a similar expression in the context of forecasting a time series by polynomial regression using the entire available time series.

so that the R_i forecast weights are

$$w_i = H_i \left(X_i' X_i\right)^{-1} X_i'. \tag{15}$$

Since X_i is a Vandermonde matrix with the first column being a column of ones (by assumption), it follows that $X_i e_1 = \iota$. Then

$$\left(X_i' X_i\right)^{-1} X_i' \iota = e_1 \equiv \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

so that $w_i \iota = 1$, where ι is the $(q_i + 1) \times 1$ vector of ones. This proves the statement. □

3.2.2 Deterministic trends DGPs

In this section we consider an expanded class of DGPs that is appropriate for applications where: 1) there is more than one pre-treatment outcome; 2) it makes sense to model the outcomes as trending over time; 3) the trend is deterministic rather than (or in addition to) stochastic. We show that the basis function regression considered in the previous section gives forecasts of the counterfactuals that remain unbiased for the expanded class of DGPs, under certain conditions.

The expanded class of DGPs always includes a deterministic trend component, possibly in addition to (either or both) the stochastic components considered in Assumption 3.

Assumption 4 (Deterministic trend DGPs)**.** *For $j = 1, 2$, let $I_j \in \{0, 1\}$ be a non-random binary indicator. The potential outcome $y_{it}(0)$ follows the process:*

$$y_{it}(0) = I_1 y_{it}^{(1)}(0) + I_2 y_{it}^{(2)}(0) + y_{it}^{(3)}(0), \tag{16}$$

where $y_{it}^{(1)}(0)$ and $y_{it}^{(2)}(0)$ are as in Assumption 3 and $y_{it}^{(3)}(0)$ is a deterministic time

trend such that $y_{it}^{(3)}(0) = \sum_{k=0}^{q_{0i}} c_{ik}^{(3)} b_k(t)$ *with* $c_i^{(3)} \in \mathbb{R}^{q_{0i}+1}$ *and known basis functions* $\{b_k(t)\}_{k=0}^{q_{0i}}$, $q_{0i} \in \{0, 1, 2, \ldots, \tau - 1\}$.

Theorem 2 below clarifies when forecasts obtained via basis function regressions satisfy the unbiasedness assumption (Assumption 1) in the presence of deterministic trends.

Theorem 2 (Unbiasedness for deterministic trend DGPs)**.** *Let Assumption 4 hold. Then,*

$$\mathbb{E}\left[\widehat{y}_{i\tau+h}^{(q_i,R_i)}(0) - y_{i\tau+h}(0)\right] = 0,$$

where $\widehat{y}_{i\tau+h}^{(q_i,R_i)}(0)$ *is defined in* (11)*, with* $q_i \geq q_{0i}$.

Proof. It is sufficient to show that for each component $y_{it}^{(r)}(0)$, $r \in \{1, 2, 3\}$,

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_i} w_{it}^{(q_i,R_i)}\, y_{it}^{(r)} - y_{i\tau+h}^{(r)}(0)\right] = 0. \tag{17}$$

For the mean stationary component ($r = 1$) and the random walk component ($r = 2$) we have $\mathbb{E}\left(y_{it}^{(r)} - y_{i\tau+h}^{(r)}(0)\right) = 0$. Multiplying this equation by $w_{it}^{(q_i,R_i)}$, summing over $t \in \mathcal{T}_i$, and using the fact that the non-random weights sum to 1 by Lemma 2, we obtain (17) for $r = 1$ and $r = 2$. To show (17) for the deterministic time trend component ($r = 3$), note that by (8), $\sum_{t\in\mathcal{T}_i} w_{it}^{(q_i,R_i)}\, y_{it}^{(3)} = \sum_{k=0}^{q_i} \widetilde{c}_{ik}^{(q_i,R_i)}\, (\tau+h)^k$, where

$$\widetilde{c}_i^{(q_i,R_i)} := \underset{c\in\mathbb{R}^{q_i+1}}{\operatorname{argmin}} \sum_{t\in\mathcal{T}_i} \left(y_{it}^{(3)} - \sum_{k=0}^{q_i} c_k\, b_k(t)\right)^2.$$

Since $q_i \geq q_{0i}$ for all i, the objective function in the last display is minimized (with value zero) at $\widetilde{c}_{ik}^{(q_i,R_i)} = c_{ik}^{(3)}$, which implies $y_{i\tau+h}^{(3)}(0) = \sum_{t\in\mathcal{T}_i} w_{it}^{(q_i,R_i)}\, y_{it}^{(3)}$, that is, (17) holds for $r = 3$ even without taking the expectation. □

Remark 5. While the stochastic components of the counterfactual in Assumption 4 are unobserved, the deterministic time trend component is assumed to be a func-

tion of the same basis functions used to obtain the forecast. This implies that the stochastic component of the data-generating process does not need to be correctly specified, but, if a deterministic trend component is present, the true basis functions must be used in estimation. This is an unusual result from the perspective of forecasting, where one typically assumes correct specification of both stochastic and deterministic parts of a model. In the next section, we discuss how the assumption of correct specification for the deterministic component could in principle be relaxed.

Remark 6. The key requirement of Theorem 2 is that q_i, the number of basis functions used in estimation, be at least q_{0i}, the true number of basis functions in the DGP. Intuitively, this means that choosing a too small number of basis functions runs the risk of delivering biased forecasts of the counterfactuals. We discuss the practical implications of this finding in the next section.

3.2.3 Choice of basis functions and tuning parameters

Our proposed method for forecasting counterfactuals in Definition 1 requires choosing: 1) the type of basis functions $b_k(t)$; 2) the number of basis functions q_i; and 3) the number R_i of pre-treatment periods used for the estimation. We discuss the tradeoffs that these choices create and offer some practical recommendations for empirical researchers.

Regarding the choice of basis functions, an implication of Theorems 1 and 2 is that this choice only matters when one is certain that the DGP has a deterministic time trend, in which case the basis functions need to be correctly specified to ensure unbiasedness (up to the order, which only needs to be larger than the true one). When the DGP is possibly stationary or possibly has a stochastic trend, the choice of basis functions does not matter for unbiasedness.

Basis functions may be chosen based on the time series properties of pre-treatment outcomes. For example, periodicity could be captured by Fourier basis functions.

Polynomial time trends appear to be a natural choice of basis functions for DGPs with deterministic trends. In DiD models it is typical to assume the presence of time trends (mostly linear) that are common between control and treatment groups.

Our results make it clear that a linear time trend can only be dealt with by either using a control group (which leads to standard DiD) or, when a control group is not available, by using (at least two) pre-treatment time periods to model the trend (which leads to our polynomial regression).

An advantage of polynomial basis functions is that it is in principle possible to relax the assumption that the basis functions are correctly specified, assuming instead that $y_{it}^{(3)}(0)$ in Assumption 4 is a continuous (but unknown) function of time. Since time in our setting is defined on a compact interval and the deterministic trend is a continuous function, $y_{it}^{(3)}(0)$ can be approximated arbitrarily well by a polynomial in time. In fact, by the Weierstrass Approximation Theorem, the approximation error approaches zero as the order of the polynomial goes to infinity. Under additional smoothness assumptions on the deterministic trend, an approximation theorem could then be used (e.g. the Polynomial Approximation Error Theorem) to derive a bound on the approximation error of $y_{it}^{(3)}(0)$ by the polynomial regression. The forecast of $y_{i\tau+h}(0)$ is biased, but we conjecture that it may be possible to do bias-correction given an expression for the bias obtained via the approximation theorem.

Our practical recommendation for empirical researchers is thus to consider polynomial basis functions, obtained by letting $b_k(t) = t^k$ in Definition 1. This is also what we focus on in the remainder of the paper.

Regarding the choice of number of basis functions q_i used for the estimation, again this only matters if the DGP has a deterministic trend component. For DGPs without a deterministic trend component, any q_i ensures unbiasedness. In one of the simulations in Section 6 we investigate the finite-sample bias and variance of the estimator for these DGPs, and find that the choice of a $q_i > 0$ has small costs in terms of variance but can help control the bias resulting from a non-stationary initial condition. If the DGP has a deterministic trend component, then q_i cannot be smaller than the true order of this trend. The true order is unknown and cannot be consistently estimated, so a trade-off emerges where a large q_i ensures unbiasedness but comes at the cost of higher variance. Cross-validation methods on pre-treatment data cannot help choose q_i if they only target accuracy, because of the necessity to ensure unbiasedness in our context. It may be possible to devise bias-correction

within cross-validation methods to select q_i, but we leave this endeavour for future work.

Our practical recommendation is to report results for a small range of values for q_i, e.g., $q_i = 0, 1, 2, 3$. In addition, plotting the time series of pre-treatment observations can provide some informal guidance on how to choose this range (for example, if the pre-treatment outcomes display a linear time trend, a polynomial of order 0 can be ruled out as it is likely to deliver a biased estimator; if the pre-treatment outcomes appear stationary, any choice of q_i is valid).

Regarding the choice of pre-treatment outcomes used for the estimation, R_i, there are several factors at play. Under mean stationarity, it makes sense to choose R_i as large as possible. On the other hand, a short R_i can guard against violation of stationarity due to parameter change. In one of the simulations in Section 6 we show that it may be advisable to select a short R_i even in the absence of parameter change. Our practical recommendation is thus to select short estimation windows, for example by letting $R_i = q_i + 1$.

3.3 Individual-specific treatment time

Note that the treatment timing τ can be individual-specific as long as it is exogenous to the potential outcomes. In this sense our approach applies to a staggered adoption setting with exogenous treatment timing. Additionally, it is possible to allow for the presence of treatment anticipation, as long as it is limited. In this case, one simply modifies the pre-treatment estimation window $\mathcal{T}_i$ in Definition 1 to include observations only up to the time $\tau - \delta_i$ at which it is still reasonable to assume that there was no treatment anticipation, that is, $\mathcal{T}_i \equiv \{\tau - \delta_i - R_i + 1, \ldots, \tau - \delta_i\}$ (and h is adjusted accordingly).

3.4 Balanced panel and pooled estimation

Assume that $R = R_i$ and $q = q_i$ are constant across i and focus on polynomial basis functions in Definition 1. The first alternative way to obtain an estimator for the ATT in our baseline setting is to consider the cross-sectional averages $\overline{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_{it}$ of

the observed outcomes in time period t. Due to linearity of the forecasting procedure, we can rewrite $\widehat{\text{FAT}}_h$ as as

$$\widehat{\text{FAT}}_h = \overline{y}_{\tau+h} - \sum_{k=0}^{q} \overline{\alpha}_k \, (\tau+h)^k, \qquad \overline{\alpha} := \underset{\alpha \in \mathbb{R}^{q+1}}{\text{argmin}} \sum_{t \in \mathcal{T}} \left(\overline{y}_t - \sum_{k=0}^{q} \alpha_k \, t^k \right)^2, \tag{18}$$

where $\mathcal{T} = \{\tau - R + 1, \ldots, \tau\}$, and we suppress the dependence on τ, q, R. Here, the cross-sectional averages for $t \leq \tau$ are used to obtain a forecast of the average counterfactual for $t = \tau + h$, which is then subtracted from the cross-sectional average observed at that time period.

The second alternative is to consider a pooled regression estimator, namely $\widehat{\text{FAT}}_h = \widehat{\beta}_h$, where

$$\left(\widehat{\beta}, \widehat{\alpha}\right) = \underset{\left\{\beta \in \mathbb{R}^h, \, \alpha \in \mathbb{R}^{n \times (q+1)}\right\}}{\text{argmin}} \sum_{i=1}^{n} \sum_{t=\tau-R+1}^{\tau+h} \left(y_{it} - \sum_{k=1}^{h} 1\{t = \tau + k\} \, \beta_k - \sum_{k=0}^{q} \alpha_{ik} \, t^k \right)^2, \tag{19}$$

which is the OLS estimator obtained from regressing y_{it} on a set of time dummies $1(t = \tau + k)$, for $k \in \{1, \ldots, h\}$, and individual-specific time trends.[11]

The alternative estimation strategies in (18) and (19) provide algebraically identical treatment effect estimates in the case of our baseline setting with $R = R_i$ and $q = q_i$. In a more general setting, however, it is possible to show that these alternative estimation strategies do *not* give the same treatment effect estimator, and may indeed give inconsistent estimates for ATT if applied incorrectly.

4 Extension: no control group, covariates

This section considers how one could incorporate covariates, including lagged outcomes, in the estimation of the FAT. We show results that prove consistency and

[11] It actually does not matter for $\widehat{\beta}$ here whether we make the coefficients α on the time trend individual-specific or not.

asymptotic normality of $\widehat{\text{FAT}}_h$. We focus throughout on the case of polynomial time trends.

4.1 Homogeneous coefficients

Suppose one considers a linear model for $y_{it}(0)$, with an individual-specific polynomial time trend of order q_i and homogeneous coefficients for the covariates:

$$y_{it}(0) = x_{it}'\,\beta + \sum_{k=0}^{q_i} c_{ik} t^k + \varepsilon_{it}, \tag{20}$$

where $x_{it} \in \mathbb{R}^{\dim x_{it}}$ is a vector of covariates (possibly including lagged outcomes), $\beta \in \mathbb{R}^{\dim x_{it}}$ and $c_{ik} \in \mathbb{R}$ are unknown parameters and $\varepsilon_{it} \in \mathbb{R}$ is such that

$$\mathbb{E}\left[\varepsilon_{it} \,\middle|\, x_{it}, x_{it-1}, \ldots, \varepsilon_{it-1}, \varepsilon_{it-2}\right] = 0. \tag{21}$$

Assume that we have estimates $\widehat{\beta}$ for the common parameters β that are consistent as $n \to \infty$ under correct model specification.[12]

Model (20) can be used to forecast the individual counterfactuals as follows:

$$\widehat{y}_{i\tau+h}^{(q_i,R_i)}(\widehat{\beta}) := x_{i\tau+h}'\,\widehat{\beta} + \sum_{k=0}^{q_i} (\tau+h)^k\, \widehat{c}_{ik}^{(q_i R_i)}(\widehat{\beta}), \tag{22}$$

$$\widehat{c}_i^{(q_i,R_i)}(\widehat{\beta}) := \underset{c \in \mathbb{R}^{q_i+1}}{\operatorname{argmin}} \sum_{t \in \mathcal{T}_i} \left(y_{it} - x_{it}'\,\widehat{\beta} - \sum_{k=0}^{q_i} t^k\, c_k \right)^2, \tag{23}$$

where $c_i = (c_{i,0}, \ldots, c_{i,q_i})$ is a $q_i + 1$ vector, and $\mathcal{T}_i = \{\tau - R_i + 1, \ldots, \tau\}$ is the set of the R_i time periods directly preceding the treatment date. The parameter

[12] For example, when $q_i = 0$ and $x_{it} = (y_{it-1}, z_{it}')'$, a consistent estimator for $\beta = (\rho, \theta')'$ can be obtained by applying an IV regression to the first-differenced model

$$y_{it} - y_{it-1} = [y_{it-1} - y_{it-2}]\;\rho + [z_{it} - z_{it-1}]'\,\theta + \varepsilon_{it} - \varepsilon_{it-1},$$

using, for example, y_{it-2} and z_{it-1} as instruments. In the Monte Carlo simulations we further extend this case to $q_i = 1$.

$R_i \in \{q_i + 1, \ldots, \tau\}$ is chosen by the researcher.

Once the individual-specific forecasts are obtained, the forecasted average treatment effect estimator is given by

$$\widehat{\text{FAT}}_h^{\text{MB}} = \frac{1}{n} \sum_{i=1}^{n} \left[y_{i\tau+h} - \widehat{y}_{i\tau+h}^{(q_i,R_i)}(\widehat{\beta}) \right]. \tag{24}$$

Here, the superscript MB refers to model-based.

Theorem 3 in the Appendix derives sufficient conditions for the consistency and asymptotic normality of $\widehat{\text{FAT}}_h^{\text{MB}}$. It is easy to verify that the same result of consistency and asymptotic normality of the model-based estimator can be obtained if one assumes that the vector process for the counterfactual outcomes and covariates can be written as the sum of potentially three components: a mean stationary process, a random walk and a polynomial time trend of order no greater than the order q_i used for the estimation.

4.2 Heterogeneous coefficients

In this section we discuss some examples of models with heterogeneous coefficients for the covariates for which one can obtain unbiased estimators of the counterfactuals.

If the model for the counterfactuals is an AR(p) with heterogeneous parameters, for example, the time series literature (e.g., Fuller and Hasza (1980), Dufour (1984), Magnus and Pesaran (1991)), has derived conditions under which forecasts from an individual AR(p) model are unbiased. The maintained assumptions are stationarity of the initial condition and symmetry of the error term.

A second example is that of strictly exogenous regressors with heterogeneous coefficients. For example, suppose that the $h = 1$ period forecast of $y_{i\tau+1}(0)$ is given

by

$$\widehat{y}_{i\tau+1}^{(q_i,R_i)} := \sum_{k=0}^{q_i} \widehat{c}_k^{(q_i,R_i)} \, (\tau+1)^{\,k} + \hat{\beta}^{(i)} x_{i\tau+1}, \tag{25}$$

$$\left(\widehat{c}^{(q_i,R_i)}, \hat{\beta}^{(i)}\right) := \underset{\alpha\in\mathbb{R}^{q_i+1},\beta\in\mathbb{R}^x}{\operatorname{argmin}} \sum_{t\in\mathcal{T}_i} \left(y_{it} - \sum_{k=0}^{q_i} c_k \, t^k - \beta x_{it}\right)^2. \tag{26}$$

The forecast (26) is unbiased provided that a Vandermonde matrix which includes functions of $(\tau - R_i + s)^j$, $j = 0, \ldots, q_i$, $s = 1, 2, \ldots, R_i$ and the covariates $x_{i\tau-R_i+s}$, $s = 1, 2, \ldots, R_i$, is invertible. This invertibility condition imposes constraints on how the covariates can change over time.

5 Extension: control group

In this section we discuss how to modify our baseline procedure when a group of individuals not exposed to the treatment is available.

Without a control group, Section 3 derived sufficient conditions ensuring that FAT_h defined in (4) is a consistent and asymptotically normal estimator of ATT_h defined in (2). These conditions are the ability to obtain forecasts of the counterfactuals using pre-treatment data that are on average unbiased (Assumption 1) and the validity of a central limit theorem (Assumption 2). As discussed above, these conditions exclude the presence of time effects such as macro shocks that affect all individuals between times τ and $\tau + h$, $h \geq 1$, and that are unforecastable using pre-treatment data. The presence of a control group allows us to weaken this assumption.

Suppose that all individuals are untreated before the implementation of the treatment at time τ and that some individuals remain untreated after τ. Let $D_i = 1$ if individual i is untreated before and after τ. The observed outcome of individual i at time t is then

$$y_{it} = D_i \left[1\left(t \leq \tau\right) y_{it}\left(0\right) + 1\left(t > \tau\right) y_{it}\left(1\right)\right] + \left(1 - D_i\right) y_{it}\left(0\right). \tag{27}$$

As before, the parameter of interest is the average treatment effect on the treated h periods after the implementation of the treatment:

$$\mathrm{ATT}_h = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(y_{i\tau+h}\left(1\right) - y_{i\tau+h}\left(0\right)\middle|\, D_i = 1\right). \tag{28}$$

Our proposed estimator is defined as:

$$\widehat{\mathrm{DFAT}}_h = \frac{1}{n_1}\sum_{i:D_i=1}\left(y_{i\tau+h} - \widehat{y}_{i\tau+h}(0)\right) - \frac{1}{n_0}\sum_{i:D_i=0}\left(y_{i\tau+h} - \widehat{y}_{i\tau+h}(0)\right), \tag{29}$$

where n_1 is the number of treated individuals at time $\tau + h$, n_0 is the number of control individuals at time $\tau + h$, and $y_{i\tau+h}$ is the observed outcome at $\tau + h$ given by (27).

Note that under (30) below, $\mathbb{E}(\widehat{\mathrm{DFAT}}_h) = \mathrm{ATT}_h$ in (28):

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(y_{i\tau+h}\left(0\right) - \widehat{y}_{i\tau+h}(0)\middle|\, D_i = 1\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(y_{i\tau+h}\left(0\right) - \widehat{y}_{i\tau+h}(0)\middle|\, D_i = 0\right). \tag{30}$$

Unlike in the baseline case, the forecast $\widehat{y}_{i\tau+h}(0)$ can be biased, as long as the average bias for the treated group equals the average bias for the control group. As a consequence, the DGP for $y_{it}(0)$ can contain additive time effects that are common across individuals such as additive macro shocks that affect both treated and control groups in the same way.

The presence of a control group allows us to substitute Assumption 4 to allow for a common shock that is not necessarily polynomial, e.g., $y_{it}\left(0\right) = \tilde{y}_{it}(0) + \gamma_t$, where $\tilde{y}_{it}(0)$ satisfies Assumption 4. Then, under assumptions similar to Assumptions 1 and 2, it is possible to show consistency and asymptotic normality of (29).

As in Section 3, we suggest using $\widehat{y}^{(q_i,R_i)}_{i\tau+h}(0)$ as an estimator for $\widehat{y}_{i\tau+h}(0)$ in (29). Here, the parameters q_i, R_i do not necessarily have to be the same for the treated and control units.

5.1 Comparison with Difference-in-Differences

Despite the apparent similarity with the difference-in-differences (DiD) framework, our method in the presence of a control group allows for DGPs for potential outcomes with more general forms of latent heterogeneity. For example, our approach allows for potential outcomes that follow fully heterogeneous autoregressive processes and/or unit root processes. In addition, it allows for the DGPs to have additive individual-specific time trends, as long as the deterministic time trend is either known or can be approximated by, e.g., a polynomial.

To see this, consider for example the following data-generating process for the potential outcomes:

$$y_{it}(0) = \rho y_{it-1}(0) + \gamma_t + k_i t + \epsilon_{it}, \mathbb{E}(\epsilon_{it}) = 0,$$

where $\rho \in [0,1]$, γ_t is a common shock, k_i is an individual-specific time trend coefficient. DiD can accommodate such specification as long as the assumption of parallel-paths holds, which requires restricting the heterogeneity of both the initial condition of the process, i.e., $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(y_{i0}(0)|D_i = 1) = \frac{1}{n}\sum_{i=1}^{n}E(y_{i0}(0)|D_i = 0)$ and the time trend coefficients, i.e., $\frac{1}{n}\sum_{i=1}^{n}k_i I(D_i = 1) = \frac{1}{n}\sum_{i=1}^{n}k_i I(D_i = 0)$, where $I(\cdot)$ is the indicator function. In contrast, DFAT_h does not require restricting the unobserved individual heterogeneity, and allows for fully heterogeneous parameters k_i. In addition, it is straightforward to include lagged pre-treatment covariates with a homogeneous autoregressive prarameter or a heterogeneous one ρ_i, which is not a possibility for difference-in-differences methods.

6 Simulation study

Throughout this section, we set $b_k(t) = t^k$ in Definition 1, so that

$$\widehat{y}^{\mathrm{PR}}_{i\tau+h} := \sum_{k=0}^{q_i} \widehat{c}^{(q_i,R_i)}_{ik} (\tau+h)^k,$$
$$\widehat{c}_i^{(q_i,R_i)} := \operatorname*{argmin}_{c\in\mathbb{R}^{q_i+1}} \sum_{t\in\mathcal{T}_i} \left(y_{it} - \sum_{k=0}^{q_i} c_k\, t^k \right)^2,$$

where we suppressed the dependence on (q_i, R_i) of $\widehat{y}^{\mathrm{PR}}_{i\tau+h}$.

We refer to the associated estimator as the *polynomial-regression* FAT:

$$\widehat{\mathrm{FAT}}^{\mathrm{PR}}_h := \frac{1}{n}\sum_{i=1}^{n} \left(y_{i\tau+h} - \widehat{y}^{\mathrm{PR}}_{i\tau+h} \right). \tag{31}$$

6.1 Polynomial-regression versus model-based FAT under misspecification

In this section, we compare the performance of the polynomial-regression estimator (31) to that of the model-based estimator (24), both under correct specification and under misspecification.

The DGP we consider here specifies the potential outcome for $i = 1, \ldots, N$ as:

$$\begin{aligned}
y_{it}(0) &= y^{(1)}_{it}(0) + y^{(3)}_{it}(0),\ t = 1, \ldots, T, \\
y^{(1)}_{it}(0) &= \mu_i + \rho y_{it-1} + u_{it},\ t \geq 1 \\
y^{(1)}_{i0}(0) &\sim \mathcal{N}(1, 2), \\
y^{(3)}_{it}(0) &= \delta_i t, \\
\mu_i &\sim \mathcal{U}[-1, 1],\ u_{it} \sim \mathcal{N}(0, 1), \\
\rho &\in \{0.2, 0.9\},\ \delta_i = 1,
\end{aligned}$$

where $y^{(1)}_{it}(0)$ is an autoregressive process with the initial condition *not* drawn from the stationary distribution, and $y^{(3)}_{it}(0)$ is a deterministic linear time trend with

homogeneous coefficients.[13] Here, $y_{it}(1) = y_{it}(0)$ at $t = \tau + 1$, so $\text{ATT}_{i\tau+1} = 0$.

We consider the case of a balanced panel with $T = 6$ periods, with $\tau = 5$, so that the first 5 periods are the "pre-treatment," and the last period is the "post-treatment" period. Hence, $h = 1$. We focus on the $h = 1$ case since most applications focus on the ATT one period after the implementation of the treatment. We show results for two sample sizes $N \in \{50, 1000\}$.

The polynomial-regression estimator (31) is computed as described in Section 3. For each i, we regress $\{y_{it}\}_{t=1}^{5}$ on a polynomial in t of order $q \leq T - 2$, and then we compute the forecast $\hat{y}_{i6}^{\text{PR}}(0) = \sum_{k=0}^{q} \hat{c}_{ik}^{(q,q+1)} 6^k$. We then use this individual forecast to compute $\widehat{\text{FAT}}_1^{\text{PR}}$ as in (31).

The model-based estimator (24) is computed as described in Section 4.1, where the common AR parameter ρ is first estimated via Anderson-Hsiao. We estimate ρ in two different ways. We use (1) y_{it-3} as an instrument when the linear time trend is correctly accounted for, and (2) y_{it-2} as an instrument when the linear time trend is not accounted for. Given the estimate $\hat{\rho}$, for each i, we then regress $\{y_{it} - \hat{\rho} y_{it-1}\}_{t=1}^{5}$ on a polynomial in t of order $q \leq T - 2$, and then compute the forecast $\hat{y}_{i6}^{\text{MB}}(0) = \hat{\rho} y_{i5} + \sum_{k=0}^{q} \hat{c}_{ik}^{(q,q+1)} 6^k$. $\widehat{\text{FAT}}_1^{\text{MB}}$ is then computed as in (24).

The set-up described in this subsection is a misspecification study: the initial condition $y_{i0}^{(1)}(0)$ is not drawn from the stationary distribution so that Assumption 4 does not hold (i.e., the DGP for the polynomial-regression FAT is misspecified), and the linear time trend is not accounted for when the model-based estimator uses the incorrect instruments for the computation of ρ in the first step.

Table 1 shows the bias and the standard error of the polynomial-regression FAT and of the model-based estimator that is correctly specified (MB) and that is misspecified by using the incorrect instruments in the estimation of the autoregressive parameter (MB missp.). The results show that the model-based estimator is sensitive to model specification and that it does not outperform the polynomial-regression estimator, even under correct model specification. The latter happens because of es-

[13] Note that here μ_i is not correlated with the initial condition. Additional Monte Carlo results, available upon request, show similar findings when the fixed effects are correlated with the initial condition.

$\rho = 0.2$		$q = 0$	$q = 1$	$q = 2$	$q = 3$
$N = 50$	PR	1.25	0.02	0.003	0.01
		(0.19)	(0.31)	(0.54)	(0.99)
	MB	1.25	0.01	0.02	0.03
		(1.26)	(0.37)	(0.54)	(0.89)
	MB	41.95	−0.86	−0.14	−0.47
	missp.	(1292.09)	(22.45)	(2.13)	(6.77)
$N = 1000$	PR	1.25	0.001	0	0.01
		(0.04)	(0.07)	(0.12)	(0.24)
	MB	1.00	0.001	−0.01	0.003
		(0.19)	(0.08)	(0.14)	(0.27)
	MB	0.23	−0.003	0	−0.02
	missp.	(0.1)	(0.11)	(0.21)	(0.41)
$\rho = 0.9$					
$N = 50$	PR	4.69	0.61	−0.06	0.01
		(0.16)	(0.21)	(0.36)	(0.65)
	MB	1.03	−0.04	−0.17	−0.25
		(3.38)	(0.69)	(0.75)	(1.42)
	MB	186.7	2.36	0.2	−0.42
	missp.	(5905.57)	(86.1)	(10.98)	(15.9)
$N = 1000$	PR	4.69	0.59	−0.06	0.01
		(0.03)	(0.05)	(0.08)	(0.15)
	MB	1.03	−0.003	−0.01	−0.01
		(0.43)	(0.14)	(0.14)	(0.28)
	MB	0.96	−4.58	0.12	−4.21
	missp.	(45.95)	(157.56)	(2.27)	(135.57)

Table 1: Bias and standard error (in parentheses) for the polynomial-regression FAT (PR), the model-based FAT that takes account of the linear time trend and uses the correct instruments in the first step (MB), and the model-based FAT that uses the incorrect instrument in the first step (MB missp.). The results are presented across different polynomial orders q and sample sizes N.

timation error in $\widehat{\rho}$, which induces more bias the more persistent is the process.

6.2 Choice of tuning parameters for polynomial-regression FAT

In this section, we compare the finite-sample performance of the polynomial-regression FAT estimator across different tuning parameters: the polynomial order, q, and the estimation window, R (the pre-treatment periods used in the polynomial regression) for different specifications of the DGP. All specifications satisfy Assumption 4, where

the process for the potential outcome is specified as the sum of up to three different components. That is, for each $i = 1, \ldots, N$:

$$
\begin{aligned}
y_{it}(0) &= I_1 y_{it}^{(1)}(0) + I_2 y_{it}^{(2)}(0) + I_3 y_{it}^{(3)}(0),\ t = 1, \ldots, T, \\
y_{it}^{(1)}(0) &= \mu_i + \rho y_{it-1}^{(1)}(0) + u_{it},\ t \geq 1, \\
y_{i0}^{(1)}(0) &\sim \mathcal{N}\left(\frac{\mu_i}{1-\rho}, \frac{1}{1-\rho^2}\right), \\
y_{it}^{(2)}(0) &= y_{it-1}^{(2)}(0) + \epsilon_{it},\ t \geq 1, \\
y_{i0}^{(2)}(0) &= 0, \\
y_{it}^{(3)}(0) &= \delta_i t, \\
\mu_i &\sim \mathcal{U}\left[-1, 1\right],\ u_{it} \sim \mathcal{N}\left(0, 1\right),\ \epsilon_{it} \sim \mathcal{N}(0, 1), \\
\rho &= 0.2,\ \delta_i = 1,
\end{aligned}
$$

and $T = 6,\ \tau = 5,\ h = 1,\ N = 1000$.

Note that the initial observation, $y_{i0}^{(1)}(0)$, is drawn from the stationary distribution of the AR(1) process $y_{it}^{(1)}(0)$, and that the time trend component, $y_{it}^{(3)}(0)$, is linear and homogeneous across individuals, so that it can be interpreted as a common shock.

Table 2 shows results for the bias and standard error of $\widehat{\text{FAT}}_1^{\text{PR}}$ across different tuning parameters. The table shows that when the potential outcome process is mean stationary (first panel) or when it is the sum of a mean stationary and a random walk (second panel), the estimator is robust to the choice of tuning parameters, in the sense that the bias and standard error of the estimator do not vary across different values of the tuning parameters. When the potential outcome process contains a linear time trend component we observe bias when the polynomial-order q is less than the true order of the time trend, that is, for $q = 0$. In this case, however, a smaller estimation window R obtains a smaller bias. When $q \geq 1$, the performance of the estimator in terms of bias is again robust to the choice of tuning parameters (with decreasing standard errors for increasing values of R).

		R				
		$q+1$	$q+2$	$q+3$	$q+4$	$q+5$
Stationary AR(1) $I_1 = 1, I_2 = 0 = I_3$	$q = 0$					
	bias	-0.0002	-0.0005	-0.0003	-0.0001	0.0005
	s.e.	0.0397	0.036	0.0354	0.0346	0.0341
	$q = 1$					
	bias	0.0047	0.0003	0.0009	0.0008	
	s.e.	0.0709	0.0565	0.0476	0.0448	
	$q = 2$					
	bias	0.0112	0.0023	0.0015		
	s.e.	0.1225	0.0907	0.0726		
Stationary AR(1) + unit root $I_1 = 1 = I_2, I_3 = 0$	$q = 0$					
	bias	-0.0029	-0.0041	-0.0045	-0.005	-0.005
	s.e.	0.0516	0.0512	0.0525	0.0547	0.0577
	$q = 1$					
	bias	-0.0004	-0.0023	-0.0023	-0.0033	
	s.e.	0.082	0.0664	0.0625	0.0606	
	$q = 2$					
	bias	0.0025	-0.0011	-0.0002		
	s.e.	0.1454	0.0997	0.0868		
Stationary AR(1) + linear trend $I_1 = 1 = I_3, I_2 = 0$	$q = 0$					
	bias	0.9998	1.4995	1.9997	2.4999	3.0005
	s.e.	0.0397	0.036	0.0354	0.0346	0.0341
	$q = 1$					
	bias	-0.0027	-0.0024	-0.0008	-0.001	
	s.e.	0.068	0.0536	0.0466	0.0442	
	$q = 2$					
	bias	-0.0032	-0.0051	-0.0022		
	s.e.	0.1225	0.0839	0.0698		
Stationary AR(1) + linear trend + unit root $I_1 = I_2 = I_3 = 1$	$q = 0$					
	bias	0.9971	1.4959	1.9955	2.4950	2.9950
	s.e.	0.0516	0.0512	0.0525	0.0547	0.577
	$q = 1$					
	bias	0.0005	-0.0015	0.001	0.0014	
	s.e.	0.0831	0.0659	0.0608	0.0621	
	$q = 2$					
	bias	0.001	-0.0047	-0.0018		
	s.e.	0.1447	0.1024	0.0873		

Table 2: Bias and standard error (s.e.) for the polynomial-regression FAT when the potential outcome is specified as indicated in the left-most column.

6.3 Heterogeneous coefficients

In this section, we compare the finite-sample behavior of the polynomial-regression FAT when the potential outcome process satisfies Assumption 4 with heterogeneous coefficients ρ_i and δ_i. That is, we consider the same specification of $y_{it}(0)$ as in the previous subsection with $I_1 = I_2 = I_3 = 1$, with the only changes being that ρ_i and δ_i vary across individuals.

Table 3 presents the results from specifying $\delta_i \sim \mathcal{U}[0, 2]$ with a homogeneous autoregressive parameter $\rho = 0.2$ (top panel) and with a heterogeneous autoregressive parameter $\rho_i \sim \mathcal{U}[0, 0.99]$ (bottom panel). We can see that the presence of heterogeneous parameters does not change the conclusions that we derived from Table 2.

7 Empirical illustrations

In this section, we replicate two analyses on the effects of different treatments. One analysis uses a standard difference-in-differences design and the other applies to a staggered adoption setting. We show that our approach can replicate the results, either numerically or qualitatively, of both empirical analyses.

Replication 1: Staggered adoption, overdose mortality and legalized medical cannabis laws

We use data from Shover et al. (2019), which analyzes the effect of legalized medical cannabis laws on opioid overdose mortality in the U.S. Shover et al. (2019) contributes to the debate about whether the adoption of such laws has decreased overdose mortality, see, e.g., Bachhuber et al. (2014). This is a staggered adoption setting: the unit of observation is at the level of state-year, with states slowly adopting legalized medical cannabis laws from 1999 to 2017. Our analysis includes 9 states that legalized medical cannabis before 2010 and 30 states that legalized medical cannabis between 2010 and 2017. The outcome of interest is the log mortality rate.

		R_i				
		$q+1$	$q+2$	$q+3$	$q+4$	$q+5$
Stationary AR(1) $I_1 = I_2 = I_3 = 1$ with $\delta_i \sim \mathcal{U}[0,2]$	$q=0$					
	bias	0.999	1.4987	2.0003	2.5012	3.0015
	s.e.	0.0544	0.0585	0.0655	0.0739	0.0826
	$q=1$					
	bias	0.0013	0.0038	0.0009	0	
	s.e.	0.0793	0.0666	0.0646	0.0635	
	$q=2$					
	bias	-0.0024	0.0065	0.0045		
	s.e.	0.1437	0.0971	0.0846		
Stationary AR(1) $I_1 = I_2 = I_3 = 1$ with $\delta_i \sim \mathcal{U}[0,2]$ and $\rho_i \sim \mathcal{U}[0,0.99]$	$q=0$					
	bias	0.9967	1.4961	1.9963	2.4958	2.9947
	s.e.	0.0554	0.0613	0.0693	0.078	0.0861
	$q=1$					
	bias	-0.0002	0.001	0.0027	0.0035	
	s.e.	0.0717	0.0629	0.0643	0.0672	
	$q=2$					
	bias	0.0019	-0.0023	-0.0005		
	s.e.	0.1264	0.0929	0.0801		

Table 3: Bias and standard error (s.e.) for the polynomial-regression FAT when the potential outcome is specified as in the left-most column. The time trend component is heterogeneous across individuals (top panel), with the addition of a cross-sectionally heterogeneous autoregressive component (bottom panel). Stationary initial condition for the AR(1) component for each i.

The original analyses in Bachhuber et al. (2014) and Shover et al. (2019) use a two-way fixed-effects estimator, which we know produces biased results in a staggered adoption setting, e.g. Goodman-Bacon (2021). We first redo the analysis to remove the bias of the original studies by using various methods, such as the staggered DiD approach of Callaway and Sant'Anna (2021) and the generalized synthetic control method of Xu (2017). The results can be found in Appendix B. We find an initial increase in overdose mortality and then a reversal, but neither is statistically significant.

We then implement FAT. Since this is a balanced panel, we start with a plot of the log mortality rate averaged across states as a function of time-to-adoption, see Figure 1, in order to get a sense for the time series properties of the outcome of interest. Given the apparent nonstationarity of the mortality rate, we choose the smallest possible estimation window for computing FAT as in (31). That is, we let $R = q + 1$. Our estimates for the ATT are stable across different polynomial orders and our results show a slight increase in mortality rates that is not statistically significant, see Figure 2. Our results corroborate those from the approaches of Callaway and Sant'Anna (2021) and Xu (2017).

Replication 2: Difference-in-Differences, refugees and far-right support

In this replication exercise, we use data from Dinas et al. (2019) which examines the relationship between refugee arrivals and support for the far right. Dinas et al. (2019) consider the case of Greece, and make use of the fact that some Greek islands (those close to the Turkish border) witnessed sudden and unexpected increases in the number of refugees during the summer of 2015, while other nearby Greek islands saw much more moderate inflows of refugees. The municipalities in the former Greek islands are considered treated, while the municipalities in the latter are considered control. The authors use a standard DiD analysis to assess whether the treated municipalites were more supportive of the far-right Golden Dawn party in the September 2015 general election. The original data set contains a total of 96 municipalities, 48

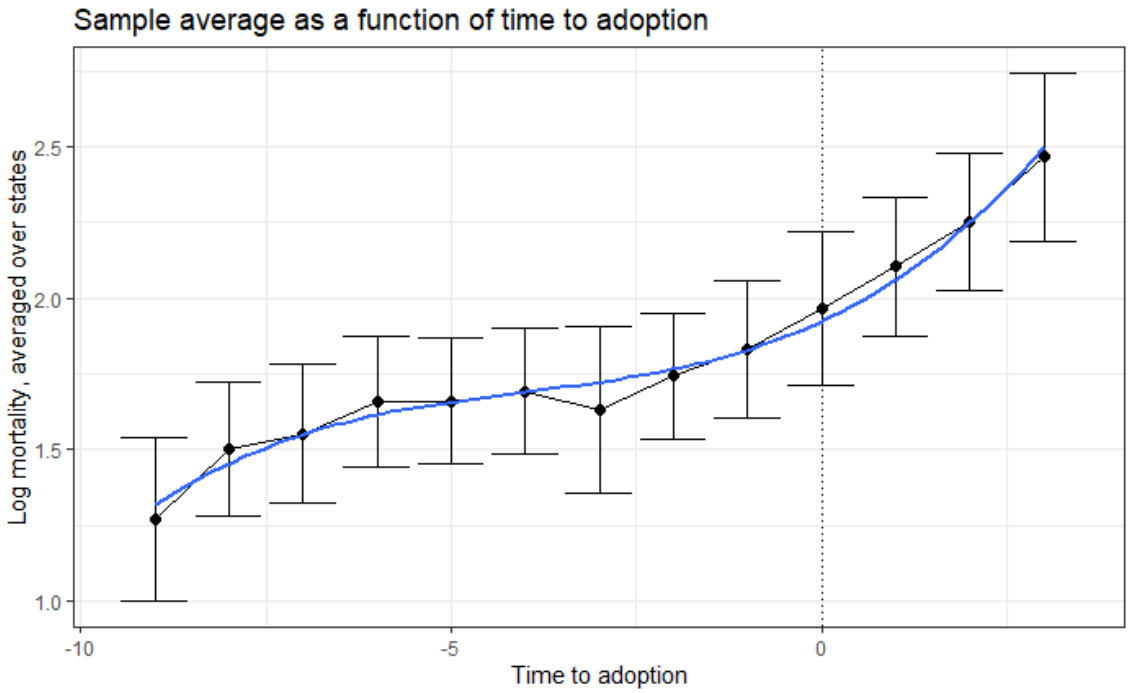

Figure 1: Log mortality rate averaged across states as a function of time-to-adoption. The blue line is a third order polynomial fit.

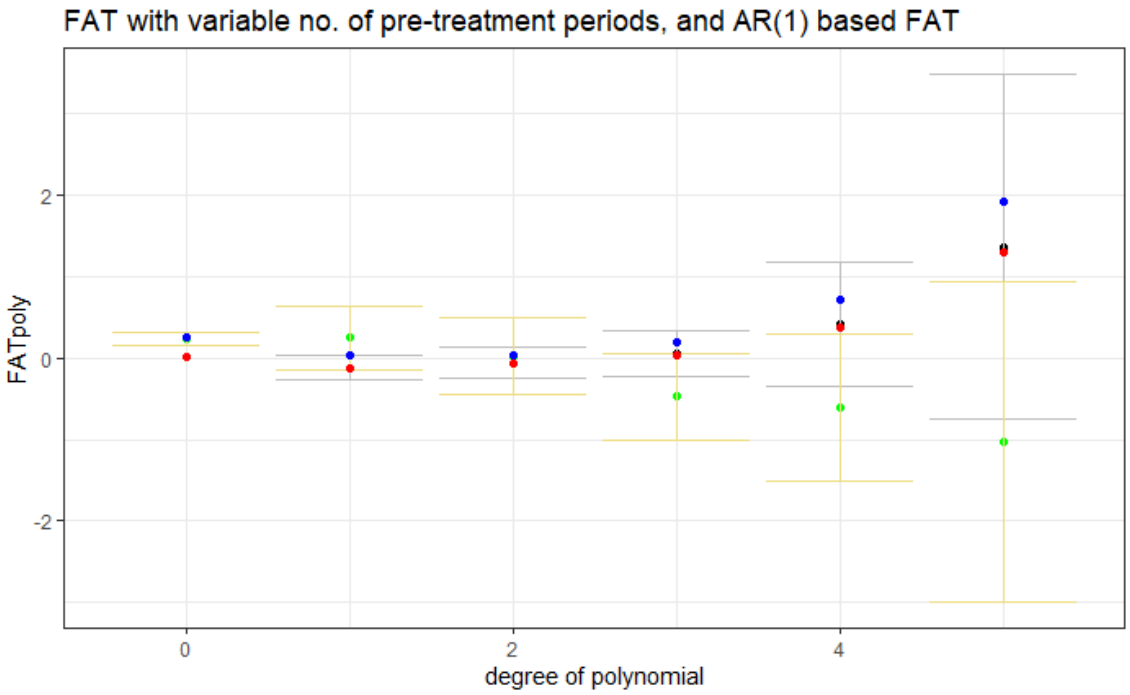

Figure 2: FAT with variable number of pre-treatment time periods. The red dots correspond to estimates of FAT for the early adopters, the blue dots to estimates of FAT for the late adopters, and the black dots to estimates of FAT for the entire sample. The gray intervals are the 95% confidence intervals corresponding to the black dots. The green dots are FAT estimates based on an AR(1) model; their corresponding 95% confidence intervals are shown in yellow.

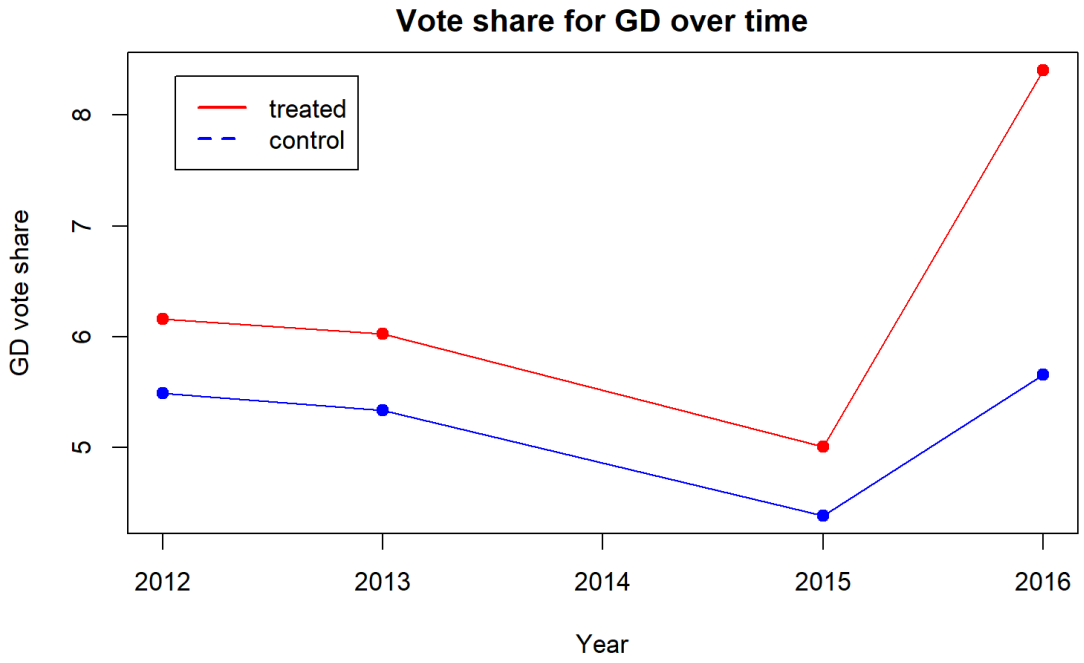

Figure 3: Vote share for Golden Dawn averaged across municipalities before and after 2015 for municipalities that were treated (red) and control (blue).

of which were treated, and data on four elections: three elections pre-treatment in 2012, 2013, 2015, and one post-treatment in 2016. The outcome of interest is the vote share for Golden Dawn (GD). Figure 3 shows the vote share for GD averaged across municipalities, treated and control, before and after the treatment time.

We use data on both the treated and the control municipalities to compute $\widehat{\text{DFAT}}_h$ with $h = 1$ and show that our estimate replicates the original DiD estimates. This application can be viewed as a "worst-case" scenario for our proposed estimator since the number of treated units is very small. We show results that use all three pre-treatment elections, in which case the order of the polynomial is $q_i = q \in \{0, 1, 2\}$, and results that use only the 2013 and 2015 pre-treatment elections, in which case the order of the polynomial is $q_i = q \in \{0, 1\}$. Note that we perform municipality-specific polynomial regressions to compute the forecasted vote share – the counterfactual outcome of interest, using the same polynomial order across all municipalities.

	FAT Treated	FAT Control	DFAT
Polynomial order		2013-2015	
$q = 0$	0.029 (0.016)	0.008 (0.009)	0.021
$q = 1$	0.054 (0.028)	0.032 (0.026)	0.022
		2012-2015	
$q = 0$	0.027 (0.012)	0.006 (0.011)	0.021
$q = 1$	0.038 (0.025)	0.017 (0.016)	0.019
$q = 2$	0.053 (0.036)	0.030 (0.027)	0.023

Table 4: DFAT under different polynomial orders and pre-treatment periods.

As Table 4 shows, our DFAT results are comparable with those in the original paper. The DiD estimates in the original paper are 0.0206 and 0.0208 when using 2013 and 2015 as pre-treatment periods and all pre-treatment periods, respectively. The two-way fixed-effects estimate is 0.021 with a standard error of 0.0393.

8 Conclusion

This paper proposed estimating average treatment effects (ATT) in the absence of a control group by forecasting individual counterfactuals using basis function regressions over a (short) time series of pre-treatment data. Forecast unbiasedness is a key requirement that is satisfied by our approach under a broad class of data-generating processes that express the individuals counterfactuals as the sum of up to three unobserved components: a stationary process, a stochastic trend and a deterministic trend. Forecasting counterfactuals using a model - even a correctly specified one - does not necessarily result in improved properties of the ATT estimator and is sensitive to misspecification bias in short time series.

A Appendix. Proofs and additional results

A.1 Theorems and proofs for Section 4.1

Consider

$$\widehat{\mathrm{FAT}}_h^{\mathrm{MB}} = \frac{1}{n}\sum_{i=1}^{n}\left[y_{i\tau+h} - \widehat{y}_h(\widehat{\beta}, y_i, x_i)\right],$$

where instead of $\widehat{y}_{i\tau+h}^{(q_i,R_i)}(\widehat{\beta})$ we write $\widehat{y}_h(\widehat{\beta}, y_i, x_i)$, making the dependence of the forecast on y_i and x_i explicit.

Theorem 3. *Assume that*

(i) The forecast is unbiased when evaluated at the true parameter value β_0, i.e.,

$$\mathbb{E}\left[\widehat{y}_h(\beta_0, y_i, x_i) - y_{i\tau+h}\,(0)\right] = 0.$$

(ii) The function $\widehat{y}_h(\beta, y_i, x_i)$ is twice continuously differentiable such that $\frac{\partial \widehat{y}_h(\beta_0, y_i, x_i)}{\partial \beta}$ has finite second moments, and for some $\delta > 0$ we have

$$R_n := \sup_{\{\beta\,:\,\|\beta-\beta_0\|\le\delta\}}\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \widehat{y}_h(\beta, y_i, x_i)}{\partial\beta\partial\beta'}\right\| = o_P(n^{1/2}).$$

(iii) The estimator $\widehat{\beta}$ satisfies

$$\widehat{\beta} - \beta_0 = \frac{1}{n}\sum_{i=1}^{n}\psi(y_i, x_i) + r_n, \tag{32}$$

where $\psi(y_i, x_i)$ has zero mean and finite variance, and $r_n = o_P(n^{-1/2})$. Together with assumption (i) this implies that $\widehat{\beta} - \beta_0 = O_P(n^{-1/2})$.

(iv) The sequence of random variables

$$u^*_{i\tau+h} := y_{i\tau+h} - \widehat{y}_h(\beta_0, y_i, x_i) - \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left[\frac{\partial \widehat{y}_h(\beta_0, y_j, x_j)}{\partial \beta'}\right] \psi(y_i, x_i) \tag{33}$$

satisfies a CLT in the sense that

$$\frac{\frac{1}{\sqrt{n}}\sum_i \left(u^*_{i\tau+h} - \mathbb{E}u^*_{i\tau+h}\right)}{\bar{\sigma}^*_n} \Rightarrow \mathcal{N}(0,1),$$

where $\bar{\sigma}^{*2}_n := Var(\frac{1}{\sqrt{n}}\sum_i u^*_{i\tau+h}) < \infty$.

Then we have that

$$\sqrt{n}\frac{\widehat{\mathrm{FAT}}^{\mathrm{MB}}_h - \mathrm{ATT}_h}{\bar{\sigma}^*_n} \Rightarrow \mathcal{N}(0,1).$$

Proof. We have

$$
\begin{aligned}
&\widehat{\mathrm{FAT}}_h^{\mathrm{MB}} - \mathrm{ATT}_h \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(y_{i\tau+h} - \widehat{y}_h(\widehat{\beta}, y_i, x_i) - \mathbb{E}\left[y_{i\tau+h} - y_{i\tau+h}\left(0\right)\right]\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(y_{i\tau+h} - \widehat{y}_h(\widehat{\beta}, y_i, x_i) - \underbrace{\mathbb{E}\left[y_{i\tau+h} - \widehat{y}_h(\beta_0, y_i, x_i)\right]}_{=\mathbb{E}\left[u^*_{i\tau+h}\right]}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[y_{i\tau+h} - \widehat{y}_h(\beta_0, y_i, x_i) - \frac{\partial\widehat{y}_h(\beta_0, y_i, x_i)}{\partial\beta'}\left(\widehat{\beta} - \beta_0\right)\right] - \mathbb{E}\left[u^*_{i\tau+h}\right] \\
&\quad + O\left(R_n\left\|\widehat{\beta} - \beta_0\right\|^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}u^*_{i\tau+h} - \mathbb{E}\left[u^*_{i\tau+h}\right] + O\left(R_n\left\|\widehat{\beta} - \beta_0\right\|^2\right) \\
&\quad - \left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\widehat{y}_h(\beta_0, y_i, x_i)}{\partial\beta'} - \frac{1}{n}\sum_{j}\mathbb{E}\left[\frac{\partial\widehat{y}_h(\beta_0, y_j, x_j)}{\partial\beta'}\right]\right\}\left(\widehat{\beta} - \beta_0\right) \\
&\quad - \frac{1}{n}\sum_{j}\mathbb{E}\left[\frac{\partial\widehat{y}_h(\beta_0, y_j, x_j)}{\partial\beta'}\right] r_n \\
&= \frac{1}{n}\sum_{i=1}^{n}u^*_{i\tau+h} - \mathbb{E}\left[u^*_{i\tau+h}\right] + o_P(n^{-1/2})
\end{aligned}
$$

Here, in the first step, we plugged in the definitions of $\widehat{\mathrm{FAT}}_h^{\mathrm{MB}}$ and ATT_h. In the second step, we used the unbiasedness of the forecast, definition (33), and assumption (iii) that $\mathbb{E}\left(\psi(y_i, x_i)\right) = 0$. In the third step, given assumption (ii), we employed a Taylor expansion of $\widehat{y}_h(\beta, y_i, x_i)$ in β around β_0. In the fourth step we decomposed $\frac{\partial\widehat{y}_h(\beta_0, y_i, x_i)}{\partial\beta'}$ into its expectation and its deviation from the expectation, and used $\widehat{\beta} - \beta_0 = \frac{1}{n}\sum_{i=1}^{n}\psi(y_i, x_i) + r_n$ and the definition of $u^*_{i\tau+h}$ in (33). In the final step we used our assumptions to conclude that the various remainder terms are all of order $o_P(n^{-1/2})$. By an application of a standard cross-sectional CLT we then obtain the conclusion of the thereom. □

B Replication 1: Overdose mortality with control units

We show here results using the data from Shover et al. (2019) from an analysis using the methods of Callaway and Sant'Anna (2021) and Xu (2017) that use some sort of control group to obtain an estimate of the ATT. Figure 4 uses the method of Callaway and Sant'Anna (2021) with not-yet-treated-states as control units, while Figure 5 uses the method of Xu (2017) to impute counterfactuals for each treated unit using a linear two-way fixed effects regression.

References

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. Journal of Economic Literature 59(2), 391–425. 2

Abouk, R. and S. Adams (2013). Texting bans and fatal accidents on roadways: Do they work? or do drivers just react to announcements of bans? American Economic Journal: Applied Economics 5(2), 179–199. 2

Aguilar, A., E. Gutierrez, and E. Seira (2021). The effectiveness of sin food taxes: Evidence from mexico. Journal of Health Economics 77, 102455. 2

Angrist, J. and J. Pischke (2009). Mostly harmless econometrics. Princeton University Press. 2

Ashenfelter, O. and D. Card (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. The Review of Economics and Statistics 67(4), 648–660. 1

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2021). Matrix completion methods for causal panel data models. Journal of the American Statistical Association 116(536), 1716–1730. 2

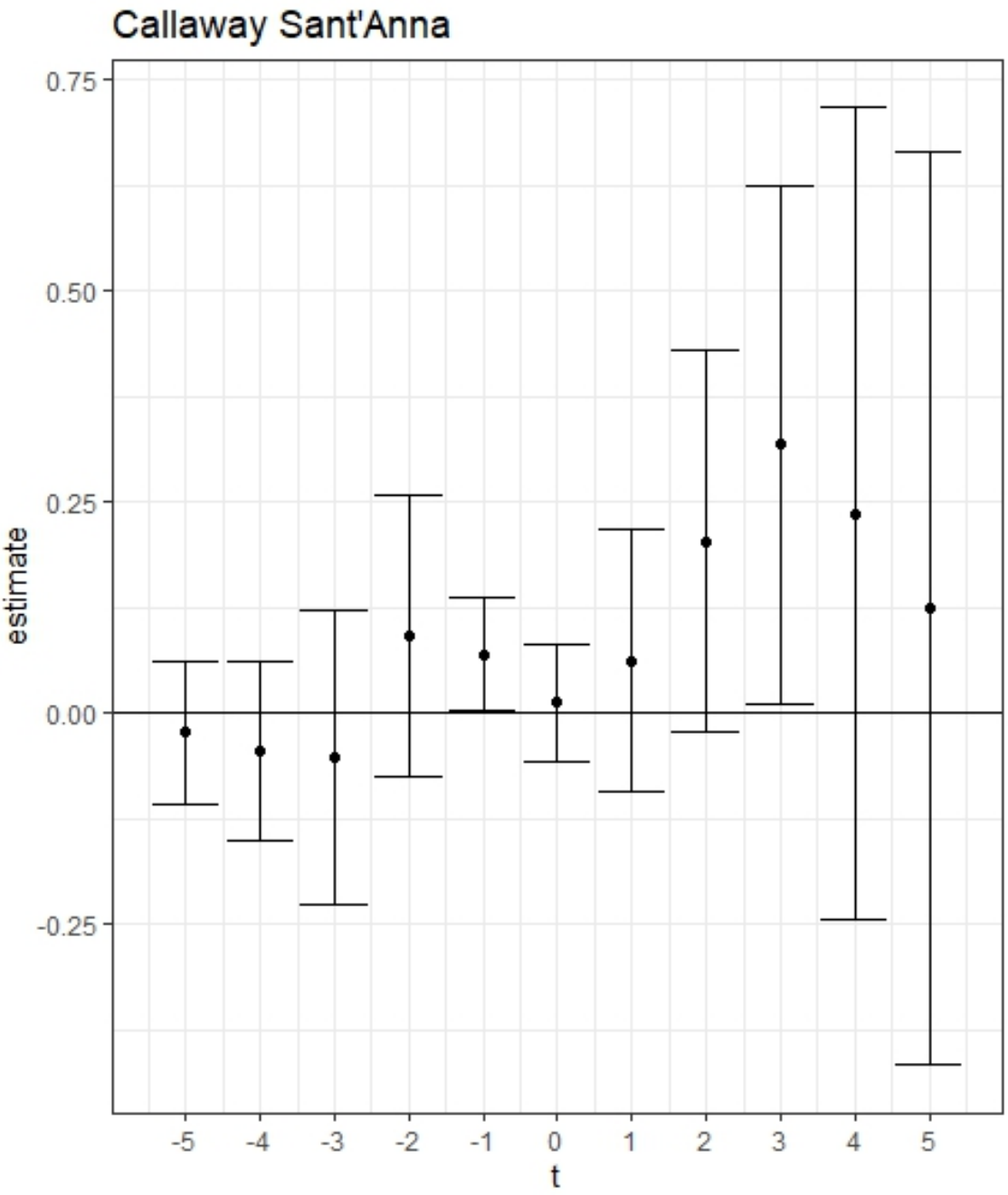

Figure 4: Overdose mortality rate as a function of time to adoption using not-yet-treated-states as control units.

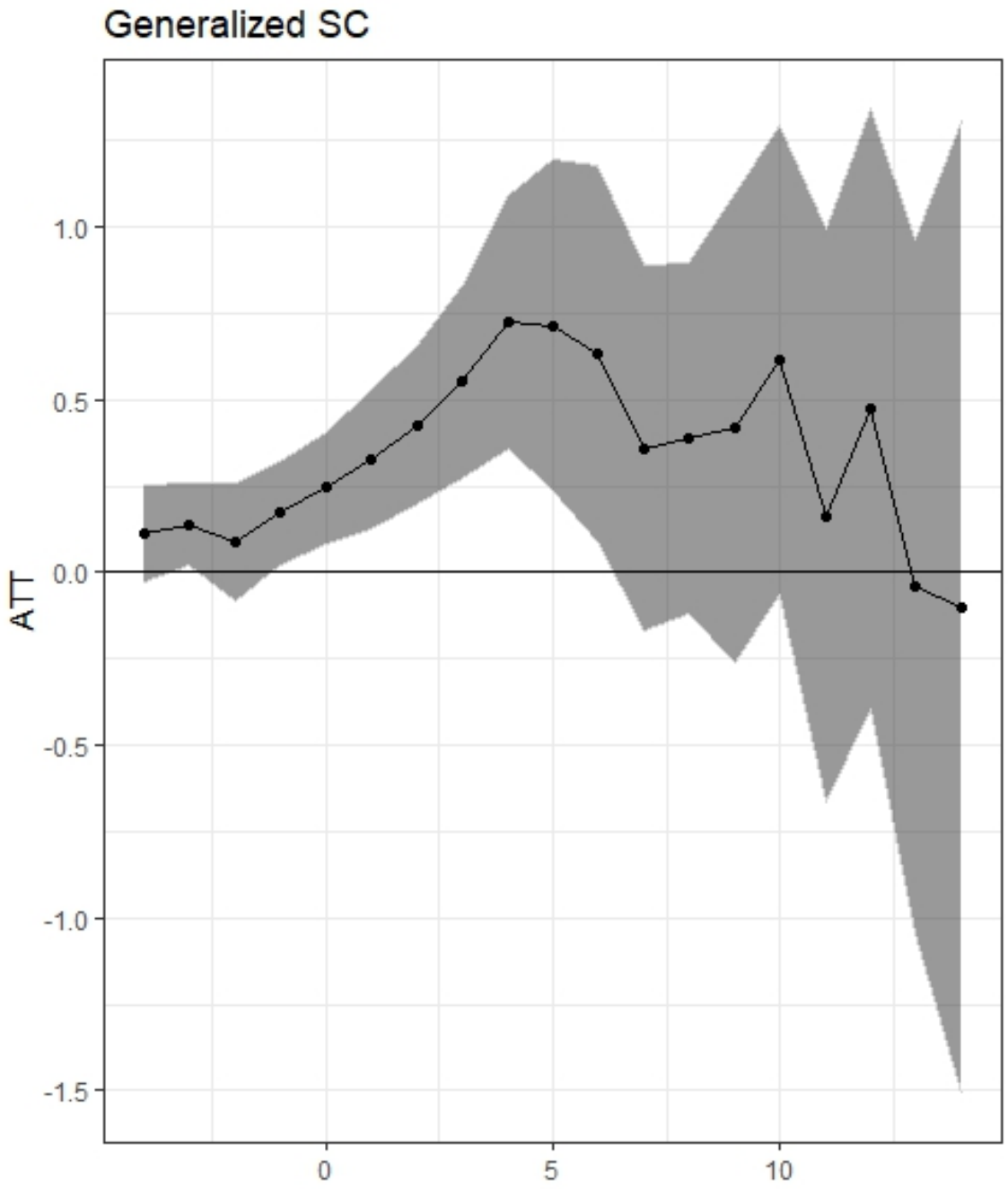

Figure 5: Overdose mortality rate as a function of time to adoption via generalized SC.

Bachhuber, M., B. Saloner, C. Cunningham, and C. Barry (2014). Medical cannabis laws and opioid analgesic overdose mortality in the united states, 1999-2010. JAMA Intern Med. 174(10), 1668–1673. 7

Bai, J. and S. Ng (2021). Matrix completion, counterfactuals, and factor analysis of missing data. Journal of the American Statistical Association 116(536), 1746–1763. 2

Baiker, K. and T. Svoronos (2019). Testing the validity of the single interrupted time series design. 2

Baker, A., D. F. Larcker, and C. C. Y. Wang (2022). How much should we trust staggered difference-in-differences estimates? Journal of Financial Economics 144(2), 370–395. 2

Baker, M., J. Gruber, and K. Milligan (2008). Universal child care, maternal labor supply, and family well-being. Journal of Political Economy 116, 709–745. 2

Baltagi, B. H. (2013). Chapter 18 - panel data forecasting. In G. Elliott and A. Timmermann (Eds.), Handbook of Economic Forecasting, Volume 2 of Handbook of Economic Forecasting, pp. 995–1024. Elsevier. 2

Bernal, J. L., S. Cummins, and A. Gasparrini (2017). Interrupted time series regression for the evaluation of public health interventions: a tutorial. International Journal of Epidemiology 46(1), 348–355. 2

Blundell, R., C. Meghir, M. Costa Dias, and J. Van Reenen (2004). Evaluating the employment impact of a mandatory job search program. Journal of the European Economic Association 2(4), 569–606. 2

Borusyak, K., X. Jaravel, and J. Spiess (2021). Revisiting event study designs: Robust and efficient estimation. 2

Brodersen, K., F. Gallusser, J. Koehler, N. Remy, and S. L. Scott (2015). Inferring causal impact using bayesian structural time-series models. Annals of Applied Statistics 1(9), 247–274. 2

Brown, S. and J. Warner (1985). Using daily stock returns: The case of event studies. Journal of Financial Economics 14(1), 3–31. 2

Callaway, B. and P. H. C. Sant'Anna (2021). Difference-in-differences with multiple time periods. Journal of Econometrics 225(2), 200–230. 2, 7, B

Cattaneo, M. and A. Titiunik (2022). Regression discontinuity designs. Annual Review of Economics 14(1), 821–851. 2

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019). The effect of minimum wages on low-wage jobs. The Quarterly Journal of Economics 134(3), 1405–1454. 2

Chan, M. and S. Kwok (2022). The PCDID approach: Difference-in-differences when trends are potentially unparallel and stochastic. Journal of Business I& Economic Statistics 40(3), 1216–1233. 2

Chen, Y. and A. Whalley (2012). Green infrastructure: The effects of urban rail transit on air quality. American Economic Journal: Economic Policy 4(1), 58–97. 2

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013). Average and quantile effects in nonseparable panel models. Econometrica 81(2), 535–580. 2

Cornelissen, T., C. Dustmann, A. Raute, and A. Schoenberg (2018). Who benefits from universal child care? estimating marginal returns to early child care attendance. Journal of Political Economy 126, 2356–2409. 2

Cryer, J., J. Nankervis, and N. Savin (1990). Forecast error symmetry in arima models. Journal of the American Statistical Association 85(411), 724–728. 2

de Chaisemartin, C. and X. D'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. American Economic Review 110(9), 2964–2996. 2

Dinas, E., K. Matakos, D. Xefteris, and D. Hangartner (2019). Waking up the golden dawn: Does exposure to the refugee crisis increase support for extreme-right parties? Political Analysis 27(2), 244–254. 7

Dufour, J.-M. (1984). Unbiasedness of predictions from estimated autoregressions when the true order is unknown. Econometrica 52(1), 209–216. 2

Fernández-Val, I., H. Freeman, and M. Weidner (2021). Low-rank approximations of nonseparable panel models. The Econometrics Journal 24(2), C40–C77. 2

Fuller, W. and D. Hasza (1980). Predictors for the first-order autoregressive process. Journal of Econometrics 13(2), 139–157. 2

Gallego, F., J.-P. Montero, and C. Salas (2013). The effect of transport policies on car use: Evidence from latin american cities. Journal of Public Economics 107, 47–62. 2

Gelman, A. and G. Imbens (2019). Why high-order polynomials should not be used in regression discontinuity designs. Journal of Business I& Economic Statistics 37(3), 447–456. 2

Ghanem, D., P. Sant'Anna, and K. Wuthrich (2022). Selection and parallel trends. 1

Gillingham, K., C. Knittel, J. Li, M. Ovaere, and M. Reguant (2020). The short-run and long-run effects of covid-19 on energy and the environment. Joule 4(7), 1337–1341. 2

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. Journal of Econometrics 225(2), 254–277. 2, 7

Greenstone, M., G. He, R. Jia, and T. Liu (2022). Can technology solve the principal-agent problem? evidence from China's war on air pollution. American Economic Review: Insights 4(1), 54–70. 2

Hausman, C. and D. Rapson (2018). Regression discontinuity in time: Considerations for empirical applications. Annual Review of Resource Economics 10, 533–552. 2

Heckman, J., H. Ichimura, and P. E. Todd (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. The Review of Economic Studies 64(4), 605–654. 1

Heckman, J., H. Ichimura, and P. E. Todd (1998). Matching as an econometric evaluation estimator. The Review of Economic Studies 65(2), 261–294. 1

Heckman, J. and E. Vytlacil (2005). Econometric evaluation of social programs. In J. Heckman and E. Leamer (Eds.), Handbook of Econometrics, Volume 6. Amsterdam: Elsevier Science. 3

Hirano, K. and G. Imbens (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. Health Services & Outcomes Research Methodology 2, 259–278. 4

Imai, K. and I. S. Kim (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? American Journal of Political Science 63(2), 467–490. 2

Kitagawa, T. and C. Muris (2016). Model averaging in semiparametric estimation of treatment effects. Journal of Econometrics 193(1), 271–289. 4

Kuhn, P. and K. Shen (2023). What happens when employers can no longer discriminate in job ads? American Economic Review 113, 1013–1048. 2

Li, P., Y. Lu, and J. Wang (2020). The effects of fuel standards on air pollution: Evidence from china. Journal of Development Economics 146, 102488. 2

Liu, L., H. Moon, and F. Schorfheide (2020). Forecasting with dynamic panel data models. Econometrica 88(1), 171–201. 2

Liu, L., Y. Wang, and Y. Xu (2023). A practical guide to counterfactual estimators for causal inferences with time-series cross-sectional data. American Journal of Political Science. doi:10.1111/ajps.12723. 2

MacKinlay, A. (1997). Event studies in economics and finance. Journal of Economic Literature 35(1), 13–39. 2

Marx, P., E. Tamer, and X. Tang (2023). Parallel trends and dynamic choices. 1

Mavroeidis, S., Y. Sasaki, and I. Welch (2015). Estimation of heterogeneous autoregressive parameters with short panel data. Journal of Econometrics 188(1), 219–235. 2

Miratrix, L. (2022). Using simulation to analyze interrupted time series designs. Evaluation Review 46(6), 750–778. 2

Nelson, C. and C. Plosser (1982). Trends and random walks in macroeconomic time series: Some evidence and implications. Journal of Monetary Economics 10(2), 139–162. 2

Oscar, J. and A. Taylor (2016). The time for austerity: Estimating the average treatment effect of fiscal policy. The Economic Journal 126, 219–255. 2

Roth, J., P. Sant'Anna, A. Bilinski, and J. Poe (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. Journal of Econometrics 235(2), 2218–2244. 4, 2

Shover, C., C. Davis, S. Gordon, and K. Humphreys (2019). Association between medical cannabis laws and opioid overdose mortality has reversed over time. PNAS 116(26), 12624–12626. 7, B

Sloczyński, T. (2020). Interpreting OLS estimands when treatment effects are heterogeneous: Smaller groups get larger weights. Review of Economics and Statistics 104(3), 501–509. 2

Sun, L. and S. Abraham (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of Econometrics 225(2), 175–199. 2

Trefler, D. (2004). The long and short of the Canada - U.S. free trade agreement. American Economic Review 94(4), 870–895. 2

Tu, M., B. Zhang, J. Xu, and F. Lu (2020). Mass media, information and demand for environmental quality: Evidence from the "Under the Dome". Journal of Development Economics 143, 102402. 2

Varian, H. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives 2(28), 3–28. 2

Watson, M. (1986). Univariate detrending methods with stochastic trends. Journal of Monetary Economics 18, 49–75. 2

White, H. (2001). Asymptotic theory for econometricians. Academic Press. 3.1

Wooldridge, J. M. (2005). Fixed-effects and related estimators for correlated random coefficient and treatment-effect panel data models. Review of Economics and Statistics 87, 385–390. 2

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. Political Analysis 25(1), 1–20. 7, B