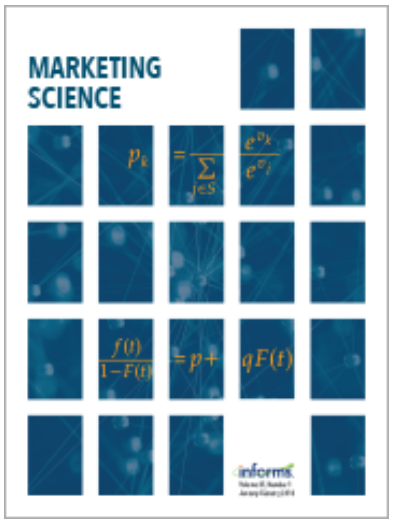


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Augmented Difference-in-Differences

Kathleen T. Li, Christophe Van den Bulte

To cite this article:
Kathleen T. Li, Christophe Van den Bulte (2023) Augmented Difference-in-Differences. Marketing Science 42(4):746-767.
https://doi.org/10.1287/mksc.2022.1406

https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Vol. 42, No. 4, July–August 2023, pp. 746–767
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Augmented Difference-in-Differences

Kathleen T. Li,[a,*] Christophe Van den Bulte[b]

[a] McCombs School of Business, University of Texas at Austin, Austin, Texas 78705; [b] The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104

*Corresponding author

Contact: kathleen.li@mccombs.utexas.edu, https://orcid.org/0000-0002-7813-3626 (KTL); vdbulte@wharton.upenn.edu, https://orcid.org/0000-0001-9708-1596 (CVdB)

Received: December 6, 2018
Revised: March 17, 2020; September 22, 2021; May 5, 2022
Accepted: May 26, 2022
Published Online in Articles in Advance: October 13, 2022

https://doi.org/10.1287/mksc.2022.1406

Abstract. Marketing scientists often estimate causal effects using data from pre/post test/control quasi-experimental settings. We propose a new, easy-to-implement augmented difference-in-differences (ADID) method that complements existing approaches to estimate the average treatment effect on the treated (ATT) from such data. Its advantage over the difference-in-differences method is that it can better handle heterogeneity between treatment and control units and, hence, requires a less stringent causal identification assumption. Its advantages over more flexible approaches like the synthetic control method are that it is easy to implement, provides easy-to-compute confidence intervals, and can be applied to data where the synthetic control and related methods cannot be applied or may not be well suited. Examples are data with short pre- and posttreatment periods or with a large number of treatment and control units. Using analytical proofs, simulations, and nine empirical applications, we document the attractive properties of ADID and provide guidance on what method(s) to use when. With the addition of ADID in their toolkit, marketers are better equipped to address important causal research questions in a wider range of data structures.

History: Avi Goldfarb served as the senior editor and Sridhar Narayanan served as associate editor for this article.
Supplemental Material: The e-companion is available at https://doi.org/10.1287/mksc.2022.1406.

Keywords: causal effects • quasi-experimental methods • inference theory • Augmented DID

1. Introduction

Addressing important managerial and public policy questions often involves identifying the average treatment effect on the treated (ATT) of programs and interventions using quasi-experimental data. Recent marketing examples include quantifying the effects of opening brick-and-mortar showrooms on online demand (Bell et al. 2018), of plain packaging on cigarettes sales (Bonfrer et al. 2020), of soda taxes on soda prices and sales (Kim et al. 2020), of legalizing medical marijuana on opioid prescriptions (Cheon et al. 2021), of using a social media platform on political donations (Petrova et al. 2021), of adopting subscription programs on customer purchases (Iyengar et al. 2022), and of adopting marketing analytics on revenue (Berman and Israeli 2022).

The fundamental problem of causal inference in quasi-experimental settings is that one wants to compare two outcomes for the same observational unit when that unit is exposed or not exposed to an intervention, yet can observe only one outcome at any given time (Holland 1986). Many quasi-experimental methods, such as the difference-in-differences (DID), synthetic control (SC), modified synthetic control (MSC), and Hsiao-Ching-Wan (HCW) methods, are widely used to estimate causal effects such as the ATT. The most popular DID method is easy to implement and offers straightforward inference but also has the most restrictive identifying assumption, making it less suited to settings where there is heterogeneity between treatment and control units. Alternative methods such as the SC, MSC, and HCW methods are better able to handle heterogeneity between treatment and control units. However, these more flexible methods have their own limitations. For example, the HCW method cannot be used when the number of control units is larger than the number of pretreatment time periods. In addition, statistical inference is not as straightforward. For example, the inference theory for the SC and MSC methods does not exist when the data are nonstationary and there are more control units than there are pretreatment time periods.

We propose a new method, augmented difference-in-differences (ADID). The basic idea is to control for the differences in both intercept and slope between treatment and control units in the pretreatment period when calculating the counter-factual posttreatment outcomes of the treated units, rather than controlling for only the difference in intercept as DID does. ADID

is among the class of flexible estimators that predict counterfactual outcomes by estimating weights on each control unit. ADID is more flexible than DID and less flexible than SC or MSC. Whereas DID and ADID share the simplicity of equal weights on control units, ADID has additional flexibility to allow weights to sum up to any value. This ability to scale the control units up or down results in better predicting the counterfactual outcomes. Even more flexible synthetic control methods estimate separate weights on each control unit. ADID does not do this, making estimation and inference straightforward.

ADID has three attractive properties. First, ADID is easy to implement because it only requires estimating two parameters (intercept and slope). Second, ADID can be used in a wide variety of data structures. The additional flexibility from including a slope adjustment makes ADID applicable to situations with heterogeneous treatment and control units where the traditional DID may be too restrictive, whereas the simplicity of estimating only two parameters makes ADID applicable to data structures where the SC, MSC, and HCW methods cannot be applied or may not be well suited. The third attractive property of ADID, shared only with DID, is that users can conveniently compute confidence intervals around the ATT point estimates. Consequently, ADID is especially useful in situations where DID is too restrictive but more flexible methods in the synthetic control family are not suitable or lack inference theory.

We establish and illustrate these three attractive properties using analytical proofs, simulations, and nine empirical applications. First, we prove that the ADID method consistently estimates ATT in situations with (i) long pre- and posttreatment time periods regardless of the number of treatment and control units, or (ii) a large number of treatment and control units regardless of the number of pre- and posttreatment time periods. Second, we develop ADID's formal inference theory and prove that its confidence intervals are easy to compute. Third, using simulations with a variety of data structures, we demonstrate that the ADID method works well in finite samples, and we also compare ADID's performance to that of extant approaches. When the pre- and posttreatment periods are very short, SC, MSC and HCW often cannot be used, whereas ADID still works well and provides inference using a t-distribution. Finally, we present three sets of empirical applications. The first five applications investigate the effects of exogenous price changes by a major retail chain on product sales. The next two applications examine how marijuana legalization affects cigarettes sales. The final two applications investigate how opening an offline showroom affected the sales of an online-first retailer.

This paper contributes to the growing literature on quasi-experimental methods. We propose a new estimator that is more flexible than DID but is simpler to implement than SC, MSC, and HCW. We develop ADID's statistical inference theory and compare its performance against that of four extant methods (DID, SC, MSC, and HCW) in simulations and empirical applications. We find that, as expected, ADID tends to outperform the other estimators in bias, precision, or both, in data structures with specific observable characteristics, but tends to be dominated by at least one alternative method in data structures with other specific observable characteristics. This shows that ADID complements rather than replaces extant approaches. Specifically, ADID tends to do especially well in data structures with a large number of treatment and control units, with short pre- and posttreatment periods, and with large amounts of heterogeneity where DID and SC parallel trends are violated. Since our theoretical, simulation, and empirical analyses indicate that no method is superior across all settings, we also provide some guidance on when to use which method. With the addition of ADID in their toolkit, marketers are better equipped to address important causal research questions in a wider range of data structures.

The remainder of the paper is organized as follows. In Section 2, we introduce the augmented DID method and the four extant methods we compare it to, focusing on the single treatment unit case. Then, we discuss causal identification assumptions and the relative merits and limitations of different methods. In Section 3, we develop the inference theory for the ADID estimator under various data structures, and extend the results to cases with multiple treatment units. In Section 4, we report simulation results comparing mean squared errors (MSEs) and coverage probabilities under different data-generating processes. Section 5 presents three sets of empirical applications. Section 6 concludes the paper. Several web appendices provide the relevant assumptions, derivations of the main results, additional simulations, empirical results, and some further theoretical analyses.

2. Estimating ATTs

We discuss how to estimate the average treatment effect on the treated (ATT) in quasi-experimental data settings with pre/post intervention time periods and treatment/control units. We consider five methods: difference-in-differences (DID), augmented difference-in-differences (ADID), synthetic control (SC), modified synthetic control (MSC), and the Hsiao-Ching-Wan (HCW) method. Section 2.1 shows how to estimate the ATT using these five methods and highlights the connections among them. Section 2.2 discusses the assumptions needed for causal identification using each of the five methods. Section 2.3 discusses their advantages and disadvantages and provides guidance on when to use which method.

All these sections focus on the case where there is only one treatment unit, and the extension to multiple treatment units is postponed until Section 3.3.

2.1. Estimation Methods

We first introduce some general notation. Let y_{it}^1 and y_{it}^0 denote unit i's potential outcome in period t with and without treatment, respectively. The treatment effect to the ith unit at time t is defined as $\Delta_{it} = y_{it}^1 - y_{it}^0$. Since one observes either y_{it}^0 or y_{it}^1, but not both, the observed data are of the following form:

$$y_{it} = d_{it}y_{it}^1 + (1 - d_{it})y_{it}^0, \qquad i = 1, \ldots, N;\ t = 1, \ldots, T, \tag{2.1}$$

where $d_{it} = 1$ if the ith unit receives a treatment at time t and $d_{it} = 0$ otherwise.

Since this section focuses on cases where only one unit receives the treatment, we assume without loss of generality that the first unit receives a treatment at time $T_1 + 1$ with $1 < T_1 < T$, while the remaining $(N - 1)$ units do not receive treatment throughout the sample period. Therefore, $y_{1t} = y_{1t}^0$ for $t = 1, \ldots, T_1$, and $y_{1t} = y_{1t}^1$ for $t \geq T_1 + 1$, whereas $y_{jt} = y_{jt}^0$ for $j = 2, \ldots, N$ and $t = 1, \ldots, T$.

In order to estimate the ATT, we need to estimate the treatment counterfactual, y_{1t}^0, in the posttreatment period, $t \geq T_1 + 1$. Let $\hat{y}_{1t}^0$ be a generic estimator of y_{1t}^0. The treatment effect at time t can then be estimated by $\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^0$ (for $t \geq T_1 + 1$) and the ATT estimator that averages $\hat{\Delta}_{1t}$ over the posttreatment period is

$$\hat{\Delta}_1 = \frac{1}{T_2} \sum_{t=T_1+1}^{T} \hat{\Delta}_{1t}, \tag{2.2}$$

where $T_2 = T - T_1$ is the number of posttreatment time periods.

2.1.1. The DID Method. DID is the most popular method to estimate the ATT in pre/post and test/control observational designs. It is very easy to implement, provides straightforward inference (standard errors and confidence intervals), and can be applied regardless of the number of treatment and control units or the number of pre- and posttreatment time periods.

For causal identification, the DID method relies on the assumption that the sample paths of the treatment unit (y_{1t}) and of the average of the control units, $\bar{y}_{co,t} = \frac{1}{N-1}\sum_{j=2}^{N} y_{jt}$, are parallel in the absence of treatment. The DID method's counterfactual outcome can be obtained via the following regression model (Web Appendix B):

$$y_{1t} - \bar{y}_{co,t} = \delta_1 + e_{1t}, \qquad t = 1, \ldots, T_1. \tag{2.3}$$

Estimating δ_1 by the least-squares method yields $\hat{\delta}_1 = T_1^{-1}\sum_{t=1}^{T_1}(y_{1t} - \bar{y}_{co,t})$. The counterfactual outcome is then estimated as $\hat{y}_{1t,DID}^0 = \hat{\delta}_1 + \bar{y}_{co,t}$. The resulting DID ATT estimator is $\hat{\Delta}_{1,DID} = T_2^{-1}\sum_{t=T_1+1}^{T}(y_{1t} - \hat{y}_{1t,DID}^0)$.

2.1.2. The ADID Method. Like DID, ADID is easy to implement, provides straightforward inference, and applies to settings regardless of the number of treatment and control units or the number of pre- and posttreatment time periods. The ADID method has a bit more flexibility than DID to satisfy the parallel trends assumption.

The ADID method relies on a simple modification to the DID method. We multiply $\bar{y}_{co,t}$ by a slope adjustment (a constant, δ_2), which leads to the following regression model:

$$y_{1t} = \delta_1 + \delta_2\bar{y}_{co,t} + e_{1t}, \qquad t = 1, \ldots, T_1. \tag{2.4}$$

Let $\hat{\delta} = (\hat{\delta}_1, \hat{\delta}_2)'$ denote the least-squares estimator of $\delta = (\delta_1, \delta_2)'$ based on (2.4) using the pretreatment data. We estimate y_{1t}^0 by $\hat{y}_{1t,ADID}^0 = \hat{\delta}_1 + \hat{\delta}_2\bar{y}_{co,t}$, and the ADID ATT by

$$\hat{\Delta}_{1,ADID} = \frac{1}{T_2}\sum_{t=T_1+1}^{T}(y_{1t} - \hat{y}_{1t,ADID}^0). \tag{2.5}$$

Forcing $\hat{\delta}_2 = 1$ reduces ADID to DID.

2.1.3. The SC Method. Proposed by Abadie and Gardeazabal (2003) and Abadie et al. (2010), the synthetic control (SC) method is another way to estimate ATTs. The inference theory was developed recently by Li (2020) and is much less straightforward than for DID and ADID. Also, there is no inference theory for data structures where there are more control units than pretreatment time periods and the data are nonstationary, which makes SC unattractive to researchers working with such data and interested in quantifying the uncertainty around their point estimates.

The SC method computes counterfactuals by first minimizing a constrained least-squares objective function. Define $z_t = (1, y_{2t}, \ldots, y_{Nt})'$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_N)'$. Let $\hat{\beta}_{SC}$ denote the SC method estimated β. Then $\hat{\beta}_{SC}$ minimizes

$$\sum_{t=1}^{T_1}(y_{1t} - z_t'\beta)^2, \tag{2.6}$$

subject to three constraints: (i) no intercept ($\beta_1 = 0$), (ii) slope coefficients sum to one, and (iii) all slope coefficients are nonnegative. There are at least three motivations for imposing these constraints. First, when the number of control units (N_{co}) is larger than the number of pretreatment time periods (T_1), the least-squares estimator of β is not defined. Imposing regularization conditions identifies the estimator in such cases. Second, many outcome variables tend to move together

across units (i.e., they are positively correlated), such that imposing nonnegativity on the weights (β_js) is expected to result in narrower confidence intervals and more accurate out-of-sample predictions. Third, when the treatment and controls are random draws from a common distribution, the constraint that the slope coefficients sum to one is correct and imposing it can help narrow the confidence intervals and boost the out-of-sample prediction accuracy. With $\hat{\beta}_{SC}$ defined above, we obtain $\hat{y}^0_{1t,SC} = z_t'\hat{\beta}_{SC}$. Therefore, the resulting SC estimator of the ATT is $\hat{\Delta}_{1,SC} = T_2^{-1}\sum_{t=T_1+1}^{T}(y_{1t} - \hat{y}^0_{1t,SC})$.

2.1.4. The MSC Method. Proposed by Doudchenko and Imbens (2016), the modified synthetic control (MSC) method is more flexible than the SC method, as it dispenses with the constraints of $\beta_1 = 0$ and $\sum_{j=2}^{N}\beta_j = 1$; that is, MSC only imposes the nonnegativity constraints on the weights β_j for $j \geq 2$ and there is no restriction on the intercept β_1. The inference theory was developed by Li (2020) but is not available for data structures where the number of control units is larger than the number of pretreatment time periods and the data are nonstationary.

The MSC method can be described as follows: Let $\hat{\beta}_{MSC}$ be the MSC estimator of β that minimizes $\sum_{t=1}^{T_1}(y_{1t} - z_t'\beta)^2$ subject to all the slope coefficients being nonnegative ($\beta_j \geq 0$ for $j = 2,\ldots,N$). The resulting MSC estimator of the ATT is $\hat{\Delta}_{1,MSC} = T_2^{-1}\sum_{t=T_1+1}^{T}(y_{1t} - \hat{y}^0_{1t,MSC})$ with $\hat{y}^0_{1t,MSC} = z_t'\hat{\beta}_{MSC}$.

2.1.5. The HCW Method. Proposed by Hsiao et al. (2012), the HCW method, which is also referred to as the OLS method, is the most flexible of the five methods. The inference theory was discussed by Hsiao et al. (2012) and expanded by Li and Bell (2017). The HCW method is not feasible when the number of control units is larger than the number of pretreatment time periods.

The HCW method uses each control unit as a separate explanatory variable and estimates the following regression model (by the least-squares method):

$$y_{1t} = z_t'\beta + \epsilon_{1t} \qquad t = 1,\ldots,T_1. \tag{2.7}$$

Let $\hat{\beta}_{OLS}$ denote the least-squares estimator of β based on (2.7) using the pretreatment data. One estimates the counterfactual outcome by $\hat{y}^0_{1t,HCW} = z_t'\hat{\beta}_{OLS}$ for $t = T_1+1,\ldots,T$. The resulting estimator of the ATT is $\hat{\Delta}_{1,HCW} = T_2^{-1}\sum_{t=T_1+1}^{T}(y_{1t} - \hat{y}^0_{1t,HCW})$.

2.1.6. Mechanical Relationships Among Methods. Figure 1 shows the connections among the five methods. The β_j in Figure 1 refer to the weights or slope coefficients where $j = 2,\ldots,N$ denotes a control unit. At the top is the HCW method, which is the least restrictive. Imposing the restriction that the weights (slope coefficients) are nonnegative results in the MSC method. Additionally imposing that the weights sum to one and the intercept is zero results in the SC method. From the MSC method, we arrive at the DID method by imposing that the weights are all equal and sum to one. Starting back on top with HCW and restricting all weights to be equal results in the ADID. Further imposing that the weights are positive and sum to one results in DID. The equal weight restriction in the DID and ADID is what makes these two methods parsimonious, which in turn makes them easy to use and results in straightforward inference.

There are other, more sophisticated estimation methods to predict the counterfactual outcome $\hat{y}^0_{1t}$ for $t \geq T_1 + 1$, including some machine learning tools (e.g., the LASSO method; see Mullainathan and Spiess 2017). Examining the inferential theory of ATT estimators based on machine learning techniques is beyond the scope of this paper.

2.2. Causal Identification and Parallel Trends

Without random assignment of units to treatment, one needs to rely on some assumptions to make causal claims. In general terms, one needs to assume that the control units' outcomes can be used to determine what the treatment units' outcomes would have been in the absence of treatment, which is the counterfactual. Different methods use different estimation techniques in order to calculate this counterfactual of interest. Often, the identifying assumption is expressed in terms of a parallel trends.

Table 1 states the identifying parallel trends assumption specific to each method (DID,[1] ADID, SC,[2] MSC, HCW). It is not possible to directly test the parallel trends assumption because we do not observe the counterfactual outcome. It is therefore useful to break

Figure 1. Mechanical Relationships Among Methods

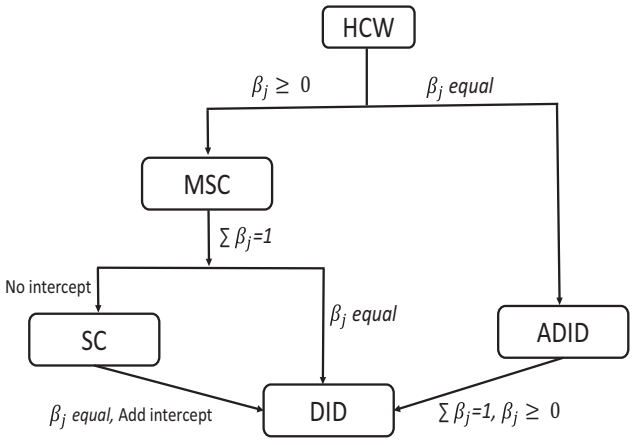


Table 1. Identifying Parallel Trends Assumption by Method

Method	Identifying parallel trends assumption
DID	Treated unit's outcome would have been parallel to the average control units' outcomes in the absence of treatment
ADID	Treated unit's outcome would have been parallel to a slope-adjusted average of the control units' outcomes in the absence of treatment
SC	Treated unit's outcome would have been the weighted average of control units' outcome in the absence of treatment
MSC	Treated unit's outcome would have been parallel to a linear combination with nonnegative weights of the control units' outcome in the absence of treatment
HCW	Treated unit's outcome would have been parallel to a linear combination of control units' outcome in the absence of treatment

down each parallel trends assumption into a testable and a nontestable part. The testable part consists of the general assumption at least holding in the pretreatment period. The nontestable part is that the correlation structure between the treatment and control units in the pretreatment time period continues to hold in the posttreatment time period in the absence of treatment. Whereas one cannot check the latter condition using the observed data, one can at least check whether the former condition is satisfied. Beyond these mechanical checks, Kahn-Lang and Lang (2020) recommend also considering to what extent the treatment and control units may differ in observable or nonobservable dimensions possibly associated with unaccounted differences in baseline outcomes in the absence of treatment, driving a wedge between the true ATT and its statistical estimate. Note that, unlike differences in baseline outcomes, differences in effect size between treatment and control units do not affect the true ATT.

2.3. Comparing the Different Methods

This section discusses the advantages and disadvantages of the newly proposed augmented DID method and the four extant ATT estimation methods (DID, SC, MSC, and HCW). It also provides specific, if somewhat mechanical, guidance on when to use which method.

The key advantages of DID are that it is very easy to implement and is very efficient when valid. The method is especially effective when there are large numbers of treatment and control units over short time periods. However, the DID method is also the most restrictive because its parallel trends assumption is more likely to be violated than those of other methods, especially when the treatment and control units are heterogeneous, that is, have different pretreatment patterns.

The SC method is more general than the DID method as it allows for different controls to have different weights. The SC method imposes a zero intercept and weights summing to one. When these restrictions hold, the SC method provides large efficiency gains over more flexible methods like MSC and HCW. When these restrictions do not hold, SC is likely to be biased. The MSC method adds an intercept and removes the weights summing to one restriction. Therefore, when the SC parallel trends assumption is violated, the MSC parallel trends assumption may still hold. When it comes to inference, both SC and MSC have a disadvantage vis-à-vis DID and ADID. Because SC and MSC involve constrained estimation, their point estimates have nonstandard distributions, making inference much more complex than for DID and ADID (Li 2020). In addition, when the number of control units is larger than the number of pretreatment time periods and the data are nonstationary, there is no formal inference theory available at all for SC and MSC. Furthermore, simulation analyses reported in Section 4.4 show that the MSC method performs poorly when the number of control units is much larger than the number of pretreatment time periods, due to a large number of parameters being estimated.

The HCW method has the most flexible weights. It can be the preferred choice when the treatment and control units are negatively correlated and the number of control units is small. However, when the number of control units is larger than the number of pretreatment time periods, HCW is not feasible. When the number of control units is smaller than the number of pretreatment time periods, but only slightly so, the HCW method has the disadvantages of likely overfitting the in-sample data and of being inefficient because the large number of parameters (weights) results in a large estimation variance.

The ADID method shares the parsimony, ease of use, efficiency, and formal inference benefits of the DID method. These benefits stem from the ADID's equal weight restriction but come at the risk of a bias if the restrictions are invalid. Therefore, the more flexible MSC and the HCW method may perform better in-sample, but not necessarily out-of-sample when compared with ADID. Simulations reported in Section 4.2 show that the ADID estimates often have a smaller mean squares error (MSE) than those obtained from more flexible methods.

Balancing ease of use, bias, precision, and the ability to perform statistical inference, and taking into

account a variety of data structures, we propose the following guidance on when to use what method(s):

- First, if the testable part of the DID parallel trends assumption holds, then use the DID method. Otherwise, or if a robustness check is desired, continue to the next steps.
- Second, assess and possibly exclude methods based on data structure.
  - If the number of control units is larger than the number of pretreatment time periods, then exclude HCW because it is not feasible.
  - If the number of control units is much larger than the number of pretreatment time periods, then exclude MSC because it cannot estimate the ATT accurately.
  - If the number of control units is larger than the number of pretreatment time periods and the data are nonstationary, then consider excluding SC and MSC because inference theory does not exist.
- Third, assess and exclude remaining candidate methods based on the validity of their respective parallel trends assumption.
  - Check the testable part of the parallel trends assumption of each remaining candidate method.
  - Exclude any method for which the testable part of the parallel trends assumption is not satisfied.
- Fourth, compare the ATTs of the remaining methods and consider the precision of the estimates.
  - If the differences among the ATTs are small, then choose the method with the most restrictive assumptions (DID $\succ$ SC $\succeq$ ADID $\succ$ MSC $\succ$ HCW) because it tends to produce the most precise estimates (narrowest confidence intervals).
  - If the differences among the ATTs are large, then calculate the prediction MSE (PMSE) for each method and choose the method with the smallest PMSE. For the definition and calculation of PMSE, see the appendix.

Note, even if the first step suggests using DID as the main method, it will often be sound practice to use one or more additional methods as a robustness check. As already noted above, sound practice also goes beyond such mechanical rules and takes into consideration institutional details of the research setting that may drive a wedge between the true ATT and its estimate (Kahn-Lang and Lang 2020).

3. Inference Theory

This section develops the inference theory for the ADID method for both stationary and nonstationary data. Sections 3.1 and 3.2 focus on the single-treatment unit case, and Section 3.3 extends the results allowing for multiple treatment units with either common or staggered treatment timing.

3.1. Stationary Data

The proposed augmented DID estimator is consistent for stationary data. That is, under mild conditions stated in Web Appendix C, the ATT estimate converges to the true ATT when the numbers of both pretreatment (T_1) and posttreatment (T_2) observations are large. Though the consistency results rely on large numbers of observations, simulations reported in Section 4 suggest that even a moderate sample size of $(T_1, T_2) = (40, 10)$ is large enough for the ADID method to work well. Additional simulations reported in Web Appendix E.4 suggest that the ADID method is still useful when T_1 and T_2 are small (even $T_1 = T_2 = 5$) provided that one uses critical values from a t_{T_1-2} distribution when computing confidence intervals.

Before the intervention, the outcome for the treatment unit is

$$y_{1t}^0 = x_t'\delta + e_{1t}, \qquad t = 1, \ldots, T_1, \tag{3.1}$$

where $x_t = (1, \bar{y}_{co,t})'$ and $\delta = (\delta_1, \delta_2)'$. We interpret (3.1) as a linear projection model, where e_{1t} is the projection error. Therefore, x_t and e_{1t} are uncorrelated by the definition of a linear projection. After treatment occurs to the first unit at time $t = T_1 + 1$, the outcome is

$$y_{1t}^1 = x_t'\delta + \Delta_{1t} + e_{1t}, \qquad t = T_1 + 1, \ldots, T. \tag{3.2}$$

As discussed above, ADID exploits the correlation between the treatment unit's outcome (y_{1t}) and the average of the control units' outcomes ($\bar{y}_{co,t}$) to estimate the counterfactual outcome for the treatment unit. The identifying assumption is that the linear projection coefficient δ remains the same for the posttreatment period in the absence of treatment; that is, the correlation between y_{1t}^0 and $\bar{y}_{co,t}$ (hence, x_t) remains the same during the posttreatment period in the absence of treatment. Therefore, one can consistently estimate δ using pretreatment data. Let $\hat{\delta}$ denote the least-squares estimator of δ based on (3.1) using the pretreatment data, and let $\hat{\Delta}_1 \equiv \hat{\Delta}_{1,ADID}$ be the ADID ATT estimate defined in (2.5). The following proposition presents the large sample distribution of $\hat{\Delta}_1$.

Proposition 3.1. *Under Assumption C1 given in Web Appendix C, we have*

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)/\sqrt{\hat{\Sigma}} \xrightarrow{d} N(0, 1), \tag{3.3}$$

where $\Delta_1 = T_2^{-1}\sum_{t=T_1+1}^{T} \Delta_{1t}$ *and* $\hat{\Sigma}$ *is defined in Appendix A.*

Proposition 3.1 implies that for $\alpha \in (0, 1)$, the $(1 - \alpha)$ confidence interval of Δ_1 is given by

$$\left[\hat{\Delta}_1 - z_{1-\alpha/2}\sqrt{\hat{\Sigma}}/\sqrt{T_2},\ \hat{\Delta}_1 - z_{\alpha/2}\sqrt{\hat{\Sigma}}/\sqrt{T_2}\right], \tag{3.4}$$

where z_α is the αth quantile of a standard normal random variable; that is, $P(N(0,1) \le z_\alpha) = \alpha$. For example, for $\alpha = 0.10$, the 90% confidence interval for Δ_1 is given by

$$\left[\hat{\Delta}_1 - 1.645\sqrt{\hat{\Sigma}}\Big/\sqrt{T_2},\ \hat{\Delta}_1 + 1.645\sqrt{\hat{\Sigma}}\Big/\sqrt{T_2}\right].$$

Our inference theory involving the standard normal distribution requires that both T_1 and T_2 are large. When T_1 and T_2 are small, we suggest using the t_{T_1-2} distribution when conducting inference. The degrees of freedom, $T_1 - 2$, equal to the pretreatment sample size (T_1) minus the number of parameters (δ_1, δ_2). Simulations reported in Web Appendix E.4 show that when T_1 and T_2 are small, the estimated confidence intervals are more accurate based on the t_{T_1-2} distribution than the $N(0, 1)$ distribution.

3.2. Nonstationary Data

The augmented DID estimator is consistent for nonstationary data, too. We first consider nonlinear trend data. Suppose that in the absence of treatment, the data are generated by the following process:

$$y_{jt}^0 = a_j + b_j f_t + u_{jt}, \qquad j = 1, \ldots, N;\ t = 1, \ldots, T, \tag{3.5}$$

where a_j and b_j are finite constants, f_t is a trend process, and u_{jt} is a zero mean weakly dependent stationary process. We allow f_t to be an arbitrary nonlinear trend process. A simple example is a quadratic trend process, $f_t = c_0 + c_1 t + c_2 t^2$. Proposition 3.2 below extends Proposition 3.1 to allow for nonlinear trend outcome variables.

Proposition 3.2. *Assuming that y_{jt} is generated by a (linear or nonlinear) trend process defined in (3.5), under Assumption C2 in Web Appendix C, we have*

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)\Big/\sqrt{\hat{\Sigma}} \xrightarrow{d} N(0,1), \tag{3.6}$$

where $\hat{\Sigma}$ is defined in Appendix A.

Proposition 3.2 states that the statistic $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)/\sqrt{\hat{\Sigma}}$ has the same standard normal distribution as in the stationary data case. Note, one does not need to know or estimate the specific nonlinear trend in the outcome variable (y_{jt}). When estimating the counterfactual outcome, one regresses y_{1t} on $(1, \bar{y}_{co,t})$, and the calculations of $\hat{\Delta}_1$ as well as $\hat{\Sigma}$ remain the same whether y_{jt} is a stationary or a nonlinear trend process.

Another type of nonstationary process is a unit-root process, $f_t = f_{t-1} + \xi_t$, where ξ_t is a zero mean, weakly dependent stationary idiosyncratic error term. Assume that in the absence of treatment, outcome variables are generated by

$$y_{jt}^0 = a_j + b_j f_t + u_{jt} \text{ with } \quad f_t = f_{t-1} + \xi_t, \tag{3.7}$$

that is, f_t follows a drift-less unit-root process. Whereas the common factor f_t in (3.5) has a nonstationary (deterministic) trend component, f_t in (3.7) follows a driftless (i.e., no intercept) unit-root process, and it does not have a deterministic trend component.

The following proposition gives the asymptotic distribution theory for the ADID estimator with unit-root nonstationary data.

Proposition 3.3. *Assuming that y_{jt} is generated by a unit-root process as in (3.7), under Assumption C3 in Web Appendix C, we have*

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)\Big/\sqrt{\hat{\Sigma}} \xrightarrow{d} N(0,1), \tag{3.8}$$

where $\hat{\Sigma}$ is defined in Appendix A.

Proposition 3.3 states that the statistic $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)/\sqrt{\hat{\Sigma}}$ has the same standard normal distribution as in the stationary data case. The reason is similar to the result that, when one estimates a cointegration relationship, a standardized t-statistic can have a standard normal distribution under fairly general conditions, even though the least-squares estimated coefficient has an asymptotic nonnormal distribution (a distribution characterized by integration of Brownian motions). See Hayashi (2000, p. 658) and Hamilton (1994, pp. 608–610) for detailed discussions.

Simulations reported in Section 4 show that the inference theories presented in Propositions 3.1–3.3 work well for a variety of data-generating processes, including stationary, nonlinear trend, and unit-root nonstationary processes.

3.3. Multiple Treatment Units

We now extend the analysis to allow for multiple treatment units. We use N_{tr} and N_{co} to denote the number of treatment and control units. We first consider a simple case where all treatment units receive a treatment at the same time and then deal with the more general case where different treatment units receive treatment at different times.

3.3.1. Common Treatment Timing. When all treatment units receive a common treatment at the same time T_1, we can first average outcomes over the treatment and control group separately to obtain $\bar{y}_{t,tr} = N_{tr}^{-1}\sum_{i=1}^{N_{tr}} y_{it,tr}$ and $\bar{y}_{t,co} = N_{co}^{-1}\sum_{i=1}^{N_{co}} y_{it,co}$, so that we have data sets for two time series, one for the treatment and one for the control group. We can then deal with this scenario as consisting of only one treatment unit.

We first consider the DID method. Under the DID parallel trends assumption, the means of the sample paths, $\bar{y}_{t,tr}$ and $\bar{y}_{t,co}$, differ by a constant (say, δ_1) in the absence of treatment. Using the pretreatment data, we estimate δ_1 based on the intercept-only regression model $\bar{y}_{t,tr} = \delta_1 + \bar{y}_{t,co} + e_t$ (where e_t is a zero mean error), which leads to $\hat{\delta}_{1,DID} = T_1^{-1}\sum_{t=1}^{T}(\bar{y}_{t,tr} - \bar{y}_{t,co})$. The counterfactual outcome is estimated by $\hat{y}^0_{t,tr,DID} = \hat{\delta}_{1,DID} + \bar{y}_{t,co}$, and the ATT estimate is given by

$$\hat{\Delta}_{DID} = \frac{1}{T_2}\sum_{t=T_1+1}^{T}(\bar{y}_{t,tr} - \hat{y}^0_{t,tr,DID}).$$

Similarly, for the ADID method, we estimate $\delta = (\delta_1, \delta_2)'$ from the regression model $\bar{y}_{t,tr} = \delta_1 + \delta_2\bar{y}_{t,co} + e_t$ for $t = 1, \ldots, T_1$ by the least-squares method. If $\hat{\delta} = (\hat{\delta}_1, \hat{\delta}_2)'$ denotes the resulting estimate of δ, then the ATT estimate is given by

$$\hat{\Delta}_{ADID} = \frac{1}{T_2}\sum_{t=T_1+1}^{T}(\bar{y}_{t,tr} - \hat{y}^0_{t,tr,ADID}), \tag{3.9}$$

where $\hat{y}^0_{t,tr,ADID} = \hat{\delta}_1 + \bar{y}_{t,co}\hat{\delta}_2$. The inference theory for $\hat{\Delta}_{ADID}$ is covered by Proposition 3.1 (for stationary data) and Propositions 3.2 and 3.3 (for nonstationary data), since we effectively deal with the same problem as one treatment unit scenario (treating $\bar{y}_{t,tr}$ as y_{1t} in Propositions 3.1–3.3).

The aforementioned ATT estimator allows N_{tr}/N_{co} to be large or small. When N_{tr} and N_{co} are both large, Web Appendix D.1 shows that $\hat{\Delta}_{ADID}$ converges to $\Delta = \frac{1}{N_{tr}T_2}\sum_{i=1}^{N_{tr}}\sum_{t=T_1+1}^{T}\Delta_{it}$ at the rate of $1/(\sqrt{T_2(N_{tr}+N_{co})})$. Therefore, even when T_2 is small, $\hat{\Delta}_{ADID}$ consistently estimates Δ, provided that N_{tr} and N_{co} are large. Simulations in Section 4.4 confirm this convergence rate. Web Appendix E.5 shows that the inference theory works well also when N_{tr} and N_{co} are large but T_2 is small. When T_1 is small as well, we suggest replacing standard normal distribution quantiles by t-distribution (with $T_1 - 2$ degree of freedom) quantiles. See simulations reported in Web Appendix E.4 for supporting evidence.

3.3.2. Staggered Treatment Timing. We now consider the case where different treatment units receive the treatment at different times. Suppose that the ith treatment unit received treatment at time $T_{1i}+1$ for $1 < T_{1i} < T$; thus, posttreatment length is $T_{2i} = T - T_{1i}$ for $i = 1, \ldots, N_{tr}$. We use $y_{it} = y_{it,tr}$ and $y_{it,co}$ to denote treatment and control outcome variables, respectively. Recall that $x_t = (1, \bar{y}_{co,t})'$, where $\bar{y}_{co,t} = N_{co}^{-1}\sum_{i=1}^{N_{co}} y_{it,co}$. For the ith treatment unit during the pretreatment period, we have

$$y_{it} = x_t'\beta_i + e_{it}, \qquad i = 1, \ldots, N_{tr};\ t = 1, \ldots, T_{1i}, \tag{3.10}$$

where $\beta_i = (\beta_{1i}, \beta_{2i})'$. We estimate β_i by $\hat{\beta}_i = (X_i'X_i)^{-1}X_i'Y_i$, where X_i and Y_i are $T_{1i} \times 2$ and $T_{1i} \times 1$ matrices with typical rows given by $x_t' = (1, \bar{y}_{co,t})$ and y_{it}, respectively.

As before, we estimate $\Delta_{it} \overset{def}{=} y^1_{it} - y^0_{it}$ by $\hat{\Delta}_{it} = y_{it} - \hat{y}^0_{it}$ for $t \geq T_{1i} + 1$, where $\hat{y}^0_{it} = x_t'\hat{\beta}_i$. The ATT estimator for the ith treatment unit is given by $\hat{\Delta}_i = T_{2i}^{-1}\sum_{t=T_{1i}+1}^{T}\hat{\Delta}_{it}$. The average treatment effect over all the treated units is obtained by averaging $\hat{\Delta}_i$ over i from 1 to N_{tr}:

$$\hat{\Delta} = \frac{1}{N_{tr}}\sum_{i=1}^{N_{tr}}\frac{1}{T_{2i}}\sum_{t=T_{1i}+1}^{T}\hat{\Delta}_{it}. \tag{3.11}$$

When both N_{tr} and N_{co} are fixed positive integers, the asymptotic distribution of $\hat{\Delta}$ is given in the following proposition.

Proposition 3.4. *Suppose that Assumption C4 in Web Appendix C holds, let $T_2 = N_{tr}^{-1}\sum_{i=1}^{N_{tr}} T_{2i}$ (the average of T_{2i}s) and $\hat{\Delta}$ be defined as in (3.11), then we have*

$$\sqrt{T_2}(\hat{\Delta} - \Delta)\Big/\sqrt{\hat{\Sigma}_{N_{tr}}} \xrightarrow{d} N(0,1),$$

where $\Delta = N_{tr}^{-1}\sum_{i=1}^{N_{tr}} T_{2i}^{-1}\sum_{t=T_{1i}+1}^{T}\Delta_{it}$, and $\hat{\Sigma}_{N_{tr}}$ is defined in the Web Appendix C.5.

The proof of Proposition 3.4 is provided in Web Appendix D.2.

Proposition 3.4 deals with the case where N_{tr} and N_{co} are fixed, and T_1 and T_2 are large. The next proposition covers the case where T_1 and T_2 are small, but N_{tr} and N_{co} are large.

Proposition 3.5. *Suppose that Assumption C5 in Appendix C holds (a short panel with many treatment and control units), and let $\hat{\Delta}$ be defined as in (3.11), then there exists a positive constant $C > 0$ such that*

$$\mathrm{MSE}(\hat{\Delta}) \leq C\left(\frac{1}{N_{tr}} + \frac{1}{N_{co}}\right),$$

where $\mathrm{MSE}(\hat{\Delta}) = E[(\hat{\Delta} - \Delta)^2]$.

The proof of Proposition 3.5 is provided in Web Appendix D.3. Proposition 3.5 implies that when N_{tr} and N_{co} are both doubled, the mean squared error (MSE) will be reduced by half. Simulations in Section 4.4 confirm this MSE convergence rate.

Proposition 3.5 gives the MSE convergence of $\hat{\Delta}$ but not the asymptotic distribution of $\hat{\Delta}$. We conjecture that it may be possible to prove, under some regularity conditions, that $\sqrt{N_{tr}}(\hat{\Delta} - \Delta)$ has an asymptotic normal

distribution with zero mean and a finite variance. We leave verifying this conjecture to future research.

Finally, as with the DID method, one can easily incorporate relevant covariates in the ADID estimation procedure. Web Appendix C.6 discusses the specific estimation steps when there are covariates.

4. Simulation Study

We now assess the performance of ADID by means of simulation. Section 4.1 describes the design. The simulation analysis covers cases where the treatment and control units are either homogeneous or heterogeneous, that is, have the same or different pretreatment patterns. The simulation also considers whether the data are stationary or nonstationary. Section 4.2 compares the performance of five different methods in terms of their mean squared error (MSE), which captures the joint effect of bias and variance. Section 4.3 assesses the confidence intervals produced by the DID and the ADID methods. Section 4.4, finally, examines the case with multiple treatment units.

4.1. Simulation Design

The simulation manipulates to what extent the treatment and control units have the same or different pretreatment patterns. In the first true data-generating process, DGP1, the treatment and control units exhibit the same patterns (homogeneous treatment and control units). In the other two data-generating processes, DGP2 and DGP3, the treatment and control units exhibit different patterns (heterogeneous treatment and control units). In DGP2, the treatment unit's slope is greater than the control units' slopes, such that the treatment unit is outside the range of the control units. In DGP3, the treatment unit's slope is within the range of the control units' slopes. Consider a study in which sales in different cities is the outcome variable. DGP1 reflects a scenario where the pretreatment sales evolution is similar between treatment and control units. For example, if the treatment and control cities have similar demographics, population growth, economic development, and other sources of temporal variation, then pretreatment sales growth will likely be similar in treatment and control cities. For DGP2, in contrast, the sales growth in the treatment unit is higher than in the control units. For instance, larger cities may exhibit higher sales growth (i.e., steeper upward trends) because of higher rates of migration into larger cities or because the demand for the product or service exhibits positive network effects. If the treatment unit is a relatively large city with an above-average sales growth rate, then the pretreatment sales evolution will differ between test and controls, as represented by DGP2. Finally, for DGP3, the treatment unit's sales growth lies within range of growth rates in the control units. An example would be the same scenario as for DGP2, but with the treatment unit being an average-size city and the control units consisting of both larger and smaller cities.

Based on the pattern of each DGP, we have some clear expectations about which methods should be most suited and perform best in each. In DGP1, all five methods (DID, ADID, SC, MSC, and HCW) are applicable because the treatment and control units follow the same patterns and the causal identification assumption is met for each method. In DGP2, only ADID, MSC and HCW are applicable because these are the only methods to allow the treatment unit's pretreatment pattern to be outside the range of the control units. In contrast, DID and SC are not applicable because their "weights summing to one" restriction is valid only when the treatment unit is within the range of the control units. In DGP3, where the pretreatment trend of the treatment unit differs from that of the individual controls yet lies within the range of the collective set, ADID, SC, MSC, and HCW are all suitable because they have enough flexibility to accommodate such a pattern. In contrast, DID should perform poorly: imposing that the weights are equal and sum to one is suitable only when the treatment's trend is equal to the average of the control units' trends.

A convenient way to manipulate the treatment-versus-control patterns in the pretreatment data is to generate the outcome data using a factor model. Generating outcome variables using some common factors results in outcomes of different units that are correlated via the common factors. We generate the $N \times 1$ vector of outcome variables in the absence of treatment, y_t^0, using the following factor model with three factors:

$$y_t^0 = a + Bf_t + u_t, \qquad t = 1, \ldots, T, \tag{4.1}$$

where $y_t^0 = (y_{1t}^0, y_{2t}^0, \ldots, y_{Nt}^0)'$, $a = (a_1, a_2, \ldots, a_N)'$, and $u_t = (u_{1t}, u_{2t}, \ldots, u_{Nt})'$ are all $N \times 1$ vectors. Moreover, $B = (b_1, b_2, \ldots, b_N)'$ is the $N \times 3$ factor-loading matrix, where b_j is a 3×1 factor loading vector for unit j, and f_t is a 3×1 vector of common factors. We choose the intercept to be a constant of ones, $(a_1, a_2, \ldots, a_N) = (1, 1, \ldots, 1)$, and error to be distributed standard normal, u_{jt} is independent and identically distributed (i.i.d.) $N(0, 1)$.

By varying the factor loadings, we can create the pattern for each DGP. To generate DGP1, where the treatment and control units follow the same patterns, we choose the same factor loadings for all units; that is, $b_i = (1, 1, 1)'$ for all $i = 1, \ldots, N$. To generate DGP2 and DGP3, where the treatment and control units follow different patterns, we choose different factor loadings for the treatment and control units; that is, $b_1 \neq b_i$ for $i = 2, \ldots, N$. In order to distinguish between DGP2 and DGP3, where the treatment unit is outside versus

within the range of the control units, we allow the factor loadings on the control units to be different. We split the control units into two groups, where the first group has factor loading b_i (for $i = 2, \ldots, (N-1)/2$) that is different from the factor loading of the second group b_j ($j = (N+1)/2, \ldots, N$). We express the factor loadings for the treatment unit and the two groups of control units[3] as follows:

$$b_1 = c_1 \mathbf{1}_{3\times 1}, b_j = c_2 \mathbf{1}_{3\times 1},\ j = 2, \ldots, \frac{N-1}{2}, b_j = c_3 \mathbf{1}_{3\times 1},\ j = \frac{N+1}{2}, \ldots, N, \tag{4.2}$$

where $\mathbf{1}_{3\times 1} = (1,1,1)'$ is a 3×1 vector of ones.

To create the three DGPs, we simply use a specific combination of c_j, $j = 1, 2, 3$:

$$DGP1: c_1 = 1,\ c_2 = 1,\ c_3 = 1,$$
$$DGP2: c_1 = 1,\ c_2 = -2,\ c_3 = 0.5,$$
$$DGP3: c_1 = 1,\ c_2 = 2,\ c_3 = -0.5.$$

For each DGP, we further examine to what extent the performance of different methods depends on the data being stationary or not (specifically, unit-root or nonlinear trend). We do so by generating outcome variables using a different combination of the following five factors: $f_{1t} = 0.8f_{1,t-1} + \epsilon_{1t}$, $f_{2t} = -0.6f_{2,t-1} + \epsilon_{2t} + 0.8\epsilon_{2,t-1}$, $f_{3t} = \epsilon_{3t} + 0.9\epsilon_{3,t-1} + 0.4\epsilon_{3,t-2}$, $f_{4t} = f_{4,t-1} + \epsilon_{4t}$, and $f_{5t} = 0.2t - 0.8\sqrt{t} + 0.8f_{5,t-1} + \epsilon_{5t}$, where ϵ_{jt} is i.i.d. $N(0, 1)$ for $j = 1, \ldots, 5$. The first three factors are stationary and the last two factors are nonstationary. Specifically, the first three factors are stationary AR(1), ARMA(1,1), and MA(2) processes, respectively. The fourth factor is a unit-root process, and the fifth factor is a nonlinear trend process. We create three types of trends for f_t: (i) stationary: $f_t = (f_{1t}, f_{2t}, f_{3t})'$, (ii) nonstationary unit-root: $f_t = (f_{4t}, f_{2t}, f_{3t})'$, and (iii) nonstationary nonlinear: $f_t = (f_{5t}, f_{2t}, f_{3t})'$.

4.2. Estimation MSE Compared Across Methods

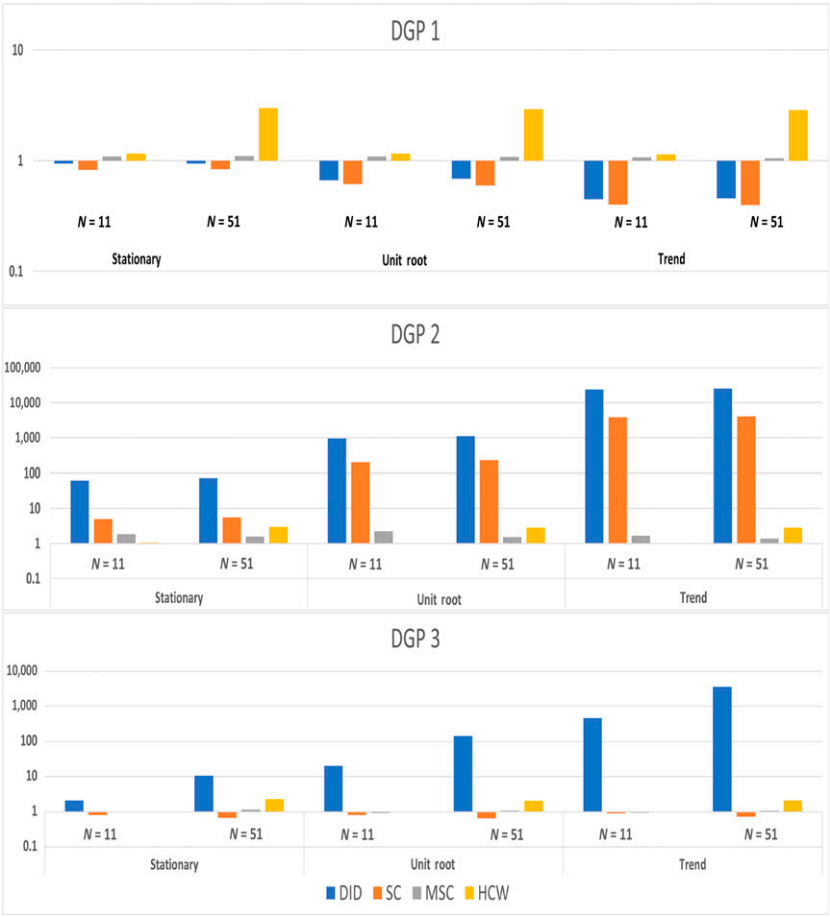
In this section, we compare the ATT estimation mean squared error (MSE) across five methods: DID, ADID, SC, MSC, and HCW. The estimation MSE is computed as $\mathrm{MSE}(\hat{\Delta}_1) = \frac{1}{M}\sum_{j=1}^{M}(\hat{\Delta}_{1,j} - \Delta_{1,j})^2$, where M = 10,000 is the number of simulation replications, $\hat{\Delta}_{1,j}$ is the calculated $\hat{\Delta}_1$ using the jth iteration of simulated data utilizing one of the five estimation methods, and $\Delta_{1,j}$ is Δ_1 computed employing the jth iteration of simulated data. We also compute the variance and squared bias defined by $\mathrm{Var}(\hat{\Delta}_1) = \frac{1}{M}\sum_{j=1}^{M}(\hat{\Delta}_{1,j} - \Delta_{1,j} - (\hat{\Delta}_1 - \Delta_1))^2$, where $\hat{\Delta}_1 = \frac{1}{M}\sum_{l=1}^{M}\hat{\Delta}_{1,j}$, $\Delta_1 = \frac{1}{M}\sum_{j=1}^{M}\Delta_{1,j}$, and $\mathrm{Bias}^2(\hat{\Delta}_1) = (\hat{\Delta}_1 - \Delta_1)^2$. It is easy to check that the standard MSE decomposition $\mathrm{MSE}(\hat{\Delta}_1) = \mathrm{Var}(\hat{\Delta}_1) + \mathrm{Bias}^2(\hat{\Delta}_1)$ holds. Note that the estimation $\mathrm{MSE}(\hat{\Delta}_1)$ is unrelated to the treatment effect, Δ_{1t}. This is because $\hat{\Delta}_1 - \Delta_1$ is unrelated to treatment effect Δ_{1t}, as the treatment effect Δ_{1t} cancels out from $\hat{\Delta}_1$ and Δ_1, leading $\hat{\Delta}_1 - \Delta_1$ to be unrelated to Δ_{1t}. Therefore, we simply choose zero treatment effect, $\Delta_{1t} = 0$, for $t \geq 1$ in all simulations.

To aid a visual comparison of MSE performance across methods, Figure 2 reports MSE ratios, benchmarked against ADID's MSE. It does so for each data pattern (DGP) and for each trend structure (stationary, unit-root, and nonlinear trend). The MSE ratios in Figure 2 are defined as MSE of a given method (DID, SC, MSC, or HCW) divided by the MSE of the ADID method. For example, in the legend for Figure 2, $DID \stackrel{def}{=} MSE_{DID}/MSE_{ADID}$. Therefore, a ratio greater than one implies that the ADID method has a smaller MSE than the given method. The bars in Figure 2 start from one and go above or below one accordingly. The vertical axis in Figure 2 is on a log scale. For Figure 2, we choose $T_1 = 80$ (number of pretreatment time periods) and $T_2 = 20$ (number of posttreatment time periods). We also consider two values of N: $n = 11$ (1 treatment unit and 10 control units) and $n = 51$ (1 treatment unit and 50 control units). The MSE tables are presented in Tables 9–11 in Web Appendix E.1, where we also report results for $N \in \{21, 31\}$ as well as $(T_1, T_2) = (40, 20)$.

We now discuss the results in Figure 2, starting with DGP1, where the treatment and control units have the same patterns. For all cases, the most flexible method (HCW) does the worst, the simplest method (DID) does quite well, but an intermediate method (SC) performs very slightly better in terms of MSE. This pattern holds for stationary, unit-root and nonlinear trend data, but the superiority of SC and DID over other methods is more pronounced for nonstationary data and especially for nonlinear trend data. This implies that, provided that the causal identification assumption holds, superfluous flexibility comes at a cost. Investigating each contributor to the MSE separately shows that, as expected, bias is negligible in DGP1. Specifically, $\mathrm{Bias}^2(\hat{\Delta})/\ \mathrm{Var}(\hat{\Delta}) \leq 0.02$ for all cases.

Next, we discuss the results for DGP2 where the treatment and control units follow different patterns and, specifically, the treatment is outside the range of the control units. DID and SC are not applicable, whereas ADID, MSC, and HCW are. Therefore, we expect ADID, MSC, and HCW methods to perform well and indeed they do. For $n = 11$, Figure 2 shows that these three methods have rather similar MSEs (the ratios are close to one), which are much lower than those of DID and SC, especially in nonstationary data. The DID method has the largest MSE, and the SC has the second largest. For $n = 51$, ADID and MSC remain the best performers, HCW loses some performance,

Figure 2. (Color online) Ratio of Given Method's MSE over ADID's MSE for Long Series ($T_1 = 80$ and $T_2 = 20$)



and DID and SC remain the worst performers by a wide margin. When T_1 and T_2 are doubled, the ADID, MSC, and HCW (when N is small) estimated MSEs are all reduced by about 50%, indicating that these three methods produce consistent ATT estimates (see Tables 9–11 in Web Appendix E.1).

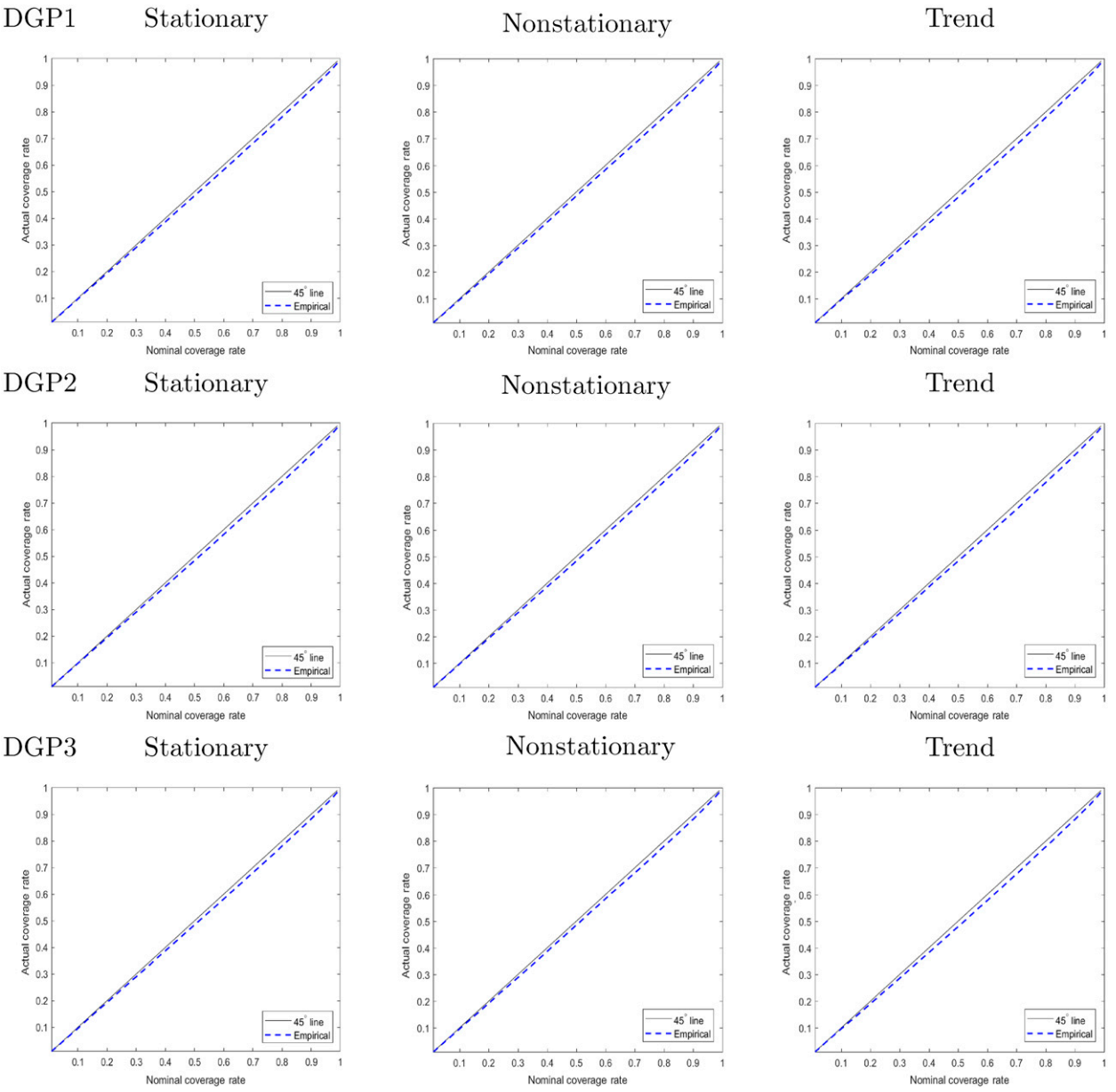
Finally, for DGP3, where the treatment and control units follow different patterns, but the treatment is within the range of the control units, all methods except for the DID are applicable on a priori grounds. As expected, the DID method performs worst in term of MSE due to its large estimation bias. Regarding the other four methods, Figure 2 shows that, for almost all cases, they perform quite similarly, though HCW performs very slightly worse when $n = 51$.

With the exception of the HCW method, the estimation MSEs are not particularly sensitive to different values of N. For HCW, the number of estimated parameters increases with N, which inflates its estimation variances. Comparing the MSEs for $(T_1, T_2) = (40, 10)$ and $(T_1, T_2) = (80, 20)$ shows that when the values for T_1 and T_2 are doubled, the MSEs are halved for most methods (see Tables 9–11 in Web Appendix E.1 for these results). This is consistent with our theoretical results that the estimated MSEs are proportional to $1/(T_1 + T_2)$.

4.3. Coverage Probabilities

Both DID and ADID provide easy-to-use and widely applicable inference theory, unlike SC, MSC, and HCW. Therefore, our analysis of the estimated coverage probabilities considers only DID and ADID. To what extent do the DID and ADID confidence intervals correctly reflect the true variability across replications? The coverage probability plots in Figures 3 and 4 provide the answer, for different data patterns (DGPs and trend structures). Each of the nine plots in these exhibits show the nominal coverage rate of the confidence intervals on the x-axis and the actual coverage rate of the confidence intervals on the y-axis. We choose the nominal coverage rate values, $x_i \in \{0.01, 0.02, \ldots, 0.99\}$, and use 100,000 simulation runs to calculate the actual coverage rate, y_i, which is the proportion of estimated confidence intervals that actually contain the true ATT. We then plot $\{x_i, y_i\}_{i=1}^{99}$ as a dashed curve. If the actual and nominal coverage rates are very close, then the $\{x_i, y_i\}_{i=1}^{99}$ (dashed) curve will be close to the 45-degree solid line. On the

Figure 3. (Color online) ADID Coverage Probability Plots



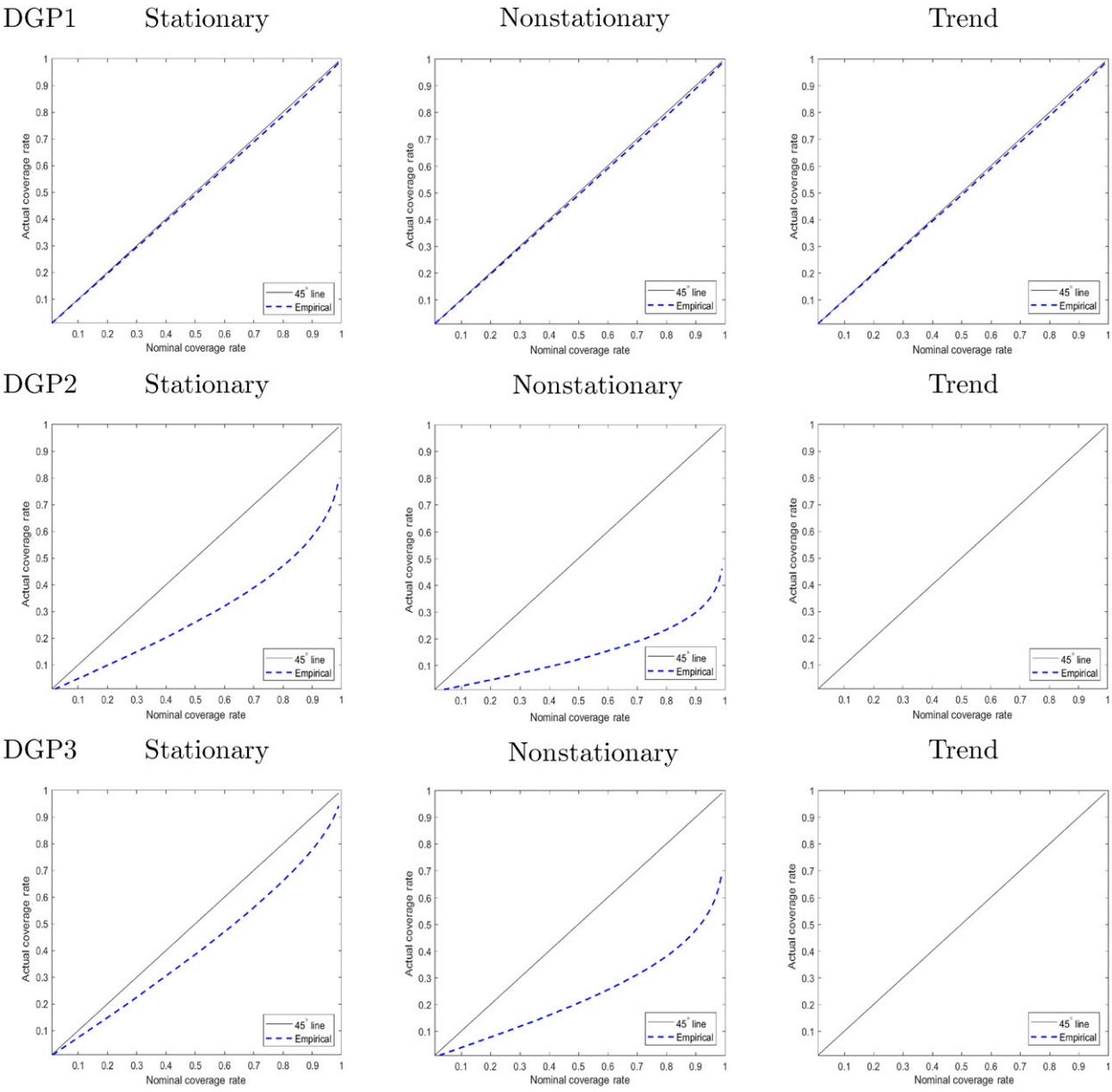
other hand, if the actual and nominal coverage rates diverge, then the dashed curve will deviate from the 45-degree line. Therefore, the plots indicate to what extent the confidence intervals provide proper coverage, over-coverage, or under-coverage.

The dashed lines are very close to the 45-degree line in each of the nine boxes in Figure 3. Hence, ADID has excellent coverage in each of the nine cases spanning different data patterns. Figure 4 shows the coverage probability plots for the DID method. The conclusion is as expected: DID has excellent coverage only for DGP1, where the treatment and control units have the same patterns, but does quite poorly for DGP2 and DGP3, where the treatment and control units have different patterns. Specifically, violation of the DID parallel lines assumption results in pronounced under-coverage. The problem is especially pronounced for nonstationary data (unit-root and nonlinear trends). In fact, for the nonlinear trend data, the under-coverage is so severe that the coverage probability is zero, making the dashed line overlap with the x-axis.

4.4. Multiple Treatment Units

We now turn to comparing DID, ADID, SC, and MSC estimates when the numbers of both treatment units (N_{tr}) and control units (N_{co}) are large. For simplicity, we consider the case where treatment is not staggered: $T_{1i} = T_1$ for all $i = 1, \ldots, N_{tr}$ so that $T_{2i} = T_2 = T - T_1$ for

Figure 4. (Color online) DID Coverage Probability Plots

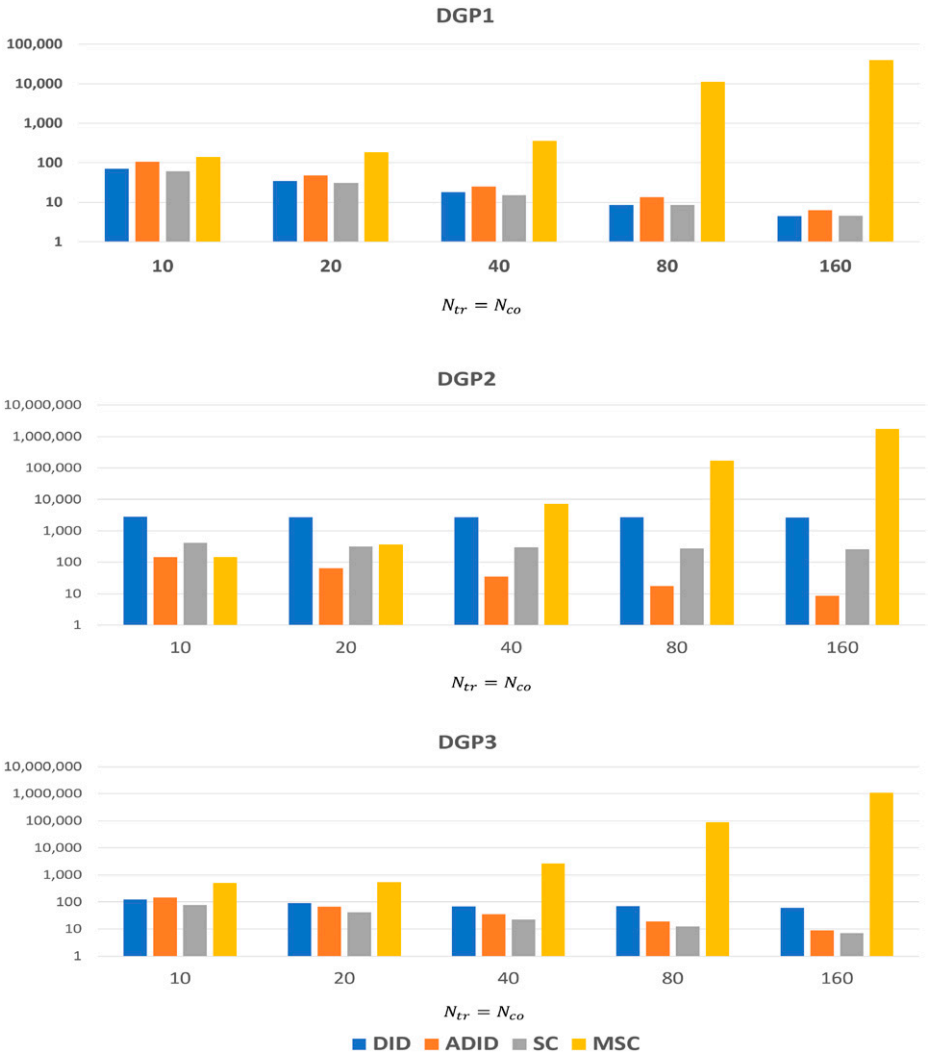


all $i = 1, \ldots, N_{tr}$. We fix $(T_1, T_2) = (10, 4)$ and choose $N_{tr} = N_{co} \in \{10, 20, 40, 80, 160\}$. We do not consider the HCW method because it breaks down in these data settings where the number of regressors (N_{co}) is larger than the pretreatment sample size (T_1). We generate the outcome variables under DGP1, DGP2, and DGP3 using stationary factors. The results using nonstationary factors (unit-root and nonlinear trend) are qualitatively similar and appear in Web Appendix E. The number of replications is 2,000 for all cases.

Figure 5 shows the MSE (multiplied by 1,000) for different methods when the time series are short. Unlike Figure 2, it reports MSE values rather than MSE ratios. It does so for two reasons. First, displaying the MSE of ADID provides confirmation that it converges to zero as $(N_{tr} + N_{co})$ increases (Proposition 3.5). Second, displaying the MSE of MSC documents that it tends to explode for large N_{tr} and N_{co}, as expected. The table corresponding to Figure 5 appears in Web Appendix E.

The DID method works well for DGP1, where the treatment and control units follow similar patterns. Its MSE decreases as N_{tr} and N_{co} increases. However, DID has large MSEs for DGP2 and DGP3, which is expected because DID is not applicable. The SC method accurately estimates the ATT for DGP1 and DGP3, and its MSE decreases quickly as N_{tr} and N_{co} increase. However, the SC method is not suitable for DGP2 and has large MSEs due to bias.

Figure 5. (Color online) MSEs (Multiplied by 1,000) by Method, for Different Large Values of $N_{co} = N_{tr}$ and Short Series ($T_1 = 10$ and $T_2 = 4$; Stationary DGPs only)



Somewhat surprisingly, the MSC method performs poorly for all DGPs. Even for DGP1, where treatment and control units have similar patterns, its MSE increases with N_{tr} and N_{co}. Because MSC estimates many parameters with a small sample size, the variance becomes very large as N_{tr} and N_{co} increase. As a result, MSC fails to reliably identify ATT when N_{tr} and N_{co} are large.

ADID is the only method that performs well for DGP1, DGP2, and DGP3. Its causal identification assumption holds for all three types of data structure, even though it has only two parameters. The MSE reduces by about one half every time N_{tr} and N_{co} are doubled, consistent with Proposition 3.5. Web Appendix E.5 presents additional simulation results showing that the ADID ATT inference theory presented in Section 3 works well for the large N_{tr} and N_{co} case.

4.5. Conclusions from the Simulation

The main conclusions of the simulation analyses are the following:

- When DID parallel trends hold and time series are long, DID and SC tend to have the lowest MSE.
- When DID parallel trends hold and time series are short, DID, ADID, and SC tend to have the lowest MSE.
- When DID parallel trends do not hold but the treatment unit's pretreatment trend is within the range of the controls' pretreatment trends, ADID and SC exhibit the lowest MSE.
- When DID parallel trends do not hold and the treatment unit's pretreatment trend is outside the range of the controls' pretreatment trends, ADID tends to exhibit the lowest MSE. This is especially so when there is only one treatment unit.
- ADID confidence intervals provide proper coverage for all nine data structures considered.

5. Illustrative Applications

We now illustrate the performance of ADID compared with DID, SC, MSC, and HCW in nine applications spanning three distinct settings: (1) how a price change affects sales of five different products in a retail chain in Brazil (Fan et al. 2022); (2) how legalizing recreational marijuana impacts cigarettes sales in two U.S. states (Bhave and Murthi 2020); and (3) how opening a showroom affects sales at a digitally native retailer in two U.S. cities (Bell et al. 2018, Li 2020).

In the first setting, the retailer randomly selected cities to receive the treatment, which was a price decrease for products I and II and a price increase for products III, IV, and V. All stores located in the same city were assigned to the same condition, though assignment varied product by product. For each product, we have daily data on quantity sold aggregated across stores for each city. In the second setting, we examine how legalizing recreational marijuana in two states, California and Washington, impacted cigarette sales in those states. We have weekly cigarettes sales by state. California and Washington are the treatment units, and 41 states that did not legalize recreational marijuana are the control units. In the third setting, we examine how opening a showroom in Boston, MA, and Columbus, OH, impacted sales of a digitally native retailer in those two cities. We have weekly sales in each city.

The data characteristics differ across these three distinct settings in terms of the number of treatment units, the number of control units, the number of pretreatment time periods (T_1), and the number of posttreatment time periods (T_2). These features of the data determine which methods (DID, ADID, SC, MSC, and HCW) are suitable for which application, even before estimating models and checking for parallel trends. Table 2 summarizes the data characteristics across the nine applications. Applying the HCW method requires that the number of pretreatment time periods is greater than the number of control units. HCW can therefore be applied only to the showroom-opening applications. These are also the only two applications allowing for inference of the SC and the MSC estimates because the inference theory for these two methods is not available if the number of control units is greater than the number of pretreatment time periods and the data are nonstationary. After excluding some methods from some applications based on these a priori considerations, we apply the remaining methods to each of the nine applications, check for additional identifying assumptions to assess if the method is appropriate for the application, and estimate the ATT, calculate its 95% confidence interval, and compute the out-of-sample or prediction MSE (PMSE).

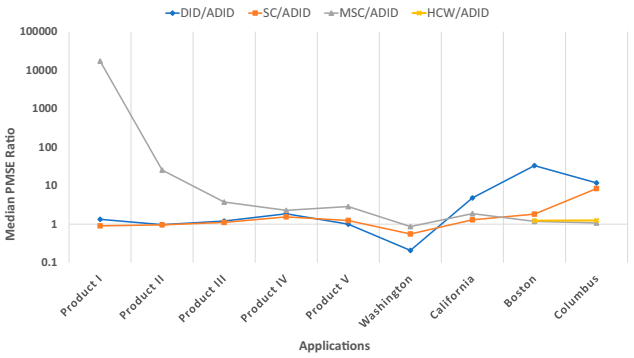
The nine applications also differ in the potential for endogeneity present in their respective settings. Since assignment to treatment and control by the Brazilian retailer was random, there are no endogeneity concerns. The second and third setting, in contrast, offer the traditional quasi-experimental design, and endogeneity may arise from the choice of treatment unit or the timing of the intervention. In the case of recreational marijuana legalization, state legislatures decided nonrandomly whether and when to pass legislation. Obviously, it is quite likely that whether a state legalized marijuana and when it did so may have been chosen in response to changing conditions or constituents' preferences. It is also conceivable that evolving conditions and preferences somehow relate to systematic differences in how cigarette sales evolved, raising endogeneity concerns. Similarly in the last two applications, endogeneity may arise from both the choice of treatment unit and the timing of the intervention.

We are not aware why selection into treatment beyond what can be detected through a violation of parallel trends would favor any one specific method we consider over any other. Hence, for the purpose of comparing the relative performance of these methods in terms of point estimates, width of confidence interval, and out-of-sample MSE, endogeneity concerns are likely moot. This likely explains why prior methodological contributions have deemed state-wide legislation and showroom openings informative to illustrate the relative merits of DID, SC, and MSC (Abadie et al. 2010, Li 2020).

Table 2. Data Characteristics of the Nine Applications

Application	T_1	T_2	Number of treatment units	Number of control units
Price reduction for Brazil product I	248	14	110	328
Price reduction for Brazil product II	248	7	100	321
Price increase for Brazil product III	248	14	97	318
Price increase for Brazil product IV	248	7	102	321
Price increase for Brazil product V	248	14	106	309
Recreational marijuana legalization in Washington	40	30	1	41
Recreational marijuana legalization in California	40	30	1	41
Showroom opening in Boston	83	27	1	10
Showroom opening in Columbus	90	20	1	10

Figure 6. (Color online) Median PMSE Ratios, Showing Each Method's Median PMSE Benchmarked Against ADID's



Before delving into each application, we provide a bird's-eye view of how well ADID performs relative to DID, SC, MSC, and HCW. We do so in terms of the out-of-sample or prediction MSE.[4] Figure 6 shows the median PMSE ratio of each method relative to ADID for each the nine applications.[5] A value greater than one indicates that the ADID method outperforms the competing method in out-of-sample prediction. In the first five applications where randomization was achieved by design and bias is likely nil, the DID, ADID, and SC methods perform similarly, whereas the MSC method performs poorly. In the four applications where the design was only quasi-experimental, the ADID method tends to outperform both the simpler DID and the more complex SC, MSC, and HCW methods. The one exception is marijuana legalization in Washington.

5.1. Price Change and Product Sales

We assess the performance of DID, ADID, and SC in investigations of how price affects sales for five products. We exclude HCW from consideration because it is not applicable when the number of pretreatment time periods is smaller than the number of control units. We also exclude MSC because it has poor out-of-sample prediction in these applications (Web Appendix F.1). This poor performance is consistent with simulation results where MSC has a large MSE when the number of control units is large relative to the number of pretreatment time periods, as is the case for these five applications.

The data set consists of daily prices and quantities sold of five products by a major retail chain in Brazil. The stores in the treatment cities all implement a price decrease for products I and II and a price increase for products III, IV, and V, while stores in the control cities keep the price as it was. We average over the stores in control versus treatment cities to calculate the average treatment effects. Fan et al. (2022) used this data set to examine heterogeneous treatment effects to each treatment city over time. Here, our aim is different, and we focus on the average change in sales across all stores in treatment cities.

Table 3. Estimated ATT for Products I–V

	ATT			ATT%		
Product	DID	ADID	SC	DID	ADID	SC
I	0.1735	0.0838	0.1162	79.24%	27.13%	33.35%
II	0.6933	0.7431	0.5651	35.41%	38.94%	23.57%
III	−3.1193	−2.953	−3.270	−18.15%	−17.35%	−19.20%
IV	−2.264	−3.327	−4.130	−7.221%	−10.26%	−10.48%
V	−1.649	−1.773	−0.8453	−14.65%	−15.58%	−7.332%

5.1.1. Treatment Effects. Because the assignment to treatment is random, treatment and control units follow similar pretreatment patterns. This scenario corresponds to DGP1 in the simulation study, where DID, ADID, and SC are all applicable and perform similarly well. We therefore expect these three methods to perform well in these five applications, too, and indeed they do. Even though the sales data display a highly nonlinear pattern, DID, ADID, and SC all fit the pretreatment data quite well. Figure 7 shows the actual and fitted/predicted sales for product I. The solid line is the actual sales, and the dashed line is the fitted sales in the pretreatment period and the predicted counterfactual sales in the posttreatment period. Similar figures for products II–V are presented in Web Appendix F.1.

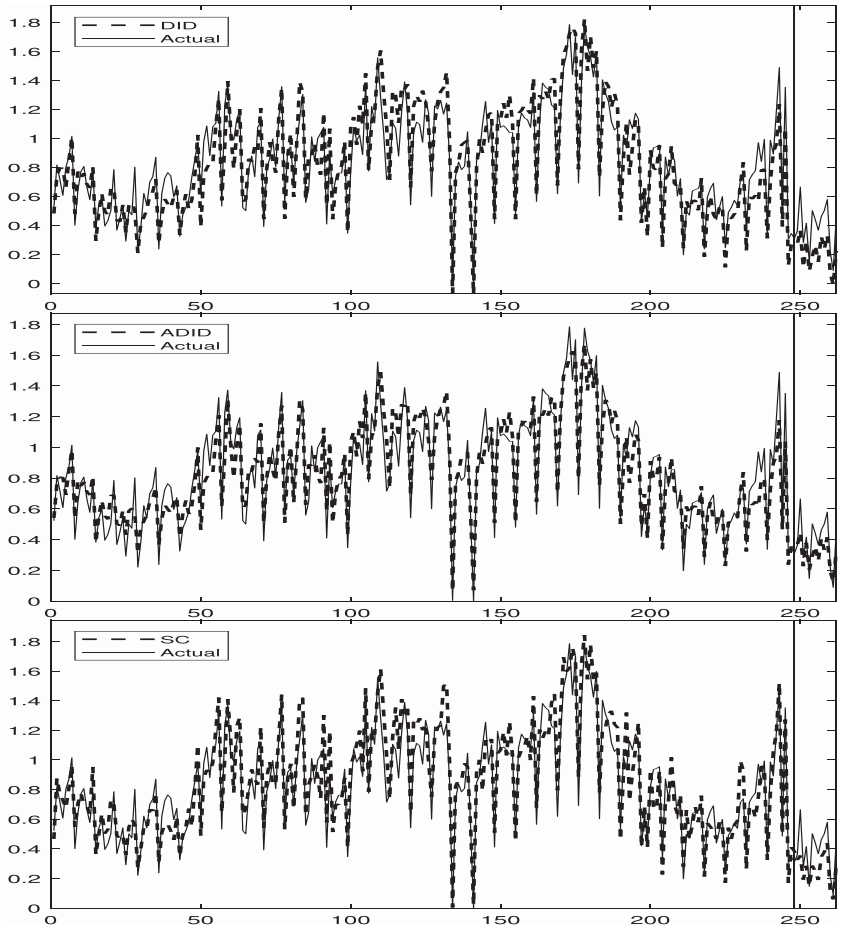
Table 3 reports the ATT estimates for the five products. They all have the expected sign (positive for the price decreases and negative for price increases), and the three methods produce similar point estimates for each product, although with some exceptions. All methods fit the in-sample data well, and they differ only slightly in out-of-sample prediction performance (Table 28 in Web Appendix F.1).

5.1.2. Confidence Intervals. Table 4 reports the 95% confidence intervals (CIs) of the DID and ADID point estimates, as well as the DID/ADID CI width ratio. We exclude SC because there is no inference theory when the number of control units is greater than the number of pretreatment time periods and the data are nonstationary. Except for product IV, both DID and ADID indicate that price changes had a statistically

Table 4. DID and ADID 95% CI for Products I–V

Product	DID CI	ADID CI	CI width ratio
I	[0.111, 0.259]	[0.0117, 0.156]	1.03
II	[0.0587, 1.33]	[0.123, 1.36]	1.02
III	[−3.69, −2.55]	[−3.48, −2.42]	1.08
IV	[−5.56, 1.03]	[−5.84, −0.810]	1.31
V	[−2.73, −0.566]	[−2.84, −0.711]	1.02

Figure 7. Sales of Product I

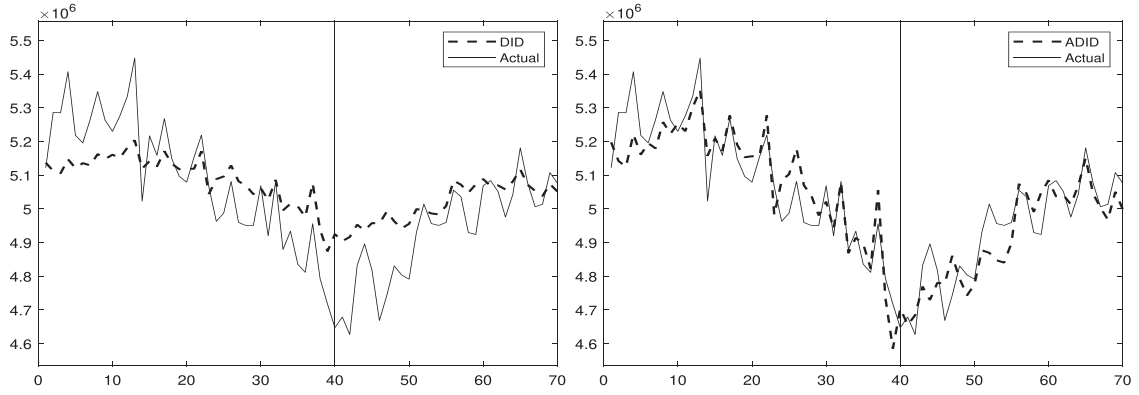


Notes. Top: DID. Middle: ADID. Bottom: SC.

significant effect on product sales. For product IV, only the ADID ATT is negative and significant at 5%. Tables 3 and 4 indicate that the statistical insignificance of the DID estimate for product IV is due to both a smaller point estimate and a wider CI. ADID has much smaller prediction MSEs than DID for that product as well (Table 28 in Web Appendix F.1).

In summary, for these five applications with an experimental rather than merely quasi-experimental design, DID, ADID, and SC perform about equally well. MSC

Figure 8. Cigarette Sales in California



Notes. Left: DID. Right: ADID.

Table 5. ATT and ATT% for DID and ADID for California and Washington

	ATT		ATT%	
State	DID	ADID	DID	ADID
California	−78,899	20,212	−1.57	0.41
Washington	−95,903	−84,757	−4.35	−3.86

performs poorly and HCW is infeasible. The main takeaway is that the additional flexibility of ADID vis-à-vis DID does not come at the cost of reduced precision or out-of-sample performance.

5.2. Marijuana Legalization and Cigarettes Sales

We now turn to how legalizing recreational marijuana affected cigarettes sales in California and Washington. Retail sales of recreational marijuana began in California on January 1, 2018, and in the state of Washington on July 1, 2014. We have 70 weeks of cigarettes sales data, with 40 weeks before and 30 weeks after treatment. The controls are 41 states in the contiguous United States that did not legalize the sale of recreational marijuana in our sample period.

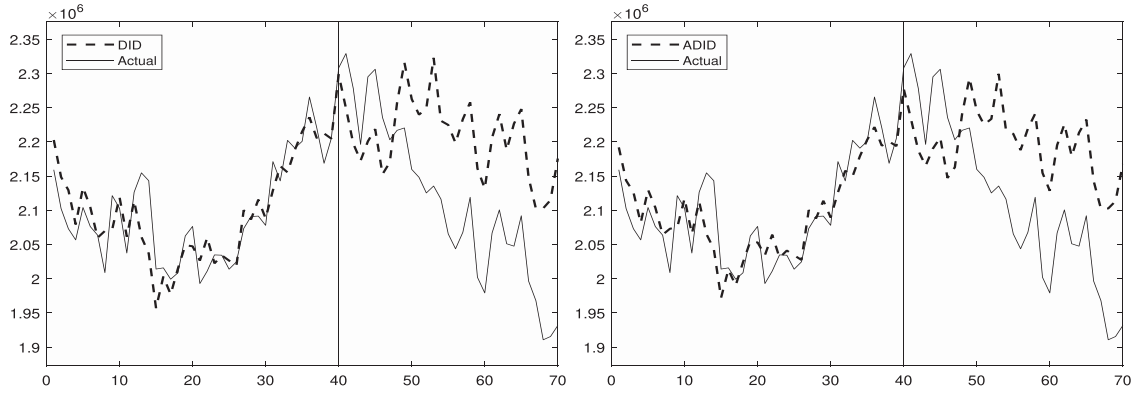
5.2.1. Treatment Effects. Since there is no inference theory for the SC and MSC methods when the number of control units is greater than the number of pretreatment time periods and data are nonstationary, we focus on the DID and the ADID methods. Web Appendix F.2 reports the SC and MSC results. The HCW method is again not applicable.

Figure 8 shows the actual and fitted/predicted sales in California obtained using DID (left panel) and ADID (right panel). The solid line is the actual sales, and the dashed line is the fitted sales in the pretreatment period and the predicted counterfactual sales in the posttreatment period. In the pretreatment periods, the downward DID trend is much less pronounced than the actual trend. This likely results in an upward bias of California's DID counterfactual sales, and a downward bias in the estimated negative ATT. The DID method suggests that weekly sales of cigarettes decreased by \$79,000, or 1.57%, after recreational marijuana was legalized in California (Table 5). In contrast, the right panel of Figure 8 shows that the ADID method fits the pretreatment data well. The ADID ATT is a negligible \$20,212, or 0.41%, increase in weekly cigarettes sales. ADID likely outperforms DID, because the pretreatment trend in California's sales is outside the range of the sales trends in the control states, the scenario corresponding to DGP2 in the simulation. ADID has the flexibility to accommodate this data pattern, whereas DID does not.

Turning our attention to Washington state, we see from Figure 9 that both the DID and ADID fit the pretreatment sales fairly tightly. Both methods perform well because the pretreatment sales pattern in Washington is similar to the average pattern in the control units, the scenario corresponding to DGP1 in the simulation study. Cigarette sales decreased by about 4% after the legalization of recreational marijuana in Washington (Table 5).

5.2.2. Confidence Intervals. Table 6 reports the 95% confidence intervals, as well as the DID/ADID CI width ratio. For California, the DID 95% CI suggests that the ATT is negative and significant at the 5% level. However, the distinct violation of the DID parallel trends assumption in Figure 8 indicates likely bias. The ADID 95% CI shows that the ATT is smaller in magnitude and not statistically different from zero at 5% significance. For Washington, in contrast, both DID and ADID indicate that legalizing recreational marijuana caused a 4% decrease in cigarettes sales, significant at the 5% level. Placebo tests reported in Web Appendix F are consistent with these results. For Washington, the DID parallel trends assumption seems to hold, and

Figure 9. Cigarettes Sales in Washington



Notes. Left: DID. Right: ADID.

Table 6. 95% CI for DID and ADID and DID/ADID CI Width Ratio for California and Washington

State	DID CI	ADID CI	CI width ratio
California	[−138,646, −19,151]	[−25,554, 65,979]	1.31
Washington	[−113,410, −7,839]	[−107,792, −6,172]	1.04

we therefore expect DID to be more precise than ADID. The former indeed outperforms the latter in out-of-sample prediction (Web Appendix F.2), but this does not translate in a narrower confidence interval (Table 6).

In summary, ADID performs better than DID for California, and the two perform about equally well for Washington. Unlike SC and MSC, both DID and ADID provide easy-to-compute confidence intervals in these two applications.

5.3. Showroom Opening and Sales

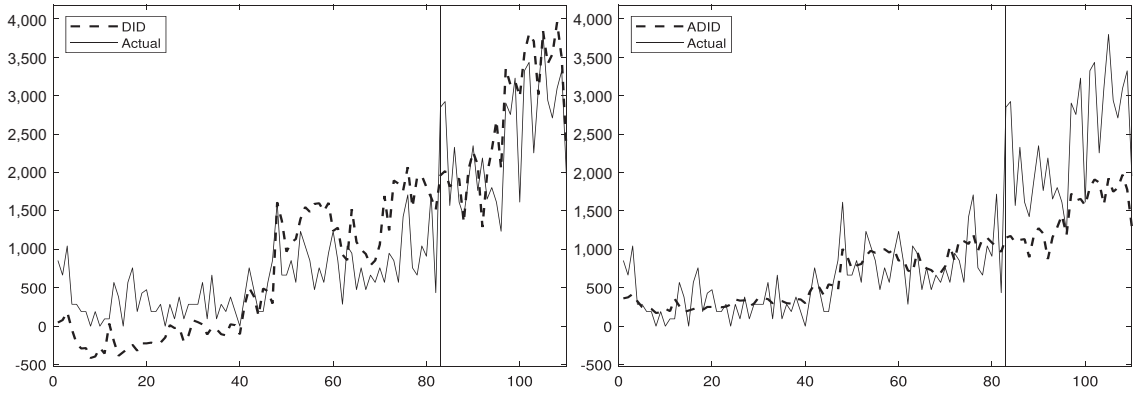
The last two applications use data from an online-first vendor offering high-quality eyeglasses at lower prices than typically encountered in the North American market ($95 vs. upward of $300). The company opened a showroom in Boston, MA, on September 22, 2011, and another in Columbus, OH, on November 10, 2011. We examine the effect of opening a brick-and-mortar showroom in a city on the sales in that same city. For the control group, we use the 10 largest markets by population without showrooms: Chicago, Houston, Portland, Seattle, Denver, Dallas, San Diego, Washington, Atlanta, and Minneapolis. The data include all transactions that occurred in those 12 cities over the 110-week period from February 2010 to March 2012. We aggregate data to the city-week level, and the dependent variable is total sales in dollars.

5.3.1. Treatment Effects. The upward trend in pretreatment sales is less steep in both Boston and Columbus than in the control cities. This pattern corresponds to DGP2 in the simulation study. Therefore, we expect ADID, MSC, and HCW to perform well because they are flexible enough to accommodate this data pattern, whereas DID and SC are not flexible enough due to their "summing to one restriction."

Figure 10 shows the actual and fitted/predicted sales for Boston using DID (left panel) and ADID (right panel). The solid line shows actual sales, and the dashed line shows the fitted sales in the pretreatment period and the predicted counterfactual sales in the posttreatment period. There are 83 pretreatment and 27 posttreatment weeks. As expected, DID does not perform well, whereas ADID does. In the pretreatment time periods, the DID fitted curve has a steeper slope than the actual sales data, which likely results in overestimating the counterfactual sales and underestimating the ATT. In contrast, ADID fits the pretreatment data well. To save space, we present plots for the SC, MSC, and HCW methods in Web Appendix F.3, and simply note here that, as expected, SC has poor in-sample fit, whereas MSC and HCW fit the in-sample data well. Table 7 reports that the three methods suitable for the data pattern in Boston produce similar point estimates of the ATT of opening a showroom.

Next, we turn to the showroom opening in Columbus. Figure 11 shows that ADID fits the pretreatment data well, whereas DID does not. The figures for the SC, MSC, and HCW methods are presented in Web Appendix F.3, and visual inspection suggests that MSC and HCW fit the pretreatment data well, whereas SC

Figure 10. Sales in Boston



Notes. Left: DID. Right: ADID.

Table 7. ATT and ATT% for Boston and Columbus by Different Methods

	ATT			ATT%		
City	ADID	MSC	HCW	ADID	MSC	HCW
Boston	946	940	973	65	64	68
Columbus	705	674	645	72	67	62

does not. We therefore focus again on ADID, MSC, and HCW in this application. The ADID estimate implies a $705 increase (72.1%) in weekly sales after a showroom is opened in Columbus (Table 7). MSC and HCW produce similar but slightly lower effect size estimates.

5.3.2. Confidence Intervals. Table 8 reports the 95% confidence intervals for the ADID, MSC, and HCW methods. The confidence intervals produced by MSC and HCW are 38%–64% wider than those obtained from ADID. This is expected because these two methods estimate more parameters without providing a notably better fit, resulting in greater estimation variances.

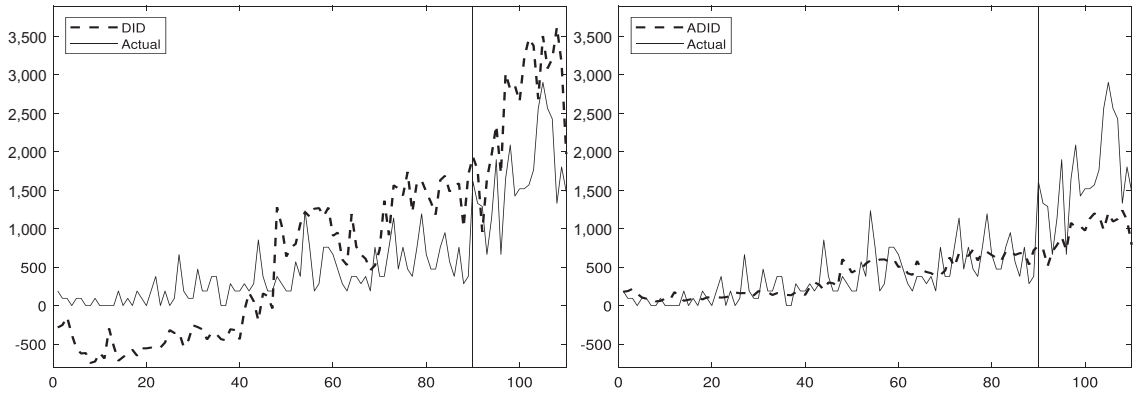
Overall, in these two final applications, ADID outperforms DID and SC in terms of bias and outperforms MSC and HCW in terms of precision. Figure 6 above and Tables 31 and 32 in Web Appendix F.3.2 show that ADID also outperforms the other methods (DID, SC, MSC, and HCW) in terms of prediction MSE in both Boston and Columbus.

6. Conclusion

Marketing scientists often estimate causal effects in pre/post test/control quasi-experimental settings. The most popular method to do so, DID, is easy to implement and provides straightforward inference but relies on a relatively restrictive assumption for causal identification. Alternatives like the SC, MSC, and HCW methods are more flexible and better able to handle differences in pretreatment trends between treatment and control units. However, this additional flexibility comes at the cost of lacking convenient inference theory or of not being applicable at all in some data structures. For example, if the number of control units is larger than the number of pretreatment time periods, then HCW is not feasible, and if the data are nonstationary in addition, then SC and MSC have no inference theory for deriving confidence intervals.

This paper introduces a new estimator, the augmented DID or ADID, that relaxes a key restriction of DID while also being simpler to implement than SC, MSC, and HCW. In addition, we develop ADID's statistical inference theory, which allows researchers to conveniently calculate confidence intervals and test hypotheses of substantive interest. Finally, we compare ADID's performance against that of DID, SC, MSC, and HCW in simulations, manipulating both the nature of heterogeneity across treatment and control units and the nature of the trend in the data (stationary, unit-root, and nonlinear), as well as in nine empirical applications. We find that ADID tends to outperform the other estimators in bias, precision, or both, in data structures with specific observable characteristics, but tends to be dominated by at least one alternative method in data structures with other specific observable characteristics. This shows that ADID complements rather than replaces extant approaches, and we provide specific guidance on what method(s) to use when. Specifically, ADID tends to be superior in data structures with large numbers of treatment and control units, with short pre- and posttreatment periods, and with large differences in pretreatment between treatment and control units. Simulations show that ADID performs well in terms of point estimates and confidence intervals for a wide range of data structures. The empirical applications show that the additional flexibility of ADID does not come at the cost of wider confidence intervals compared with DID when the latter is valid. With the addition of ADID in their toolkit,

Figure 11. Sales in Columbus



Notes. Left: DID. Right: ADID.

Table 8. 95% CI and CI Width Ratio for Boston and Columbus by Different Methods

	95% CI			CI width ratio	
City	ADID	MSC	HCW	MSC/ADID	HCW/ADID
Boston	[661, 1,229]	[632, 1,415]	[507, 1,438]	1.38	1.64
Columbus	[448, 961]	[393, 1,087]	[260, 1,030]	1.53	1.50

marketers are better equipped to address important causal research questions in a wider range of data structures.

Competing Interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or nonfinancial interest in the subject matter or materials discussed in this manuscript. The authors have no funding to report.

Acknowledgments

The authors are indebted to David Bell for stimulating discussions in the early stages of the research. They thank the review team, Eric Bradlow, Dylan Small, Venkatesh Shankar, and Raji Srinivasan for valuable comments; Sachin Sridhar for excellent research assistance; and Marcelo Medeiros and an eyewear company for sharing data. Some of the analyses presented here were calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at the University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researchers and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

Appendix. Variance Estimator and Prediction MSE Calculation

A.1. Variance Estimator

This appendix provides expressions of variance estimator $\hat{\Sigma}$ under different conditions. Web Appendix C proves that, for Proposition 3.1, the asymptotic variance of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$ consists of two parts: $\Sigma = \Sigma_1 + \Sigma_2$, where Σ_1 is related to pretreatment estimation error and Σ_2 corresponds to the posttreatment idiosyncratic errors. We can consistently estimate Σ_1 and Σ_2 by

$$\hat{\Sigma}_1 = (T_2/T_1)B\hat{V}B, \quad \hat{\Sigma}_2 = T_1^{-1}\sum_{t=1}^{T_1}\sum_{s=1,\,|s-t|\le l_2}^{T_1}\hat{e}_{1t}\hat{e}_{1s}, \qquad (A.1)$$

where $B = T_2^{-1}\sum_{t=T_2+1}^{T} x_t'[T_1^{-1}\sum_{t=1}^{T_1} x_t x_t']^{-1}$, $\hat{V} = T_1^{-1}\sum_{t=1}^{T_1}\sum_{s=1,|s-t|\le l_1}^{T_1}\hat{e}_{1t}\hat{e}_{1s}x_t x_s'$, $\hat{e}_{1t} = y_{1t} - x_t'\hat{\delta}$ ($\hat{\delta}$ is defined below (2.10)), $l_1 = O(T_1^{1/4})$ and $l_2 = O(T_2^{1/4})$.

If e_{1t} only displays moving average correlation of order q, that is, $E(e_{it}e_{1s}) = 0$ if $|t-s| > q$, then, $\hat{V}$ and $\hat{\Sigma}_2$ simplify to $\hat{V} = T_1^{-1}\sum_{t=1}^{T_1}\sum_{s,|s-t|\le q}\hat{e}_{1t}\hat{e}_{1s}x_t x_s'$ and $\hat{\Sigma}_2 = T_1^{-1}\sum_{t=1}^{T_1}\sum_{s,|s-t|\le q}\hat{e}_{1t}\hat{e}_{1s}$.

If e_{1t} is serially uncorrelated, then $\hat{V}$ and $\hat{\Sigma}_2$ further simplify to $\hat{V} = T_1^{-1}\sum_{t=1}^{T_1}\hat{e}_{1t}^2 x_t x_t'$ and $\hat{\Sigma}_2 = T_1^{-1}\sum_{t=1}^{T_1}\hat{e}_{1t}^2$.

For the DID ATT estimate, following the same steps in the proof of Proposition 3.1 we can show that, if the DID parallel trends assumption holds, then $\sqrt{T_2}(\hat{\Delta}_{1,DID} - \Delta_1)/\sqrt{\hat{\Sigma}_{DID}} \xrightarrow{d} N(0,1)$, where $\hat{\Sigma}_{DID} = \hat{\sigma}_{DID}^2(T_2/T_1 + 1)$, $\hat{\sigma}_{DID}^2 = T_1^{-1}\sum_{t=1}^{T_1}\hat{e}_{1t,DID}^2$, $\hat{e}_{1t,DID} = y_{1t} - \hat{y}_{1t,DID}^0$.

A.2. Prediction MSE

We compare the prediction performance of different methods using backdating (Li 2020, Abadie 2021). This method uses the pretreatment data ($t \le T_1$) and proceeds as if the intervention occurred at an earlier time period, $T_0 + 1$, where $1 < T_0 < T_1$. Therefore, the first T_0 time periods make up the new pretreatment period, and the remaining $T_1 - T_0$ time periods are the new posttreatment time period in this backdating procedure. However, in actuality, there is no treatment during the posttreatment time, and the predicted counterfactual outcomes should be close to the actual outcomes. As a metric of how close they are, we use the prediction mean squared error (PMSE), which reflects the mean squared difference between the actual outcomes and the predicted counterfactuals over the last $T_1 - T_0$ time periods.

We first describe how to compute the PMSE when there are multiple treatment units, $N_{tr} > 1$ (as with the Brazilian product price changes). Because there is no treatment for $t \le T_1$, we can select a $T_0 \in \{1, \ldots, T_1 - 1\}$ and treat $T_0 + 1$ as a pseudo treatment time and set T_1 as the terminal period of the sample. We then fit the model using pretreatment data $1 \le t \le T_0$ and predict the counterfactual outcome $\bar{y}_{t,tr}^0 = N_{tr}^{-1}\sum_{i=1}^{N_{tr}} y_{it,tr}$ for $t \in T_0 + 1, \ldots, T_1$. Let $\hat{y}_{t,tr}^0$ denote the predicted value of $\bar{y}_{t,tr}^0$. Since there is no treatment for $t \le T_1$, we actually observe $\bar{y}_{t,tr}^0$ for $t \in \{T_0 + 1, \ldots, T_1\}$. Therefore, we can compute prediction MSE (PMSE) by

$$\text{PMSE} = \frac{1}{T_1 - T_0}\sum_{t=T_0+1}^{T_1}(\bar{y}_{t,tr} - \hat{y}_{t,tr}^0)^2, \qquad (A.2)$$

Where $\hat{y}_{t,tr}^0$ is one of the four predictors using the DID, ADID, SC, and MSC methods. We use PMSE_{DID}, $\text{PMSE}_{\text{ADID}}$, PMSE_{SC}, and PMSE_{MSC} to denote the resulting PMSEs.

For empirical applications with one treatment unit ($N_{tr} = 1$), we use $i = 1$ to denote the treatment unit and need to replace $\bar{y}_{t,tr}^0$ and $\hat{y}_{tr,t}^0$ above by y_{1t}^0 and $\hat{y}_{1t}^0$, respectively. Then, $PMSE = \frac{1}{T_1 - T_0}\sum_{t=T_0+1}^{T_1}(y_{1t}^0 - \hat{y}_{1t}^0)$. We also calculate the PMSE for a wide range of values of T_0. We choose the number of pretreatment time periods (T_0) to be greater than the number of posttreatment periods ($T_1 - T_0$) and set the difference between the different values of the pretreatment time periods (T_0) to be no

larger than 10. Note that when $T_0 > N_{co}$, we can also compute PMSE for the HCW method.

Endnotes

[1] Recently, Rambachan and Roth (2022) propose a robust inference method for a DID estimator when the DID parallel trends assumption is mildly violated. It may be possible to generalize Rambachan and Roth's (2022) result to cover the ADID ATT estimate when the ADID's parallel pretrend assumption is mildly violated. We leave this interesting and challenging topic to a possible future research topic.

[2] Because the SC method does not include an intercept, it is not enough to require that the treatment would have been parallel to the weighted average of the control units' outcomes. For simplicity, we refer to the SC assumption as the *SC parallel trends assumption*, although it is a much stronger assumption involving overlap.

[3] Web Appendix E.3 reports additional simulations allowing factor loadings b_i, $i \geq 2$, to be random draws from a nondegenerate distribution, so that b_i varies with i for $i = 2, \ldots, N$. Results show that the ADID method leads to consistent estimation results. This further documents that ADID allows the treatment and controls to follow different patterns.

[4] The MSE calculation in the simulation pertained to the difference between the estimated and the true treatment effects. In the empirical applications where the true effect is unknown, the PMSE pertains to the difference between the predicted and observed sales levels. We describe how we calculate PMSE in Appendix A.2 and report the results in the tables in Web Appendix F.

[5] Each PMSE ratio reported in Figure 6 is the median of several PMSE ratios computed over different out-of-sample window widths. The individual window-specific ratios are reported in Web Appendix F.

References

Abadie A (2021) Using synthetic controls: Feasibility, data requirements, and methodological aspects. *J. Econom. Literature* 59(2): 391–425.

Abadie A, Gardeazabal J (2003) The economic costs of conflict: A case study of the Basque Country. *Amer. Econom. Rev.* 93(1): 113–132.

Abadie A, Diamond A, Hainmueller J (2010) Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *J. Amer. Statist. Assoc.* 105(490): 493–505.

Bell D, Gallino S, Moreno A (2018) Offline showrooms and customer migration in omni-channel retail. *Management Sci.* 64(4): 1629–1651.

Berman R, Israeli A (2022) The value of descriptive analytics: Evidence from online retailers. *Marketing Sci.* Forthcoming.

Bhave A, Murthi BPS (2020) A study of the effects of legalization of recreational marijuana on consumption of cigarettes. Preprint, submitted January 13, https://doi.org/10.2139/ssrn.3508422.

Bonfrer A, Chintagunta PK, Roberts JH, Corkindale D (2020) Assessing the sales impact of plain packaging regulation for cigarettes: Evidence from Australia. *Marketing Sci.* 39(1):234–252.

Cheon H, Guo T, Manchanda P, Sriram S (2021) The impact of medical marijuana legalization on opioid prescriptions. Preprint, submitted October 12, https://doi.org/10.2139/ssrn.3917975.

Doudchenko N, Imbens GW (2016) Balancing, regression, difference-in differences and synthetic control methods: A synthesis. NBER Working Paper No. 22791, National Bureau of Economic Research, Cambridge, MA.

Fan J, Masini R, Medeiros MC (2022) Do we exploit all information for counterfactual analysis? Benefits of factor models and idiosyncratic correction. *J. Amer. Statist. Assoc.* 117(538):574–590.

Hamilton J (1994) *Time Series Analysis* (Princeton University Press, Princeton, NJ).

Hayashi F (2000) *Econometrics* (Princeton University Press, Princeton, NJ).

Holland P (1986) Statistics and causal inference. *J. Amer. Statist. Assoc.* 81(396):945–960.

Hsiao C, Ching H, Wan S (2012) A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong Kong with mainland China. *J. Appl. Econometrics* 27(5):705–740.

Iyengar R, Park Y-H, Yu Q (2022) The impact of subscription programs on customer purchases. *J. Marketing Res.* Forthcoming.

Kahn-Lang A, Lang K (2020) The promise and pitfalls of differences-in-differences: Reflections on '16 and Pregnant' and other applications. *J. Bus. Econom. Statist.* 38(3):613–620.

Kim S, Lee C, Gupta S (2020) Bayesian synthetic control methods. *J. Marketing Res.* 57(5):831–852.

Li K (2020) Statistical inference for average treatment effects estimated by synthetic control methods. *J. Amer. Statist. Assoc.* 115(532):2068–2083.

Li K, Bell D (2017) Estimation of the average treatment effect with panel data: Asymptotic theory and implementation. *J. Econometrics* 197:65–75.

Mullainathan S, Spiess J (2017) Machine learning: An applied econometric approach. *J. Econom. Perspect.* 31(2):87–106.

Petrova M, Sen A, Yildirim P (2021) Social media and political contributions: The impact of new technology on political competition. *Management Sci.* 67(5):2997–3021.

Rambachan A, Roth J (2022) A more credible approach to parallel trends. Working Paper, Harvard University, Cambridge, MA.