Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods

Kathleen T. Li

To cite this article: Kathleen T. Li (2020) Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods, Journal of the American Statistical Association, 115:532, 2068-2083, DOI: 10.1080/01621459.2019.1686986

To link to this article: https://doi.org/10.1080/01621459.2019.1686986

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 532, 2068–2083: Theory and Methods
https://doi.org/10.1080/01621459.2019.1686986

Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods

Kathleen T. Li

Marketing Department, McCombs School of Business, University of Texas at Austin, Austin, TX

ABSTRACT

The synthetic control (SC) method, a powerful tool for estimating average treatment effects (ATE), is increasingly popular in fields such as statistics, economics, political science, and marketing. The SC is particularly suitable for estimating ATE with a single (or a few) treated unit(s), a fixed number of control units, and large pre and post-treatment periods (which we refer as "long panels"). To date, there has been no formal inference theory for SC ATE estimator with long panels under general conditions. Existing work mostly use placebo tests for inference or some permutation methods when the post-treatment period is small. In this article, we derive the asymptotic distribution of the SC and modified synthetic control (MSC) ATE estimators using projection theory. We show that a properly designed subsampling method can be used to obtain confidence intervals and conduct inference whereas the standard bootstrap cannot. Simulations and an empirical application that examines the effect of opening a physical showroom by an e-tailer demonstrate the usefulness of the MSC method in applications. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2018
Accepted October 2019

KEYWORDS
Inference; Projection theory; Subsampling; Synthetic control method

1. Introduction

Identifying average treatment effects (ATE) from quasi-experimental data has become one of the most important endeavors of social scientists over the last three decades. It has proven to be one of the most challenging as well. The difficulty lies in accurately estimating the counterfactual outcomes for the treated units in the absence of treatment. Early literature on examining treatment effects focused on evaluating the effectiveness of education and labor market programs (Ashenfelter 1978; Ashenfelter and Card 1985) and the effect of minimum wage on unemployment (Card and Krueger 1994). More recently, researchers have used quasi-experimental data to evaluate many diverse topics such as the effect of Internet information on financing terms for new cars (Busse, Silva-Risso, and Zettelmeyer 2006); effect of school term length on student performance (Pischke 2007); price reactions to rivals' local channel exits (Ozturk, Venkataraman, and Chintagunta 2016); offline bookstore openings' effect on sales at Amazon (Forman, Ghose, and Goldfarb 2009), and consumer relocation on brand preferences (Bronnenberg, Dubé, and Gentzkow 2012). Others have used difference-in-differences (DID) methods to examine various treatment effects, especially in digital environments, such as executional strategies for display advertising (Goldfarb and Tucker 2011); online information on consumers' strategic behavior (Mantin and Rubin 2016); and how offline stores drive online sales (Wang and Goldfarb 2017). See Imbens and Wooldridge (2009) for more examples.

DID and the propensity score matching methodologies are perhaps the most popular approaches used to estimate treatment effects. These methods are especially effective when there are a large number of treatment and control units over short time periods. One crucial assumption for the DID method is that outcomes of the treated and control units follow parallel trends in the absence of treatment. Violation of this parallel trends assumption in general will result in biased estimates. For panel data with a relatively large number of time series observations, alternative methods may be better suited than DID for estimating counterfactual outcomes. For example, the synthetic control (SC) method proposed by the seminal work of Abadie and Gardeazabal (2003), and Abadie, Diamond, and Hainmueller (2010) can be used to successfully estimate ATE. This method has many attractive features: First, it is more general than the conventional DID method because it allows for different control units to have different weights (individual specific coefficients) when estimating the counterfactual outcome of the treated unit. Second, the SC method restricts the weights assigned to the control group to be nonnegative (because outcome variables are likely to be positively correlated) and may lead to better extrapolation. Third, when there exists a large number of control units, the SC method can reduce the estimation variance by dropping some controls that are (often weakly) negatively correlated or uncorrelated with the treated unit. In fact, Athey and Imbens (2017) describe the SC method as "arguably the most important innovation in the evaluation literature in the last 15 years."

CONTACT Kathleen T. Li *kathleen.li@mccombs.utexas.edu* Marketing Department, McCombs School of Business, University of Texas at Austin, 2110 Speedway Stop B6700, Austin, TX 78712-1275.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

Abadie, Diamond, and Hainmueller (2010) suggested that the potential applicability of the SC method to comparative case studies is very broad (Abadie, Diamond, and Hainmueller 2010, p. 493). However, this method is not without some limitations. For example, the restriction that the sum of the weights assigned to the controls equal to one implicitly requires that outcomes for the treated unit and a weighted average of control units follow parallel trends over time in the absence of treatment (e.g., Abadie 2005, Assumption 3.1). For many social science datasets, this "parallel trends" assumption may not hold.

Doudchenko and Imbens (2016) proposed a modified synthetic control (MSC) method. Their proposed modifications are adding an intercept and dropping the slope coefficients sum-to-one restriction in the original synthetic control (OSC) model. Dropping these restrictions makes the modified method applicable to a wider range of data settings. We show that even when the "parallel line" assumption is violated and the OSC method cannot be used to consistently estimate ATE, it is likely that the MSC can be used to deliver reliable ATE estimation results because the MSC method can adjust the slope of a linear combination of the control outcomes and make it parallel to the treated unit's outcome sample path (in the absence of treatment). We use simulated and real data to demonstrate the improvement of the MSC over the OSC method when the treated and control units are drawn from heterogeneous distributions.

To date, there has been no formal inference theory for the SC and MSC ATE estimators with long panels under general conditions. Currently, applications of SC mostly use placebo tests that rely on the assumption that the treatment units are randomly assigned or other permutation methods that can only be applied when the post-treatment sample size is small. Hahn and Shi (2017) showed that the validity of using placebo tests requires a strong normality distribution assumption for the idiosyncratic error terms under a factor model data generating framework. Conley and Taber (2011) and Ferman and Pinto (2016, 2018) proposed rigorous inference methods for DID and SC ATE estimators under different conditions. Conley and Taber (2011) assumed that there is only one treated unit and a large number of control units and that the idiosyncratic errors from the treated and the control units are identically distributed (a sufficient condition for this is random assignment of the treated unit). Conley and Taber (2011) showed that proper inference for the DID ATE estimator can be conducted using the control units' information. Their method allows for both the pre and the post-treatment periods to be small. Assuming instead that the pretreatment period is large and the post-treatment period is small, Ferman and Pinto (2016, 2018) showed that Andrews' (2003) end-of-sample instability test can be used to conduct inference for ATE estimators without requiring the random assignment to the treated unit assumption. Chernozhukov, Wuthrich, and Zhu (2017) recently proposed a general inference procedure for a number of different ATE estimators, including DID, SC, and a factor-model-based method. They analyze two situations: (1) assuming that the idiosyncratic error term satisfies an exchangeability condition (e.g., iid), the authors use a permutation inference method for achieving exact finite sample size; (2) if the data are dynamic and serially correlated, they instead use an inference procedure that achieves approximate uniform size control for the case of a large pretreatment sample and a small post-treatment sample. The exchangeability assumption is strong and may not be plausible in many applications. Further, for many data settings, the post-treatment sample period may not be particularly small when compared to the pretreatment sample. Therefore, for this type of data, inference methods based on small post-treatment sample size will be invalid.

In this article, we focus on a different set up. We consider the case where there is only one (or a few) treated unit(s), a fixed number of control units, and large pre- and post-treatment sample sizes (long panel). We make three contributions. First, and most importantly, using projection theory (Zarantonello 1971; Fang and Santos 2018), we derive the formal inference theory for the SC and MSC method ATE estimator with long panels. The asymptotic distribution is nonnormal and nonstandard, and standard bootstrapping breaks down (Andrews 2000; Fang and Santos 2018). Second, we show that a carefully designed subsampling method, that is, applying the subsampling method only to part of the statistic, provides valid inferences. Finally, we provide a simple sufficient condition under which the SC and MSC estimator is uniquely determined and show via simulations and an empirical example that a MSC method, which is robust to "nonparallel trends" situations, can greatly enhance the applicability of the SC method to estimating ATE. Therefore, our work complements the existing inference work based on small post-treatment sample size (e.g., Andrews 2003; Ferman and Pinto 2016; Chernozhukov, Wuthrich, and Zhu 2017).

A closely related approach to the SC method is the one proposed by Hsiao, Ching, and Wan (2012, hereafter, HCW) which uses the unconstrained least squares method to replace SC's constrained estimation procedure. The asymptotic theory of HCW's ATE estimator were developed by Hsiao, Ching, and Wan (2012) and Li and Bell (2017). Since the limiting distribution of the (modified) SC estimator is a projection of the limiting distribution of unconstrained least-squares (HCW) estimator onto a convex set (referred to as a tangent cone), we can also interpret our inference theory as an extension of Hsiao, Ching, and Wan (2012) and Li and Bell (2017) to the case of using nonnegative weights in constructing the counterfactual estimate.

2. Estimating ATE Using Panel Data

We start by introducing some notation. Let y_{it}^1 and y_{it}^0 denote unit i's outcome in period t with and without treatment, respectively. The treatment effect from intervention for the ith unit at time t is defined as $\Delta_{it} = y_{it}^1 - y_{it}^0$. However, we do not simultaneously observe y_{it}^0 and y_{it}^1. The observed data is in the form $y_{it} = d_{it}y_{it}^1 + (1 - d_{it})y_{it}^0$, where $d_{it} = 1$ if the ith unit is under the treatment at time t, and $d_{it} = 0$ otherwise.

We consider the case where there is a finite number of treated and control units and the treated units are drawn from heterogeneous distributions (i.e., they are not randomly assigned). Also, the treatment time occurs at different times for different treated units. In this type of situation, it is reasonable to estimate treatment effects for each treated unit separately. Under the assumption that the treatment effects $\Delta_{it} = y_{it}^1 - y_{it}^0$ follow a

stationary process, we can define the ATE as $\Delta_i = E(\Delta_{it})$, where the expectation is with respect to the stationary distribution of Δ_{it}. In this way, we can obtain ATE for each treated unit. To obtain ATE over all the treated units, we can average (possibly with different weights) over all treated units. Hence, in this article, we focus on the case where there is one treated unit that receives a treatment at time $T_1 + 1$. Without loss of generality, we assume that it is the first unit. We want to estimate ATE for the first unit: $\Delta_1 = E(\Delta_{1t})$. The difficulty in estimating the treatment effects is that y_{1t}^0 is not observable for $t \geq T_1 + 1$. Specific methods for estimating y_{1t}^0 are discussed in subsequent sections. For now, let $\hat{y}_{1t}^0$ be a generic estimator of y_{1t}^0. Then ATE is estimated by averaging over the post-treatment period,

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

where $T_2 = T - T_1$ is the post-treatment sample size.

2.1. The Synthetic Control Method

We examine the scenario where a treatment was administered to the first unit at $t = T_1 + 1$. Thus, the remaining $N - 1$ units are control units. To use unified notation to cover both the SC and the MSC methods, we add an intercept to the OSC method. Therefore, utilizing the correlation between y_{1t} and y_{jt} where $j = 2, \ldots, N$, we can estimate the SC counterfactual outcome y_{1t}^0 based on the following regression model:

$$y_{1t} = x_t' \beta_0 + u_{1t}, \quad t = 1, \ldots, T_1, \tag{1}$$

where $x_t = (1, y_{2t}, \ldots, y_{Nt})'$ is an $N \times 1$ vector of a constant (of one) and the control units' outcome variables, $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,N})'$ is an $N \times 1$ vector of unknown coefficients, and u_{1t} is a zero mean, finite variance idiosyncratic error term. Essentially, we can think of all the outcomes as correlated with some common factors. Our analysis below also applies to the OSC method by removing the intercept from (1), that is, redefine x_t as $(y_{2t}, \ldots, y_{Nt})'$ with β being an $(N-1) \times 1$ vector of coefficients.

Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010) proposed a SC method that uses a weighted average of the control units to approximate the sample path of the treated unit. The weights are selected by best fitting the outcome of the treated unit using pretreatment data, and the weights are nonnegative and sum to one. Specifically, $\beta = (\beta_1, \ldots, \beta_N)'$ is selected via the following constrained minimization problem:

$$\hat{\beta}_{T_1,\mathrm{SC}} = \arg \min_{\beta \in \Lambda_{\mathrm{SC}}} \sum_{t=1}^{T_1} \left[y_{1t} - x_t' \beta \right]^2, \tag{2}$$

where $\Lambda_{\mathrm{SC}} = \{\beta \in \mathcal{R}^N : \beta_j \geq 0 \text{ for } j = 2, \ldots, N \text{ and } \sum_{j=2}^{N} \beta_j = 1\}$. With $\hat{\beta}_{T_1,\mathrm{SC}}$ defined as the minimizer to (2), the SC fitted/predicted curve is

$$\hat{y}_{1t,\mathrm{SC}}^0 = x_t' \hat{\beta}_{T_1,\mathrm{SC}}, \quad t = 1, \ldots, T_1, T_1 + 1, \ldots, T. \tag{3}$$

For $t = 1, \ldots, T_1$, $\hat{y}_{1t,\mathrm{SC}}^0$ is the in-sample fitted curve, and for $t = T_1+1, \ldots, T$, $\hat{y}_{1t,\mathrm{SC}}^0$ is the predicted counterfactual outcome of y_{1t}^0. The ATE is estimated by $\hat{\Delta}_{1,\mathrm{SC}} = T_2^{-1} \sum_{t=T_1+1}^{T} (y_{1t} - \hat{y}_{1t,\mathrm{SC}}^0)$. Abadie, Diamond, and Hainmueller (2010) also suggested using covariates to improve the fit when relevant covariates are available. Adding covariates to the model is straightforward. To focus on the main issue of the article, we will consider the case without any relevant covariates and discuss how to add relevant covariates in Supplementary Appendix G.

2.2. The Modified Synthetic Control Method

For many quasi-experimental data used in economics, political science, marketing, and other social science fields, the treated unit and the control units may exhibit substantial heterogeneity and the treated unit's outcome and a weighted average (with weights summing to one) of the control units' outcomes may not follow parallel trends in the absence of treatment. In this section, we consider two simple modifications advocated by Doudchenko and Imbens (2016). Specifically, we add an intercept and remove the coefficients sum to one restriction in the OSC model, that is, we still keep the nonnegative constraints: $\beta_j \geq 0$ for $j = 2, \ldots, N$ but drop the restriction $\sum_{j=2}^{N} \beta_j = 1$. When the sum of the estimated weights (coefficients) is far from one, we suggest not imposing the sum-to-one restriction. Therefore, the MSC method is the same as (2) except that the sum-to-one restriction on the slope coefficients is removed, that is, solve the following (constrained) minimization problem:

$$\hat{\beta}_{T_1,\mathrm{MSC}} = \arg \min_{\beta \in \Lambda_{\mathrm{MSC}}} \sum_{t=1}^{T_1} \left[y_{1t} - x_t' \beta \right]^2, \tag{4}$$

where x_t and β are defined the same way as in (2), and $\Lambda_{\mathrm{MSC}} = \{\beta \in \mathcal{R}^N : \beta_j \geq 0 \text{ for } j = 2, \ldots, N\}$. Let X be the $T_1 \times N$ matrix with its tth row given by $x_t' = (1, y_{2t}, \ldots, y_{Nt})$. We show in the Supplementary Appendix B that when X has full column rank (which requires that $T_1 \geq N$), the SC minimizers, $\hat{\beta}_{T_1,\mathrm{SC}}$ and $\hat{\beta}_{T_1,\mathrm{MSC}}$, are uniquely defined. With $\hat{\beta}_{T_1,\mathrm{MSC}}$ defined in (4), the counterfactual outcome is estimated by $\hat{y}_{1t,\mathrm{MSC}}^0 = x_t' \hat{\beta}_{T_1,\mathrm{MSC}}$ for $t = T_1 + 1, \ldots, T$, and the ATE is estimated by $\hat{\Delta}_{1,\mathrm{MSC}} = T_2^{-1} \sum_{t=T_1+1}^{T} \left[y_{1t} - \hat{y}_{1t,\mathrm{MSC}}^0 \right]$.

We would like to emphasize here that Equation (1) should be viewed as a projection model. The model allows for $E(u_{1t}|x_t) \neq 0$. The usual endogeneity estimation bias problem does not arise here. The interpretation of $x_t' \beta_0$ is that it is a projection of y_{1t} onto the space of linear function of x_t (i.e., $x_t' \beta_0$) subject to $\beta_0 \in \Lambda_{\mathrm{MSC}}$. Under the assumption that $u_t = y_t - x_t' \beta_0$ is a zero mean, weakly dependent stationary process, we can show that $\hat{\beta}_{T_1,\mathrm{MSC}} = \arg\min_{\beta \in \Lambda_{\mathrm{MSC}}} \sum_{t=1}^{T_1} (y_{1t} - x_t' \beta)^2$ consistently estimates $\beta_0 = \arg\min_{\beta \in \Lambda_{\mathrm{MSC}}} E[(y_{1t} - x_t' \beta)^2]$. As a result, $\hat{\Delta}_{1,\mathrm{MSC}}$ consistently estimates ATE, Δ_1. We can drop (partially) negatively correlated controls without affecting MSC's consistent ATE estimation results. This differs from conventional regression models where biased and inconsistent estimation results result from omitting some relevant explanatory variables.

However, the additional slope coefficients sum-to-one constraint required by the SC method can lead to biased estimation results if such a constraint is improper to impose. Consider a simple case where one regressor has a time trend component. For example, $x_{2t} = t + v_{2t}$, where v_{2t} is weakly dependent

stationary process, and $\beta_{0,2} > 1$. The SC method necessarily restricts $\hat{\beta}_{2,T_1,\mathrm{SC}} \in [0,1]$. Then even when $\Delta_1 = 0$, we have $\hat{\Delta}_{1,\mathrm{SC}} = T_2^{-1}\sum_{t=T_1+1}^{T}(y_{1t} - x_t'\hat{\beta}_{T_1,\mathrm{SC}}) \approx T_2^{-1}\sum_{t=T_1+1}^{T} t(\beta_{0,2} - \hat{\beta}_{2,T_1,\mathrm{SC}}) = (1/2)(2T_1 + T_2 + 1)(\beta_{0,2} - \hat{\beta}_{2,T_1,\mathrm{SC}}) \to \infty$ as T_1 or $T_2 \to \infty$ because $\beta_{0,2} > 1 \geq \hat{\beta}_{2,T_1,\mathrm{SC}}$. Thus, the SC ATE estimation result is biased and inconsistent in such a case.

2.3. The Least Squares Based Method

Hsiao, Ching and Wan's (2012, HCW) suggested estimating β_0 in model (1) with the least squares method. Let $\hat{\beta}_{T_1,\mathrm{OLS}}$ denote the least squares estimator of β_0 using the pretreatment period data, that is,

$$\hat{\beta}_{T_1,\mathrm{OLS}} = \arg\min_{\beta\in\mathcal{R}^N} \sum_{t=1}^{T_1} \left[y_{1t} - x_t'\beta\right]^2. \tag{5}$$

Then HCW's ATE estimator is given by $\hat{\Delta}_{1,\mathrm{HCW}} = T_2^{-1}\sum_{t=T_1+1}^{T}(y_{1t} - \hat{y}^0_{1t,\mathrm{HCW}})$, where $\hat{y}^0_{1t,\mathrm{HCW}} = x_t'\hat{\beta}_{T_1,\mathrm{OLS}}$.

HCW method can be an effective approach to estimate ATE when the number of control units is small. However, when the number of control units is larger than the number of pretreatment time periods, a unique weight vector, β, that minimizes (5) may not exist. In such cases, adding restrictions (by the SC or MSC methods) such as regularizing the weights to be nonnegative and sum to one, may reduce the number of parameters and help identify the model. Another rationale for imposing nonnegativity restriction is that in most applications, y_{jt}'s are positively correlated with each other and therefore tend to move up or down together. The sum-to-one restriction $\sum_{j=2}^{N}\beta_j = 1$ introduced by Abadie, Diamond, and Hainmueller (2010) implicitly assumes that a weighted average of the control units' outcomes and the treated unit's outcome would have followed parallel trends over time in the absence of treatment. The restriction that the slope coefficients sum to one can improve the out-of-sample extrapolation when the "parallel trends" assumption holds. However, in general, the slope coefficient sum to one restriction should be considered on its merit rather than a rule, as discussed in Doudchenko and Imbens (2016).

3. Distribution Theory

3.1. A Projection of the Unconstrained Estimator

To study the distribution theory of the SC ATE estimator, we first show that we can express the constrained estimator as a projection of the unconstrained (the ordinary least squares) estimator onto a constrained set. Then we use the theory of projection onto convex sets to derive the asymptotic distribution of the SC ATE estimator.

Let $\hat{\beta}_{\mathrm{OLS}}$ denote the ordinary least squares estimator of β_0 using data $\{y_{1t}, x_t\}_{t=1}^{T_1}$. We show in the Supplementary Appendix B that the constrained estimator $\hat{\beta}_{T_1} = \arg\min_{\beta\in\Lambda}\sum_{t=1}^{T_1}(y_{1t} - x_t'\beta)^2$ can be obtained as a projection of $\hat{\beta}_{\mathrm{OLS}}$ onto the convex set Λ, where $\Lambda = \Lambda_{\mathrm{SC}}$ or $\Lambda = \Lambda_{\mathrm{MSC}}$.

We first define some projections. For $\theta \in \mathcal{R}^N$, we define two versions of projection of θ onto a convex set Λ as follows:

$$\Pi_{\Lambda,T_1}\theta = \arg\min_{\lambda\in\Lambda}(\theta - \lambda)'(X'X/T_1)(\theta - \lambda), \tag{6}$$

$$\Pi_{\Lambda}\theta = \arg\min_{\lambda\in\Lambda}(\theta - \lambda)'E(x_t x_t')(\theta - \lambda). \tag{7}$$

Here, we use the notation Π_Λ to denote a projection onto the set Λ. The first projection Π_{Λ,T_1} is with respect to a random norm $\|a\|_X = \sqrt{a'(X'X/T_1)a}$ while the second projection Π_Λ is with respect to a nonrandom norm $\|a\|_E = \sqrt{a'E(x_t x_t')a}$, that is, $\Pi_{\Lambda,T_1}\theta = \arg\min_{\lambda\in\Lambda}||\lambda - \theta||_X^2$ and $\Pi_\Lambda\theta = \arg\min_{\lambda\in\Lambda}||\lambda - \theta||_E^2$. The first projection will be used to connect $\hat{\beta}_{T_1}$ and $\hat{\beta}_{\mathrm{OLS}}$, and the second projection relates the limiting distributions of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ and $\sqrt{T_1}(\hat{\beta}_{\mathrm{OLS}} - \beta_0)$.

With the above definition of the projection operator Π_{Λ,T_1}, we show in the Supplementary Appendix B that

$$\begin{aligned}\hat{\beta}_{T_1} &= \arg\min_{\beta\in\Lambda}(\hat{\beta}_{\mathrm{OLS}} - \beta)'(X'X/T_1)(\hat{\beta}_{\mathrm{OLS}} - \beta)\\ &= \Pi_{\Lambda,T_1}\hat{\beta}_{\mathrm{OLS}}.\end{aligned} \tag{8}$$

Equation (8) states that the constrained estimator is a projection of the unconstrained estimator onto the constrained set Λ. It is easy to check that when $X'X/T_1$ is a diagonal matrix, there is a simple closed form solution to the constrained minimization problem (8). Let $\hat{\beta}_{\mathrm{OLS}}$ denote the least squares estimator of β. Then it is easy to see that the closed form solution is $\hat{\beta}_{T_1,j} = \hat{\beta}_{\mathrm{OLS},j}$ if $\hat{\beta}_{\mathrm{OLS},j} \geq 0$ and $\hat{\beta}_{T_1,j} = 0$ if $\hat{\beta}_{\mathrm{OLS},j} < 0$ for $j = 2,\ldots,N$, that is, the projection simply keeps the positive component as it is and maps the negative component to zero. However, when $X'X/T_1$ is not a diagonal matrix, a simple noniterative closed form solution does not exist. Nevertheless, we show in the Supplementary Appendix B that when $X'X/T_1$ is positive definite, the objective function is globally convex and there is a unique solution to the constrained minimization problem.

3.2. The Asymptotic Theory: The Stationary Data Case

The asymptotic distribution of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ (hence, for $\hat{\Delta}_1$) can be characterized as a projection of the limiting distribution of an unconstrained estimator into a convex set, which is the so-called tangent cone of Λ evaluated at β_0. We use T_{Λ,β_0} to denote the tangent cone. The formal definition of the tangent cone T_{Λ,β_0} is given in Supplementary Appendix A.

To study the asymptotic distribution of $\hat{\Delta}_1$, we first need to know the asymptotic distribution of the (modified) SC estimator $\hat{\beta}_{T_1}$. The next theorem gives the needed result. But we will first list some regularity conditions that will be used in proving the main results of this article.

Assumption 1. The data $\{x_t\}_{t=1}^{T}$ is a weakly dependent stationary process so that laws of large numbers hold: $T_1^{-1}\sum_{t=1}^{T_1} x_t \xrightarrow{p} E(x_t)$ and $(X'X/T_1) \equiv T_1^{-1}\sum_{t=1}^{T_1} x_t x_t' \xrightarrow{p} E(x_t x_t')$. $E(x_t x_t')$ is positive definite, where X is the $T_1 \times N$ matrix with its tth row given by $x_t' = (1, y_{2t}, \ldots, y_{Nt})$. Let $\phi = \lim_{T_1,T_2\to\infty}\sqrt{T_2/T_1}$. Then, ϕ is a finite nonnegative constant.

Assumption 2. $\{u_{1t}\}_{t=1}^{T}$ has zero mean and satisfies $T_1^{-1/2}\sum_{t=1}^{T_1} x_t u_{1t} \xrightarrow{d} N(0,\Sigma_1)$, where $\Sigma_1 = \lim_{T_1\to\infty} T_1^{-1}\sum_{t=1}^{T_1}\sum_{s=1}^{T_1} E(u_{1t}u_{1s}x_t x_s')$.

Assumption 3. Let $v_{1t} = \Delta_{1t} - \Delta_1 + u_{1t}$. We assume that v_{1t} has zero mean and satisfies a central limit theorem: $T_2^{-1/2}\sum_{t=T_1+1}^{T} v_{1t} \xrightarrow{d} N(0,\Sigma_v)$, where $\Sigma_v = \lim_{T_2\to\infty} T_2^{-1}\sum_{t=T_1+1}^{T}\sum_{s=T_1+1}^{T} E(v_{1t}v_{1s})$.

Assumption 4. Let $w_t = (y_{1t}, y_{2t}, \ldots, y_{Nt}, \Delta_{1t}d_t)$ for $t = 1,\ldots,T$, where $d_t = 0$ if $t \le T_1$ and $d_t = 1$ if $t \ge T_1+1$. Assume that $\{w_t\}_{t=1}^{T_1}$ and $\{w_t\}_{t=T_1+1}^{T}$ are both weakly dependent stationary processes. Define $\rho(\tau) = \max_{1\le t\le T}\max_{1\le i,j\le N+1} |\mathrm{cov}(w_{it}, w_{j,t+\tau})/\sqrt{\mathrm{var}(w_t)\mathrm{var}(w_{j,t+\tau})}|$. Then there exists some finite positive constants $C > 0$ and $0 < \lambda < 1$ such that $\rho(\tau) \le C\lambda^{\tau}$.

Assumptions 1 and 2 imply that $\sqrt{T_1}(\hat\beta_{\mathrm{OLS}} - \beta_0) \xrightarrow{d} N(0, A^{-1}\Sigma_1 A^{-1})$, where $A = E(x_t x_t')$ and Σ_1 is defined in Assumption 2. Assumption 3 requires that a central limit theorem applies to a partial sum of v_{1t}. Assumption 4 is also used in Li and Bell (2017). This assumption ensures that the estimator $\hat\beta_{T_1}$ obtained using the pretreatment data is asymptotically independent with a quantity that involves the post-treatment sample average of the de-meaned treatment effects and the idiosyncratic error.

Theorem 3.1. Let Z_1 denote the limiting (normal) distribution of $\sqrt{T_1}(\hat\beta_{\mathrm{OLS}} - \beta_0)$. Then under Assumptions 1–4, we have

$$\sqrt{T_1}(\hat\beta_{T_1} - \beta_0) \xrightarrow{d} \Pi_{T_{\Lambda,\beta_0}} Z_1. \tag{9}$$

Theorem 3.1 states that the limiting distribution of the constrained estimator can be represented as a projection of the limiting distribution of the unconstrained (least squares) estimator onto the tangent cone T_{Λ,β_0}. There is a simple interpretation of the above result. The range of Z_1 is $\mathcal{R}^N$. However, when the constraints are binding, the range of $\sqrt{T_1}(\hat\beta_{T_1} - \beta_0)$ is a convex subset of $\mathcal{R}^N$. It can be shown that the asymptotic (as $T_1 \to \infty$) range of $\sqrt{T_1}(\hat\beta_{T_1} - \beta_0)$ is exactly T_{Λ,β_0}, the tangent cone of Λ at β_0. Therefore, the asymptotic distribution of $\sqrt{T_1}(\hat\beta_{T_1} - \beta_0)$ is the projection of the limiting distribution of $\sqrt{T_1}(\hat\beta_{\mathrm{OLS}} - \beta_0)$ onto the asymptotic range of $\sqrt{T_1}(\hat\beta_{T_1} - \beta_0)$.

With the help of Theorem 3.1, we derive the asymptotic distribution of $\sqrt{T_2}(\hat\Delta_1 - \Delta_1)$ as follows.

Theorem 3.2. Under the same conditions as in Theorem 3.1, we have

$$\sqrt{T_2}(\hat\Delta_1 - \Delta_1) \xrightarrow{d} -\phi E(x_t')\Pi_{T_{\Lambda,\beta_0}} Z_1 + Z_2, \tag{10}$$

where $\hat\Delta_1 = \hat\Delta_{1,\mathrm{SC}}$ or $\hat\Delta_{1,\mathrm{MSC}}$, $\Delta_1 = E(\Delta_{1t})$, $\phi = \lim_{T_1,T_2\to\infty}\sqrt{T_2/T_1}$, Z_1 is defined in Theorem 3.1, Z_2 is independent with Z_1 and distributed as $N(0,\Sigma_v)$, $\Sigma_v = \lim_{T_2\to\infty} T_2^{-1}\sum_{t=T_1+1}^{T}\sum_{s=T_1+1}^{T} E(v_{1t}v_{1s})$, $v_{1t} = \Delta_{1t} - E(\Delta_{1t}) + u_{1t}$, and u_{1t} has zero mean and is defined in (1).

The proof of Theorem 3.2 is given in Supplementary Appendix A.

Although we can use projection theory to characterize the asymptotic distribution of $\sqrt{T_1}(\hat\Delta_1 - \Delta_1)$, the inference is not straightforward as we have to know β_0 to calculate the tangent cone T_{Λ,β_0}. We show in Section 4 that a carefully designed subsampling method can be used to conduct valid inference. In particular, we do not need to know β_0 (or the tangent cone) when using the subsampling method for inference.

3.3. The Trend-Stationary Data Case

Up until now, we have only considered the stationary data case. However, many datasets, especially panel data with a long time dimension, exhibit some trending behaviors. For example, new product sales may increase over time due to word of mouth. In this subsection, we extend the stationary data result to the trend-stationary data case.

We add a time trend regressor to the regression model and obtain

$$y_{1t} = \alpha_0 t + z_t'\beta_0 + u_{1t}, \quad t = 1,\ldots,T_1, \tag{11}$$

where $z_t = (1, \eta_{2t}, \ldots, \eta_{Nt})'$, and η_{jt} is the de-trended data from y_{jt} for $j = 2,\ldots,N$. Let $\hat\alpha_{T_1}$ and $\hat\beta_{T_1}$ be the constrained least squares estimators of α_0 and β_0 subject to $\beta_j \ge 0$ for $j = 2,\ldots,N$ and $\sum_{j=2}^{N}\beta_j = 1$ for the SC estimator (or without the sum to one restriction for the MSC estimator) using the pretreatment data. We estimate y_{1t}^0 by $\hat y_{1t}^0 = \hat\alpha_{T_1} t + z_t'\hat\beta_{T_1}$ and the ATE is estimated by

$$\hat\Delta_1 = \frac{1}{T_2}\sum_{t=T_1+1}^{T}\hat\Delta_{1t}, \tag{12}$$

where $\hat\Delta_{1t} = y_{1t} - \hat y_{1t}^0$. We still maintain the assumption that Δ_{1t} is a stationary process so that the focal point is still $E(\Delta_1)$. To derive the asymptotic distribution of $\sqrt{T_2}(\hat\Delta_1 - \Delta_1)$, we need first present the theory for the unconstrained least squares estimator of $\gamma_0 = (\alpha_0, \beta_0')'$. Let $\hat\gamma_{\mathrm{OLS}}$ denote the ordinary least squares estimator of γ_0. Define $M_{T_1} = \sqrt{T_1}\mathrm{diag}(T_1, 1, \ldots, 1)$, which is an $(N+1)\times(N+1)$ diagonal matrix with its first diagonal element equals to $T_1^{3/2}$ and all other diagonal elements equal to $\sqrt{T_1}$. Then, it is well established that (e.g., Hamilton 1994)

$$M_{T_1}(\hat\gamma_{\mathrm{OLS}} - \gamma_0) \xrightarrow{d} N(0,\Omega), \tag{13}$$

where Ω is a $(N+1)\times(N+1)$ positive definite matrix the explicit definition of which is presented in Chapter 16 of Hamilton (1994).

We still use Λ to denote constrained sets for $\hat\gamma_{T_1}$ for trend-stationary data case. Now γ is an $(N+1)\times 1$ vector whose first component is the time trend coefficient and whose second component is the intercept. Hence, the constrained sets for the standard and the MSC models are $\Lambda_{\mathrm{SC}} = \{\gamma \in \mathcal{R}^{N+1} : \gamma_j \ge 0 \text{ for } j = 3,\ldots,N+1,\ \sum_{j=3}^{N+1}\gamma_j = 1\}$; and $\Lambda_{\mathrm{MSC}} = \{\gamma \in \mathcal{R}^{N+1} : \gamma_j \ge 0 \text{ for } j = 3,\ldots,N+1\}$. Define the SC estimator

$$\hat\gamma_{T_1} = \arg\min_{\gamma\in\Lambda}\sum_{t=1}^{T_1}(y_{1t} - w_t'\gamma)^2, \tag{14}$$

where $w_t = (t, z_t')'$ and $\Lambda = \Lambda_{\mathrm{SC}}$ or Λ_{MSC}. Then similar to Theorem 3.1, we have the following result.

Theorem 3.3. Let Z_3 denote the limiting distribution of $\sqrt{T_1}(\hat{\gamma}_{\text{OLS}} - \gamma_0)$ as described in (13). Then under the assumptions D1–D3 presented in Supplementary Appendix D, we have

$$\sqrt{T_1}(\hat{\gamma}_{T_1} - \gamma_0) \xrightarrow{d} \Pi^{tr}_{T_{\Lambda,\gamma_0}} Z_3,$$

where $\Pi^{tr}_{\Lambda}\theta$ denotes the projection of $\theta \in \mathcal{R}^{N+1}$ onto the convex set Λ defined in (D.11) of Supplementary Appendix D, T_{Λ,γ_0} is the tangent cone of Λ ($\Lambda = \Lambda_{\text{SC}}$ or Λ_{MSC}) evaluated at γ_0, and Z_3 is the weak limit of $M_{T_1}(\hat{\gamma}_{\text{OLS}} - \gamma_0)$ as described in (13).

With Theorem 3.3, we can derive the asymptotic distribution of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$.

Theorem 3.4. Under the same conditions as in Theorem 3.3, we have

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) \xrightarrow{d} -c'\Pi^{tr}_{T_{\Lambda,\gamma_0}} Z_3 + Z_2,$$

where $\hat{\Delta}_1$ is defined in (12), $\Delta_1 = E(\Delta_{1t})$, $c = (\sqrt{\phi}(2+\phi), \sqrt{\phi}E(z_t'))'$, $\phi = \lim_{T_1,T_2\to\infty}\sqrt{T_2/T_1}$, Z_3 is the limiting distribution of $M_{T_1}(\hat{\gamma}_{\text{OLS}} - \gamma_0)$ as defined in (13), and Z_2 is independent of Z_3 and normally distributed with zero mean and variance Σ_v.

The proof of Theorem 3.4 is given in Supplementary Appendix D.

3.4. The Unit-Root Nonstationary Data Case

In this section, we consider the case where the outcome variables follow drift-less unit root processes. We assume that, in the absence of treatment, the outcome variables are generated via

$$y^0_{it} = y^0_{i,t-1} + \eta_{it}, \quad t = 1,\ldots,T;\ i = 1,\ldots,N, \tag{15}$$

where η_{it} is a zero mean weakly dependent stationary process. For example, if y^0_{it} is generated by some common factors that are unit-root processes, then y^0_{it}, for $i = 1,\ldots,N$, also follows a unit-root process. The treated and control units form a co-integration relationship, that is,

$$y_{1t} = x_t'\beta_0 + u_{1t}, \quad t = 1,\ldots,T_1, \tag{16}$$

where $x_t = (1, y_{2t},\ldots,y_{Nt})'$, β is an $N \times 1$ vector of constant coefficients, and u_{1t} is a zero mean stationary process. The unknown vector of parameters, β_0, is a co-integration vector. As in the stationary data case, β_0 does not have an universal constant vector interpretation. In practice, we can use different control units to consistently estimate ATE. If N is large, we may impose restrictions such as $\beta_{0,j} \geq 0$ for $j = 2,\ldots,N$ to reduce the number of control units used in estimating the above co-integration model. We will continue to consider the MSC, which imposes nonnegativity slope coefficients restriction and the SC method, which imposes both the nonnegativity slope coefficients restriction and the coefficients sum-to-one constraint. The estimation method is the same as with the stationary data case. We estimate β_0 by the constrained least squares method based on (16). Let $\hat{\beta}_{T_1}$ denote the resulting constrained estimator of β_0. The counterfactual outcome is estimated via $\hat{y}^0_{1t} = x_t'\hat{\beta}_{T_1}$ for $t = T_1+1,\ldots,T$. The post-treatment period outcome is generated by $y_{1t} = y^0_{1t} + \Delta_{1t} = x_t'\beta_0 + u_{1t} + \Delta_{1t}$ for $t = T_1+1,\ldots,T$. Hence, ATE is estimated by

$$\hat{\Delta}_1 = \frac{1}{T_2}\sum_{t=T_1+1}^{T}\left[y_{1t} - \hat{y}^0_{1t}\right] = -\frac{1}{T_2}\sum_{t=T_1+1}^{T} x_t'(\hat{\beta} - \beta) + \bar{\Delta}_1 + \frac{1}{T_2}\sum_{t=T_1+1}^{T} u_{1t}, \tag{17}$$

where $\bar{\Delta}_1 = T_2^{-1}\sum_{t=T_1+1}^{T}\Delta_{1t}$.

Rewriting (17) leads to

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) = -\left(\frac{1}{T_1\sqrt{T_2}}\sum_{t=T_1+1}^{T} x_t'\right)[T_1(\hat{\beta} - \beta_0)] + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t}, \tag{18}$$

where $v_{1t} = u_{1t} + \Delta_{1t} - E(\Delta_{1t})$.

Since $\hat{\beta}_{T_1}$ is a projection of $\hat{\beta}_{\text{OLS}}$ onto the tangent cone, we need the asymptotic distribution theory for $T_1(\beta_{\text{OLS}} - \beta_0)$. Define an $N \times N$ diagonal matrix $D_{T_1} = T_1\,\text{diag}(T_1^{-1/2}, 1, \ldots, 1)$. Then the limiting distribution of the least squares estimator of the co-integration vector is well established:

$$D_{T_1}(\beta_{\text{OLS}} - \beta_0) \xrightarrow{d} Z_4,$$

where the limiting distribution of Z_4 is characterized by integration of some Brownian motions and its specific definition is presented in Supplementary Appendix D. In Supplementary Appendix D, we further show that $-(1/\sqrt{T_2})\sum_{t=T_1+1}^{T} x_t'D_{T_1}^{-1} \xrightarrow{d} Z_5$, where Z_5 is also an integration of a Brownian motion whose definition is given in Supplementary Appendix D. Hence, we obtain the limiting distribution for $\sqrt{T_2}(\hat{\Delta}_1 - \bar{\Delta}_1)$ as follows.

Theorem 3.5. Under the assumptions D4 and D5 given in Supplementary Appendix D, we have

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) \xrightarrow{d} Z_5\Pi^{I}_{T_{\Lambda,\beta_0}} Z_4 + Z_2, \tag{19}$$

where $\Pi^{I}_{\Lambda}\theta$ denotes the projection of $\theta \in \mathcal{R}^N$ onto the convex set Λ ($\Lambda = \Lambda_{\text{SC}}$ or Λ_{MSC}) defined in (D.18) of Supplementary Appendix D, Z_2 is distributed as $N(0,\Sigma_v)$, $\Sigma_v = \lim_{T_2\to\infty} T_2^{-1}\sum_{t=T_2+1}^{T}\sum_{s=T_1+1}^{T} E(v_{1t}v_{1s})$.

3.5. The Case of Large N

Up until now, we have focused on the case where N is fixed and T_1 and T_2 are large. In practice, there may exist a large number of control units, that is, N may not be small relative to T_1 and T_2. Since the control units' outcomes serve as regressors in our projection equation (4), it is well known that a large number of regressors lead to large estimation variance. Therefore, in practice when N is large, using all the control units may be suboptimal to using a subset of the control units in estimating the counterfactual outcome for the treated unit. In fact, the nonnegativity

coefficients and coefficients sum-to-one constraints suggested in the SC approach is partially motivated by this reason. Imposing these restrictions helps reduce the number of regressors and regularize the SC estimator. There is a large statistical literature on estimating high-dimensional models, even for models where the number of regressors is much larger than the sample size. In such cases, we have to impose some sparsity assumptions to obtain consistent estimation result. In our analysis, if the control units are drawn from a common distribution, sparsity may not be a reasonable assumption. Portnoy (1985) considered a regression model with iid data and allowed for the number of regressors to increase with sample size. In our notation, Portnoy (1985) assumed that $N \to \infty$ and $N^{3/2}\log(N)/T_1 \to 0$ as $T_1 \to \infty$ and established the asymptotic normal distribution result that $a'_{T_1}(\hat{\beta}_{\text{OLS}} - \beta_0)/b_{T_1} \xrightarrow{d} N(0,1)$, where a_{T_1} is an $N \times 1$ vector that satisfies $||a_{T_1}||$ is finite, and b_{T_1} is a scalar normalization factor.

Portnoy's (1985) result cannot be applied to our setting directly because we consider a projection model and unlike Portnoy (1985), we do not assume that $E(u_{1t}|x_t) = 0$. However, we conjecture that Portnoy's (1985) result may be extended to the weakly dependent stationary data case and also cover the case of a projection model. Then under the assumptions that $N \to \infty$ and $(N^{3/2}\log N)/T_1 \to 0$ as $T_1, T_2 \to \infty$, we conjecture that the results of this article (such as for the stationary data case) may remain valid for the large N case. In Sections 5.4 and 5.5 as well as in Supplementary Appendix E, we use simulations to examine cases where N is larger than T_2 and comparable to T_1. The simulation results show that in these large N scenarios, the MSC method performs better than the least squares method suggested by HCW (2012).

4. Inference Theory

We discuss inference methods for the ATE estimator $\hat{\Delta}_1$. For ease of exposition, we will focus on inference for the stationary data case. For trend-stationary data, we can first de-trend the data and then use the inference method discussed in this section for the de-trended data. The inference steps for unit-root stationary data are the same as the stationary data case.

As discussed in Section 3, the inference theory for the SC estimator is complicated. The asymptotic distribution of $\hat{\beta}_{T_1}$ depends on whether $\beta_{0,j} = 0$ or $\beta_{0,j} > 0$ for $j = 2, \ldots, N$. When $\beta_{0,j} > 0$ for all $j = 2, \ldots, N$, asymptotically, the constraints are nonbinding and the asymptotic theory of the constrained estimator is the same as that of the unconstrained ordinary least squares estimator. However, when the constraints are binding for some $j \in \{2, \ldots, N\}$, the asymptotic distribution of the constrained estimator is much more complex (e.g., Equation (9)). The asymptotic distribution of the SC coefficient estimators depends on whether or not the true parameters take values at the boundary. In practice, we do not know which constraints are binding and which are not, making it more difficult to use the asymptotic theory for inference. Moreover, when parameters fall on the boundary of the parameter space, the standard bootstrap method does not work (Andrews 2000; Fang and Santos 2018). We resolve this difficulty by proposing an easy-to-implement subsampling method. The proposed method works whether constraints are binding, partially binding, or nonbinding. That is, the subsampling method is adaptive in the sense that we do not need to know whether constraints are binding and if they are binding, we do not need to know on which coefficients they are binding. Hong and Li (2018) showed that numerical differentiation bootstrap method can consistently estimate the limiting distribution in many cases where the conventional bootstrap is known to fail. We can also use Hong and Li (2018) method to conduct inference for the SC estimator. In this article, we focus on the simple subsampling method.

We use m to denote the subsample size. We show that $\hat{\Delta}_1$ can be decomposed into two terms: the first term is related to the constrained estimator, $\hat{\beta}_{T_1}$, and the second term is unrelated to $\hat{\beta}_{T_1}$ but depends on T_2. Brute-force application of the subsampling method will not work in general. The correct procedure is to apply the subsampling method only to the $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ term and apply the bootstrap method to the remaining term that is unrelated to the constrained estimator $\hat{\beta}_{T_1}$.

For the whole sample period, the outcome y_{1t} is generated by

$$y_{1t} = x_t'\beta_0 + d_t\Delta_{1t} + u_{1t}, \quad t = 1, \ldots, T_1, T_1+1, \ldots, T, \tag{20}$$

where d_t is the post-treatment time period dummy so that $d_t = 0$ if $t \le T_1$ and $d_t = 1$ if $t \ge T_1 + 1$.

Substituting (20) into the left-hand-side of (10), we obtain

$$\begin{aligned}\hat{A} &\overset{\text{def}}{=} \sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)\\ &= -\sqrt{\frac{T_2}{T_1}}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} x_t'\right]\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t}\\ &= \hat{A}_1 + \hat{A}_2,\end{aligned} \tag{21}$$

where $v_{1t} = \Delta_{1t} - \Delta_1 + u_{1t}$, $\hat{A}_1 = -\sqrt{\frac{T_2}{T_1}}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} x_t'\right]\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$, and $\hat{A}_2 = \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t}$.

Now we impose an additional assumption that u_{1t} and v_{1t} are both serially uncorrelated, which greatly simplifies the subsampling method that will be discussed below. This assumption can be easily tested in practice. When this assumption is violated, we conjecture that more sophisticated methods such as block subsampling method may be used to deliver valid inferences.

Expression (21) suggests that we only need to apply the subsampling method to $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ because only this term is related to the constrained estimator. We now describe the subsampling steps. In Supplementary Appendix A, we show that when v_{1t} is serially uncorrelated, we can consistently estimate Σ_v by $\hat{\Sigma}_v = T_2^{-1}\sum_{t=T_1+1}^{T}\hat{v}_{1t}^2$, where $\hat{v}_{1t} = \hat{\Delta}_{1t} - \hat{\Delta}_1$. We generate $v_{1t}^* \sim$ iid $N(0, \hat{\Sigma}_v)$ for $t = T_1 + 1, \ldots, T$. Next, let m be the subsample size that satisfies the conditions that $m \to \infty$ and $m/T_1 \to 0$ as $T_1 \to \infty$. For $t = 1, \ldots, m$, we randomly draw (y_{1t}^*, x_t^*) from $\{y_{1t}, x_t\}_{t=1}^{T_1}$ with replacement (subsampling). Then we use the subsample $\{y_{1t}^*, x_t^*\}_{t=1}^{m}$ to estimate β_0 by the constrained least squares method, that is, $\hat{\beta}_m^* = \arg\min_{\beta\in\Lambda}\sum_{t=1}^{m}(y_{1t}^* - x_t^{*\prime}\beta)^2$. The subsampling-bootstrap ver-

sion of the statistic $\hat{A}$ is given by

$$\hat{A}^* = -\sqrt{\frac{T_2}{T_1}}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} x_t'\right]\sqrt{m}(\hat{\beta}_m^* - \hat{\beta}_{T_1}) + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t}^*. \quad (22)$$

We repeat the above process for a large number of times (J times). Using $\{\hat{A}_j^*\}_{j=1}^J$, we can obtain confidence intervals for $\hat{A}$. Specifically, we sort the subsampling-bootstrap statistics such that $\hat{A}_{(1)}^* \le \hat{A}_{(2)}^* \le \cdots \le \hat{A}_{(J)}^*$. Then the $1-\alpha$ confidence interval for Δ_1 is given by

$$[\hat{\Delta}_1 - T_2^{-1/2}\hat{A}_{((1-\alpha/2)J)}^*,\ \hat{\Delta}_1 - T_2^{-1/2}\hat{A}_{((\alpha/2)J)}^*]. \quad (23)$$

We show that the above method indeed gives consistent estimation of the confidence intervals for Δ_1 in the next theorem.

Theorem 4.1. Under the same conditions as in Theorem 3.2 and the assumptions that u_{1t} and v_{1t} are serially uncorrelated and that $m \to \infty$ and $m/T_1 \to 0$ as $T_1 \to \infty$, the $(1-\alpha)$ confidence interval of Δ_1 can be consistently estimated by $[\hat{\Delta}_1 - T_2^{-1/2}\hat{A}_{((1-\alpha/2)J)}^*,\ \hat{\Delta}_1 - T_2^{-1/2}\hat{A}_{((\alpha/2)J)}^*]$ for any $\alpha \in (0,1)$.

Remark 4.1. We apply the subsampling method only to part of the statistic, $\hat{A}_1$, because $\hat{A}_1$ depends on the constrained estimator $\hat{\beta}_{T_1}$ and it is known that bootstrap methods do not work when the true parameters are at the boundary of the parameter space. We apply a bootstrap method rather than a subsampling method to the other term, $\hat{A}_2$. This is important because it is difficult to apply a subsampling method to $\hat{A}_2$ (which depends on T_2) as T_2 is usually smaller than T_1. Subsampling methods applied to $\hat{A}_2$ with subsample sizes much smaller than T_2 usually do not work well in practice.

Remark 4.2. Even though we randomly draw (y_t^*, x_t^*) from $\{y_s, x_s\}_{s=1}^{T_1}$ for $t = 1, \ldots, m$, we do not require that $\{y_s, x_s\}_{s=1}^{T_1}$ be a serially uncorrelated process. In fact, they can have arbitrary serial correlation, for example, $\{y_{jt}\}_{j=1}^N$ is generated by some serially correlated common factors. We only need that the idiosyncratic error u_{1t} in (1) is serially uncorrelated. This can be easily tested given data. In Section 5, we generate y_{jt} using a three factor model, where the three factors follow AR, ARMA, and MA processes, respectively. Simulations show that the above proposed subsampling method works well. When u_{1t} is serially correlated, we conjecture that random subsampling method can be replaced by a block subsampling method. We leave the formal justification of using block subsampling method as a future research topic.

Remark 4.3. In the subsampling literature, the choice of subsample size m is a key issue. Bickel and Sakov (2008) propose a data-driven method to select m. In general, a value of m that is too small or too large does not work well. When m falls into an appropriate interval, the performance should be stable and acceptable. For our model, because $\beta_{0,j} > 0$ for some $j \in \{2, \ldots, N\}$, and the statistic

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) = -\sqrt{\frac{T_2}{T_1}}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} x_t'\right]\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t} \quad (24)$$

contains a term $\hat{A}_2 = T_2^{-1}\sum_{t=T_1+1}^{T} v_{1t}$, which is not related to $\hat{\beta}_{T_1}$, the subsampling method works reasonably well for a wider range of m. Even for $m = T_1$ (the bootstrap method), size distortions are quite mild indicating that although the bootstrap method does not lead to valid inference theoretically, it may still have practical value in applications. We provide evidence supporting this argument in Section 5.3 and in Supplementary Appendix F.

The subsampling method is a powerful tool for inference. It works under quite general conditions even when the regular bootstrap method does not work as in the case of the SC ATE estimator. Politis, Romano, and Wolf (1999) provide proofs and arguments showing that subsampling method works under very weak regularity conditions. Although we only theoretically investigate the subsampling method without replacement, it is well known that choosing a subsample size m out of the original T_1 data with or without replacement are asymptotically equivalent under mild conditions including $m/T_1 \to 0$, $m \to \infty$ as $T_1 \to \infty$. See Bickel and Sakov (2008). One advantage of using the "with replacement" method is that, when the bootstrap method works, the subsampling method also works when $m = T_1$ whereas the "without replacement" method breaks down when $m = T_1$.

The result of Theorem 4.1 can be used to conduct hypothesis tests regarding Δ_1. For example, we can test the null hypothesis

$$H_0\colon \Delta_1 \le 0 \text{ against the alternative hypothesis } H_1\colon \Delta_1 > 0.$$

By Theorem 4.1, a level α test for this hypothesis can be constructed by comparing $\hat{A}|_{\Delta_1=0} = \sqrt{T_2}\hat{\Delta}_1$ to the critical value

$$\hat{c}_{1-\alpha} \stackrel{\text{def}}{=} \inf\left\{c:\ P\left(\hat{A}^* \le c \,|\, \{y_{1t}, x_t\}_{t=1}^{T_1}\right) \ge 1-\alpha\right\}, \quad (25)$$

where $\hat{A}^*$ is defined in (22).

In Lemma C.3 presented in Supplementary Appendix C, we show that Theorem 4.1 holds uniformly locally at β_{0,T_1} as β_{0,T_1} approaches the boundary of the set Λ (as $T_1 \to \infty$), that is, our analysis delivers inference procedures with reliable size control.

5. Simulation Results

We use similar T_1, T_2, and N that appear at our empirical data to examine the performance of subsampling method inferences through simulations.

5.1. A Three Factor Data Generating Process

We conduct simulation studies using the same data generating process as in Hsiao, Ching, and Wan (2012) and Du and Zhang

(2015). We consider the following three factor data generating process:

$$y_t^0 = a + Bf_t + u_t, \quad t = 1, \ldots, T, \tag{26}$$

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, $B = (b_1, b_2, \ldots, b_N)'$ is the $N \times 3$ loading matrix where b_j is a 3×1 loading vector for unit j, $f_t = (f_{1t}, f_{2t}, f_{3t})'$ is the 3×1 vector of common factors, $f_{1t} = 0.8f_{1t-1} + \epsilon_{1t}$, $f_{2t} = -0.6f_{1t-1} + \epsilon_{2t} + 0.8\epsilon_{2t-1}$, $f_{3t} = \epsilon_{3t} + 0.9\epsilon_{3t-1} + 0.4\epsilon_{3t-2}$, and ϵ_{jt} is iid $N(0, 1)$. We choose $(a_1, a_2, \ldots, a_N) = (1, 1, \ldots, 1)$. We use two different distributions for u_{it}: iid $N(0, 1)$ and iid uniform$[-\sqrt{3}, \sqrt{3}]$. The simulation results are virtually identical for these two settings. For conciseness, we only report results for when u_{it} is iid uniform$[-\sqrt{3}, \sqrt{3}]$.

We use a set up similar to our empirical data by setting $T_1 = 90$, $T_2 = 20$, $T = T_1 + T_2 = 110$, and $N = 11$ (with 10 control units). We consider the following two sets for factor loadings:

$$\begin{aligned}
\text{DGP1}: \ & b_1 = \mathbf{1}_{3\times 1}; \ \ b_j = \mathbf{1}_{3\times 1} \text{ for } j = 2, \ldots, 7; \ \text{ and} \\
& b_j = \mathbf{0}_{3\times 1} \text{ for } j = 8, \ldots, 11, \\
\text{DGP2}: \ & b_1 = 2(\mathbf{1}_{3\times 1}); \ \ b_j = \mathbf{1}_{3\times 1} \text{ for } j = 2, \ldots, 7; \ \text{ and} \\
& b_j = \mathbf{0}_{3\times 1} \text{ for } j = 8, \ldots, 11,
\end{aligned}$$

where $\mathbf{1}_{3\times 1}$ and $\mathbf{0}_{3\times 1}$ denote 3×1 vectors of ones and zeros, respectively.

For both DGP1 and DGP2, 6 out of 10 control units have nonzero loadings and the remaining 4 control units have zero loadings. For DGP1, all nonzero factor loadings are set to be ones so that the treated and the control units (with nonzero loadings) are drawn from a common distribution. In contrast, for DGP2, loadings for the treated unit all equal to 2, and the controls units' loadings (with nonzero loadings) are all equal to 1. Thus, the treated and control units are drawn from two heterogeneous distributions.

We generate the following treatment effects Δ_{1t}:

$$\Delta_{1t} = \alpha_0 \left[\frac{e^{z_t}}{1 + e^{z_t}} + 1 \right], \quad t = T_1 + 1, \ldots, T, \tag{27}$$

where $z_t = 0.5z_{t-1} + \eta_t$ and η_t is iid $N(0, 0.5^2)$. For post-treatment period, $y_{1t} = y_{1t}^1 = y_{1t}^0 + \Delta_{1t}$, where y_{1t}^0 are generated as described earlier and Δ_{1t} is generated by (27). When $\alpha_0 = 0$, there is no treatment effect and when $\alpha_0 > 0$, the treatment effect is positive.

5.2. The Estimated Coverage Probabilities: The Stationary Data Case

In this section, we focus on the accuracy of the estimated coverage probabilities. We consider four estimation methods: (i) the OSC method (which does not have an intercept); (ii) the SC method defined in (2) and (3) which adds an intercept to the OSC method; (iii) the MSC method; and (iv) the (least squares) HCW method. We choose $\alpha_0 = 0$ (no treatment effects) and $\alpha_0 = 1$ (positive treatment effects). The estimation results are virtually identical for these two cases. For brevity, we only report the case of $\alpha_0 = 0$ in this section.

Since we have $N = 11$ parameters in the regression model, we need to choose a subsample size $m > N$. We select $m = 20$, 40, 60, 80, and 90 and include the case where the subsample size m equals the full sample size, $m = T_1 = 90$, for the reason discussed in Remark 4.3. The number of simulations are 1000 for each setup, and 400 subsamples are generated within each simulation. For conciseness, we only report results for $m = 20$, 60, 90. Also, since regular bootstrap method works for the HCW method, we only report bootstrap ($m = T_1$) results for HCW method.

The top of Table 1 reports estimated coverage probabilities for DGP1. All four estimation methods result in reasonably good estimated coverage probabilities. Next, we examine the results for DGP2 where the treated and control units are drawn from heterogeneous distributions so that the "parallel trends" assumption is violated. The bottom of Table 1 shows that both the original SC and the SC (which adds an intercept to the original SC) suffer from undercoverage problems. Both the MSC and the HCW methods work well, resulting in coverage probabilities close to their nominal values.

5.3. The Estimated Coverage Probabilities: The Unit-Root Data Case

In this subsection, we consider the unit-root nonstationary data process. Specifically, we replace the first common factor by an unit-root process, $f_{1t} = f_{1,t-1} + \epsilon_{it}$, and keep the second and the third common factors as before. The new data-generating processes, DGP3 and DGP4, are defined as follows:

Table 1. Coverage probabilities (common distribution).

	DGP1 (common distribution)									
m	20	60	90	20	60	90	20	60	90	90
	Original SC			SC			Modified SC			HCW
50%	0.444	0.498	0.488	0.499	0.462	0.482	0.517	0.488	0.493	0.484
80%	0.752	0.770	0.738	0.767	0.762	0.778	0.785	0.786	0.790	0.780
90%	0.856	0.864	0.868	0.883	0.879	0.885	0.894	0.882	0.883	0.888
95%	0.922	0.914	0.934	0.940	0.940	0.936	0.942	0.940	0.938	0.952
	DGP2 (heterogeneous distributions)									
	Original SC			SC			Modified SC			HCW
50%	0.266	0.262	0.294	0.294	0.314	0.306	0.474	0.492	0.470	0.452
80%	0.502	0.480	0.518	0.526	0.522	0.540	0.776	0.770	0.738	0.778
90%	0.618	0.610	0.636	0.658	0.638	0.666	0.884	0.876	0.866	0.882
95%	0.712	0.692	0.718	0.752	0.720	0.754	0.936	0.930	0.926	0.924

Table 2. Coverage probabilities for the unit-root process.

	DGP3									
	Original SC			SC			Modified SC			HCW
m	20	60	90	20	60	90	20	60	90	90
50%	0.423	0.464	0.442	0.483	0.470	0.476	0.517	0.452	0.548	0.469
80%	0.722	0.750	0.730	0.752	0.750	0.776	0.738	0.830	0.758	0.751
90%	0.832	0.857	0.853	0.892	0.838	0.876	0.882	0.837	0.952	0.885
95%	0.897	0.930	0.917	0.954	0.916	0.938	0.934	0.905	0.984	0.938
	DGP4									
	Original SC			SC			Modified SC			HCW
50%	0.033	0.046	0.054	0.070	0.064	0.072	0.445	0.429	0.432	0.466
80%	0.074	0.086	0.088	0.132	0.140	0.127	0.728	0.750	0.737	0.746
90%	0.097	0.100	0.108	0.168	0.169	0.175	0.852	0.855	0.849	0.847
95%	0.117	0.116	0.128	0.216	0.193	0.203	0.911	0.921	0.908	0.917

DGP3 is identical to DGP1 except that f_{1t} is replaced by a unit-root process,

DGP4 is identical to DGP2 except that f_{1t} is replaced by a unit-root process.

The estimated coverage probabilities are given in Table 2. For DGP3, the treated and the control units (with positive loadings) are drawn from a common distribution. All four methods work well. However, since the treated and controls are drawn from heterogeneous distributions for DGP4, the OSC and the SC (with an intercept) significantly underestimate the coverage probabilities. The MSC and the HCW methods still work well as they do not require data to be drawn from a homogeneous distribution.

5.4. Estimation MSE for a Wide Range of N

Our asymptotic inference theory is developed under the assumptions that N is fixed, and T_1 and T_2 are large. As discussed in Section 3.5, we conjecture that the asymptotic results of the article might be extended to allow for the case where N increases with T_1 but at a rate slower than T_1 diverges to infinity. In this section, we use simulations to examine the performance of the MSC method for a wide range of values of N, including cases that N is greater than T_2 and is comparable to T_1. We again fixed $T_1 = 90$ and $T_2 = 20$ and vary N from $\{11, 21, 31, 51, 81\}$. We compute MSE of $\hat{\Delta}_1$ as follows:

$$\text{MSE}(\hat{\Delta}_1) = \frac{1}{M}\sum_{j=1}^{M}(\hat{\Delta}_{1,j} - \Delta_1)^2,$$

where $\hat{\Delta}_{1,j}$ is the estimated value of Δ_1 at the jth replication, and $M = 10{,}000$ is the number of Monte Carlo simulation replications.

We use the same three (stationary) factor data generating processes as discussed in Section 5.1, and we consider four different sets of factor loadings in $b_i' f_t = b_{1i} f_{1t} + b_{2i} f_{2t} + b_{3i} f_{3t}$, where b_i's, for $i = 1, \ldots, N$, are given by

$$\text{DGP5}: \; b_1 = \mathbf{1}_{3\times 1}; b_j = (2)\mathbf{1}_{3\times 1} \text{ for } j = 2, \ldots, \frac{N+1}{2};$$
$$b_j = (-0.5)\mathbf{1}_{3\times 1} \text{ for } j = \frac{N+3}{2}, , \ldots, N;$$

$$\text{DGP6}: \; b_1 = \mathbf{1}_{3\times 1}; \; b_j = (-2)\mathbf{1}_{3\times 1} \text{ for } j = 2, \ldots, \frac{N+1}{2};$$
$$b_j = (0.5)\mathbf{1}_{3\times 1} \text{ for } j = \frac{N+3}{2}, \ldots, N;$$

$$\text{DGP7}: \; b_1 = (\mathbf{1}_{3\times 1}); \; b_j = (0.5)\mathbf{1}_{3\times 1} \text{ for } j = 2, \ldots, \frac{N+1}{2};$$
$$b_j = (0.2)\mathbf{1}_{3\times 1} \text{ for } j = \frac{N+3}{2}, \ldots, N;$$

$$\text{DGP8}: \; b_1 = (0.2)\mathbf{1}_{3\times 1}; \; b_j = \mathbf{1}_{3\times 1} \text{ for } j = 2, \ldots, \frac{N+1}{2};$$
$$b_j = (0.5)\mathbf{1}_{3\times 1} \text{ for } j = \frac{N+3}{2}, \ldots, N.$$

We use the short-hand-notation OSC, SC, and MSC to denote the original synthetic control, the synthetic control method that adds an intercept to the OSC and the MSC methods, respectively. Table 3 reports estimation MSE for $\hat{\Delta}_1$. For DGP5, the treated and the control units' overall correlation is positive, and the coefficients sum-to-one constraint is a correct restriction. In this case, the OSC performs the best, followed by the synthetic control method with intercept (SC), the MSC and the HCW method.

For DGP6, the overall correlation between the treated and control units is negative. Moreover, the OSC and SC's constraint that the slope coefficients sum-to-one is incorrect, which leads to large estimation MSEs. When $N \leq 21$, HCW performs the best because it can explore the strong negative correlation to get accurate counterfactual estimates while the MSC drops control units that are (partially) negatively correlated and uses the (relatively weakly) positive correlated control units to estimate counterfactuals. Therefore, its MSEs are larger than HCW's MSE when N is not too large. However, when $N \geq 31$, the MSC's MSEs are smaller than HCW's. The difference become larger as N increases.

For DGP7, although all units are positively correlated with the common factors, the treated unit's correlation with the common factors is much stronger than the control units' correlation. In this case, the MSC has the smallest MSEs for all cases. Moreover, its MSE decreases monotonically as N increases. This suggests that when the correlation between the treated and control units are weak, the MSC method can take advantage of the existence of a large number of control units and drop those control units that are (partially) negatively correlated with

Table 3. MSE($\hat{\Delta}_1$).

N	11	21	31	51	81	11	21	31	51	81
	DGP5					DGP6				
OSC	0.0614	0.0592	0.0581	0.0591	0.0582	0.3897	0.3792	0.3782	0.3822	0.3746
SC	0.0749	0.0727	0.0715	0.0722	0.0716	0.4667	0.4514	0.4598	0.4583	0.4531
MSC	0.0714	0.0733	0.0761	0.0868	0.1041	0.1381	0.1017	0.0933	0.0902	0.1031
HCW	0.0748	0.0844	0.0974	0.1506	0.7059	0.0749	0.0832	0.0988	0.1448	0.7174
	DGP7					DGP8				
OSC	0.3793	0.3856	0.3808	0.3749	0.3852	0.1825	0.1779	0.1751	0.1718	0.1728
SC	0.4603	0.4629	0.4621	0.4519	0.4673	0.2202	0.2140	0.2123	0.2107	0.2078
MSC	0.1267	0.1030	0.0971	0.0931	0.0937	0.0667	0.0672	0.0668	0.0685	0.0693
HCW	0.1299	0.1068	0.1138	0.1672	0.7684	0.0730	0.0838	0.0962	0.1496	0.7334

the treated unit to exploit maximum positive correlation of the treated unit with the remaining control units. For the HCW method, its MSE also decreases as N goes up initially. However, when N goes up further, HCW's MSE increases drastically due to estimating a large number of parameters (control units) when computing counterfactual outcomes.

Finally, for DGP8, the treated unit is strongly correlated with common factors, but control units are weakly correlated with these factors. In this case, the MSC method performs the best for all N values considered. Moreover, its MSEs are stable for a wide range of N values. HCW's MSEs are higher than the MSC method and differences increase quickly as N grows.

In summary, when the treated unit and control units are drawn from heterogeneous distribution, MSC dominates the OSC and the SC method while HCW also dominates the OSC and the SC when N is not too large. For DGP6 with $N = 11$ and $N = 21$, the HCW method has smaller MSE than the MSC method. For all other cases the HCW method has larger MSE than the MSC method. Indeed HCW can have substantially larger estimation MSEs than the MSC when N is large.

To find out whether estimation MSEs are mainly from squared biases or variances, we decompose MSE into squared bias and variance using the identity: $\text{MSE}(\hat{\Delta}_1) = (\text{Bias}(\hat{\Delta}_1))^2 + \text{var}(\hat{\Delta}_1)$, where $\text{MSE}(\hat{\Delta}_1) = M^{-1}\sum_{j=1}^{M}(\hat{\Delta}_{1,j} - \Delta_1)^2$, $\text{Bias}(\hat{\Delta}_1) = \bar{\hat{\Delta}}_1 - \Delta_1$, $\text{var}(\hat{\Delta}_1) = M^{-1}\sum_{j=1}^{M}(\hat{\Delta}_{1,j} - \bar{\hat{\Delta}}_1)^2$, and $\bar{\hat{\Delta}}_1 = M^{-1}\sum_{j=1}^{M}\hat{\Delta}_{1,j}$. For brevity, we report the ratios of $\text{var}(\cdot)/\text{MSE}(\cdot)$ in Supplementary Appendix E, where the results show that ratios are greater than 0.96 for all cases considered. Thus, the variance is the dominating component of MSE and the squared bias is negligible compared to variance. This explains why the MSC method often has smaller MSE than the HCW method especially when N is large because large N (large number of parameters that need to be estimated) lead to large estimation variance for the HCW method. The MSC method removes negatively correlated control units which leads to smaller estimation variance than that of the HCW method. In Supplementary Appendix E, we also report MSE for estimating counterfactual outcome y_{1t}^0. The relative performances among different methods are the same as the case of $\text{MSE}(\hat{\Delta}_1)$.

5.5. Coverage Probabilities: The Large N Case

Next, we report coverage probabilities for the large N set up. We consider $N = 31, 51, 81$. We consider subsample sizes $m = 40, 60, 90$ for $N = 31$; $m = 70, 80, 90$ for $N = 51$; and $m = 90$ for $N = 81$. To be concise, we will only report the results for the MSC and the HCW methods. Also, we only report results for DGP5 and DGP6 and put results for DGP7 and DGP8 in Supplementary Appendix E. The estimated coverage probabilities are reported in Table 4. We notice that the MSC methods work reasonably well except for DGP6 and $N = 81$, where it overestimates the coverage probabilities. The HCW method does not work as well due to its large estimation variances (since it needs to estimate a large number of parameters) and overestimates the coverage probabilities.

In practice, what *finite* T_1, T_2, and N values can lead to reliable inference (for the MSC method) is a finite sample issue. Researchers can use simulations to examine this problem for their specific combination of T_1, T_2, and N. The simulation results reported in this section suggest that our inference theory works satisfactorily for $T_1 = 90$, $T_2 = 20$, and $N = 11$. If we allow for N to diverge to $+\infty$ along with $T_1 \to \infty$, we conjecture that the rate that N diverges has to constrained such that $N/T_1 = o(1)$ or even $N^{3/2}/T_1 = o(1)$ (e.g., Portnoy 1985). A rigorous theoretical investigation for allowing large N with stationary data is beyond the scope of this article.

6. Empirical Application

We illustrate the usefulness of the MSC method in practice with empirical data. We calculate the ATE based on the MSC method and the confidence intervals using the subsampling method.

6.1. Eyewear Brand and ATE Estimation

We have data from an iconic online first eyewear brand that provides of high quality eyewear at modest prices ($95 instead of $300+range). In February 2010, the company opened its first physical showroom in New York City. Later, they opened showrooms in several cities hoping that opening physical showrooms would significantly promote sales. They opened a showroom in Columbus, Ohio on November 10, 2011. In this section, we want to evaluate how the showroom opening in Columbus affected Columbus' average weekly sales (the ATE) in the post-treatment period. As discussed in Section 2 on estimating treatment effects using panel data, we estimate the counterfactual sales for Columbus by letting the sales of Columbus be the dependent variable and the control cities' sales (sales in cities without showrooms) be the explanatory variables. We run a con-

Table 4. Coverage probabilities for large N.

	DGP5									
	Modified SC control							HCW		
N	$N = 31$			$N = 51$			$N = 81$	$N = 31$	$N = 51$	$N = 81$
m	40	60	90	70	80	90	90	90	90	90
50%	0.534	0.507	0.491	0.502	0.528	0.567	0.532	0.567	0.867	0.781
80%	0.833	0.775	0.781	0.831	0.831	0.832	0.833	0.871	0.996	0.992
90%	0.910	0.874	0.900	0.928	0.904	0.927	0.923	0.946	1.00	1.00
95%	0.951	0.934	0.946	0.971	0.953	0.962	0.960	0.969	1.00	1.00
	DGP6									
	Modified SC control							HCW		
N	$N = 31$			$N = 51$			$N = 81$	$N = 31$	$N = 51$	$N = 81$
m	40	60	90	70	80	90	90	90	90	90
50%	0.531	0.521	0.517	0.552	0.479	0.532	0.578	0.648	0.846	0.785
80%	0.812	0.813	0.817	0.836	0.807	0.821	0.876	0.891	0.993	0.994
90%	0.904	0.905	0.918	0.917	0.906	0.907	0.945	0.956	1.00	1.00
95%	0.945	0.959	0.947	0.960	0.945	0.958	0.982	0.993	1.00	1.00

strained regression, that is, we regress weekly sales of Columbus on sales of control cities to obtain the estimated coefficients under the restriction that the coefficients are nonnegative. Then, using these estimated coefficients, together with the post-treatment period sales for the control group cities, we compute the counterfactual of what sales would be for Columbus in the absence of the showroom opening. The 10 largest cities (by population) that do not have showrooms were selected as the control group cities. These control cities are Atlanta, Chicago, Dallas, Denver, Houston, Minneapolis, Portland, San Diego, Seattle, and Washington.

We estimate ATE using three versions of the SC method: (i) the OSC method, where the slope coefficients are nonnegative and sum to one and does not include an intercept; (ii) the SC method which adds an intercept to the OSC method; and (iii) the MSC method which drops the restriction that the slope coefficients sum to one.

In Figure 1, the solid line is Columbus' actual sales, the dotted line is the in-sample-fit (for $t \leq T_1$) and the out-of-sample (for $t > T_1$) counterfactual forecast curve computed via the OSC method. Figure 1 shows that the OSC method fits the in-sample data poorly as the fitted curve is mostly above the actual sales. One reason for this is that there is no intercept for the OSC method. Adding an intercept to the model would solve the problem that almost all of fitted data lies above the actual data during the pretreatment period.

Figure 2 plots Columbus' actual sales (solid line) and the in-sample-fit and out-of-sample counterfactual forecast (dotted

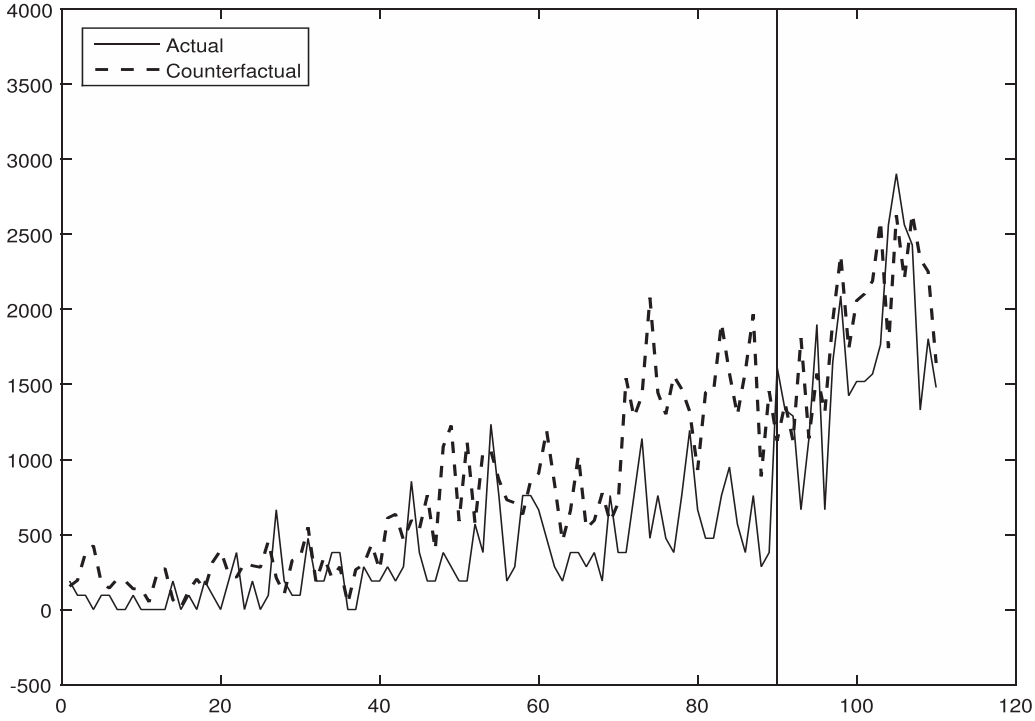

Figure 1. Columbus: The original synthetic control fitted curve.

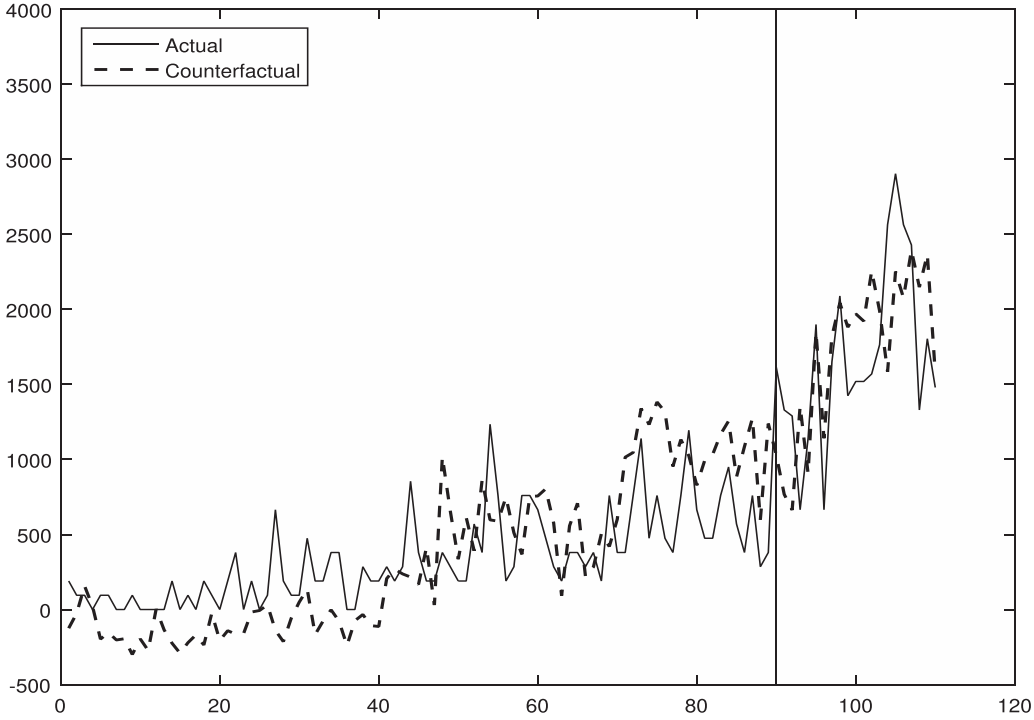

Figure 2. Columbus: Add an intercept to the synthetic control fitted curve.

line) curve computed by adding an intercept to the OSC method (i.e., using (3)). Introducing an intercept to the model allows the fitted curve and the actual data to have the same in-sample mean. Unlike the OSC method, the fitted data no longer lies mostly above the actual data. However, the in-sample fit is still poor as the fitted curve underestimates the actual sales for the first half of the in-sample data and overestimates the actual sales for the second half of the in-sample data. Applying the SC method to to Columbus' sales dataset overestimates the counterfactual outcome, which results in an underestimation of ATE. The reason for this is that without the restriction of coefficients sum-to-one, the sum of the slope coefficients is 0.234. The SC method imposes the restriction that the slope coefficients sum-to-one, which inflates the slope of fitted curve to be larger than the slope of the actual data. The estimated intercept moves the fitted curve down parallel in an attempt to make the fitted curve and the actual data have the same sample mean (for pretreatment period data). This leads the fitted curve to be below the actual data for the first half of the pretreatment time period and above the actual data for the second half of the pretreatment time period, that is, the "parallel line" assumption is violated. Hence, it leads to a significant overestimation of the out-of-sample counterfactual sales, which in turn leads to a severely downward biased estimated ATE.

The above analysis suggests that restricting the slope coefficients sum to one is the reason for a large estimation bias of the SC method. Therefore, we further relax the weights sum-to-one restriction, that is, only keep the nonnegativity of the weights but drop the sum-to-one constraint (and include an intercept to the model). The estimation results are plotted in Figure 3. We observe a greatly improved in-sample-fit. Unlike Figure 2, the fitted curve in Figure 3 does not appear to have any systematic estimation bias (for $1 \leq t \leq T_1$). Our estimation result shows that opening a showroom in Columbus on November 10, 2011 leads to an average 67% increase in weekly sales. In the next subsection, we show that the estimated positive ATE is highly statistically significant.

6.2. Confidence Intervals for the ATE

We use the subsampling method discussed in Section 4 to estimate confidence intervals (CI) for the ATE (Δ_1). Since our proposed subsampling method requires that the idiosyncratic error u_{1t} defined in (1) and v_{1t} defined in Theorem 3.3 are serially uncorrelated, we first test whether these assumptions hold. Our test statistics are based on the sample analogues of $\sqrt{T_1}\rho_u = \sqrt{T_1}E(u_{1t}u_{1,t-1})/E(u_{1t}^2)$ and $\sqrt{T_2}\rho_v = \sqrt{T_2}E(v_{1t}v_{1,t-1})/E(v_{1t}^2)$. The p-values of these tests are 0.467 and 0.0963, respectively. Therefore, we do not reject the null hypotheses that u_{1t} and v_{1t} are serially uncorrelated at the 5% significance level.

To conduct the subsampling inference, we choose subsample sizes $m = 20, 40, 60, 80$, and 90. For each value of m, we conduct 10,000 subsampling simulations to obtain $\{\hat{\Delta}_1 - T_2^{-1/2}\hat{A}_j^*\}_{j=1}^{10,000}$ (see Equation (23)). We then sort these 10,000 statistics to obtain $\alpha/2$ and $(1-\alpha/2)$ percentiles for $\alpha = 0.2, 0.1, 0.05$ and 0.01. The results are given in Table 5.

First, Table 5 shows that the estimated confidence intervals are similar for different subsample sizes including the case of $m = T_1$ (recall that $T_1 = 90$). The empirical data further verifies that due to the reason discussed in Remark 4.3 and further illustrated in Supplementary Appendix F, the subsampling method works well for a wide range of values of m. Next, we notice that the lower bound of these intervals are all positive and far above zero for all values of m. This implies that the estimated ATE value of 673.91 is positive and significantly different from zero for all conventional significant levels. In fact, if we conduct

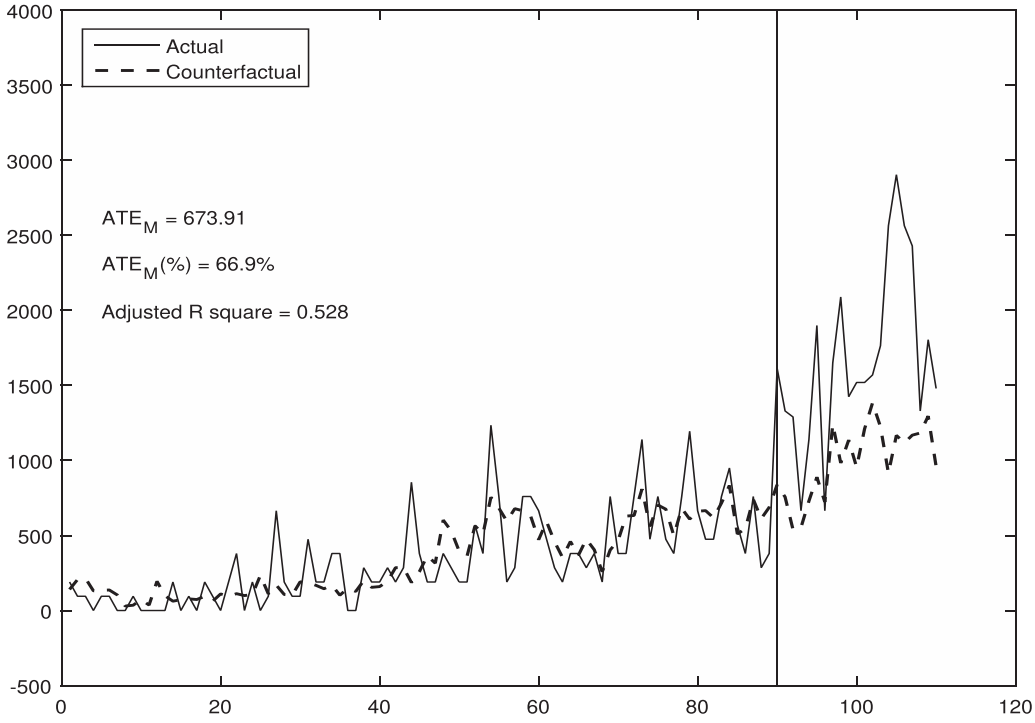

Figure 3. Columbus: Modified synthetic control ATE estimation.

a 5% level one-sided test, we reject $\Delta_1 = 430$ and in favor of $\Delta_1 > 430$ because $P(\Delta_1 \leq 430) < 0.05$ or equivalently, $P(\Delta_1 > 430) > 0.95$ for all values of m considered. Thus, opening a showroom in Columbus significantly increased the company's eyewear sales.

When testing the null that v_{1t} is serially uncorrelated, we obtain a p-value of .0963, which actually rejects the null of zero serial correlation at the 10% level. Thus, it suggests that there may be nonzero correlation in v_{1t}. Therefore, we compute the estimated confidence intervals for ATE while allowing for v_{1t} to follow an AR(1) process in Supplementary Appendix G.4. The results there show that the confidence intervals are similar to, but slightly wider than those reported in Table 5.

For comparison, we also compute confidence intervals for the HCW method. The HCW's ATE point estimate is 645.3, and the bootstrap (i.e., $m = T_1 = 90$) estimates of the confidence intervals are given in Table 6. We observe that the lengths of these intervals are similar to the MSC CIs except that they all move to the left because HCW's ATE estimate is smaller than the MSC estimate.

6.3. Robustness Checks

We conduct the following robustness checks:

1. Change the treatment date from $T_1 = 90$ to a pseudo treatment date $T_0 = T_1 - 10 = 80$.
2. Compare forecasting performance with the unconstrained (HCW) estimation method.
3. Add three covariates (monthly data linear interpolated to weekly data): unemployment rate, labor force, and average weekly earnings for all employees in private sector.
4. Select control units based on covariates matching.

For conciseness, we only report robustness check 1 here and report robustness checks 2 through 4 in Supplementary Appendix G.

6.3.1. Change the Treatment Date

The Columbus showroom was opened in week 90 ($T_1 = 90$). We change the treatment date to be 10 weeks earlier as if the showroom had been opened at $t = 80$. Using data from $t = 1$ to

Table 5. Confidence intervals (MSC, based on 10,000 simulations).

	$m = 20$	$m = 40$	$m = 60$	$m = 80$	$m = 90$
80% CI	[489.6, 880,1]	[487.4, 870.1]	[491.7, 876.5]	[487.8, 871.4]	[488.4, 876.5]
90% CI	[436.3, 941.9]	[431.5, 927.8]	[432.9, 926.4]	[437.5, 921.6]	[433.9, 929.9]
95% CI	[395.1, 996.0]	[389.6, 975.5]	[390.9, 978.4]	[392.2, 967.6]	[387.4, 977.6]
99% CI	[295.6, 1110.1]	[309.8, 1068.1]	[299.0, 1074.1]	[302.1, 1069.0]	[297.6,1079.5]

Table 6. Confidence intervals (HCW, based on 10,000 simulations).

Coverage probability	80%	90%	95%	99%
Confidence interval	[433.1, 821.6]	[383.0, 878.0]	[342.6, 930.7]	[264.6, 1035.6]

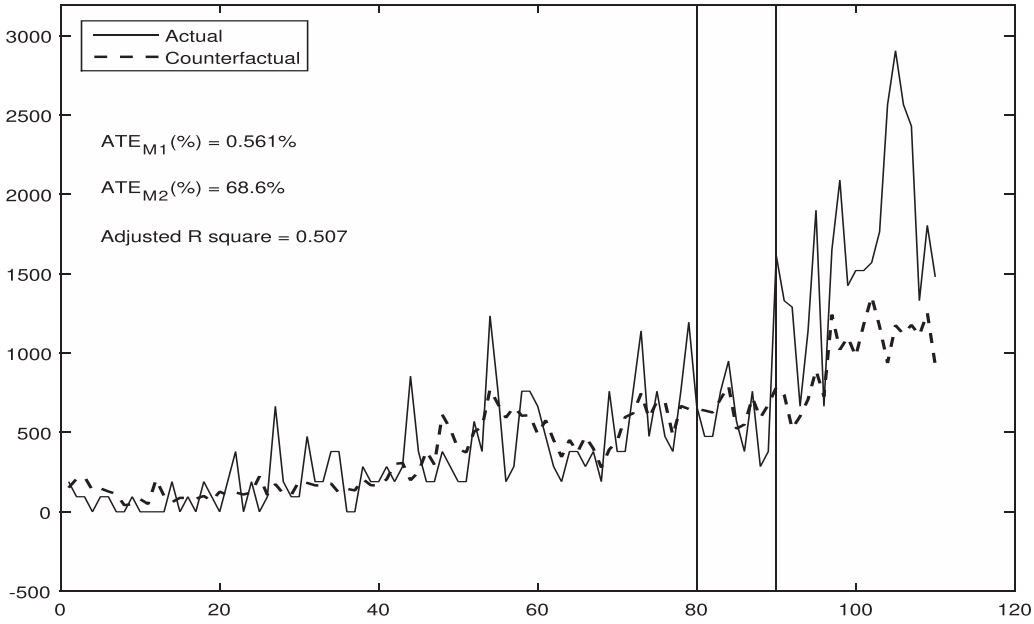

Figure 4. Columbus: Modified synthetic control ATE: different "T_1."

Table 7. Confidence intervals (MSC, based on 10,000 simulations).

	$m = 20$	$m = 40$	$m = 60$	$m = 80$	$m = 90$
80% CI	[−132.7, 196.1]	[−136.2, 176.6]	[−136.3, 169.9]	[−137.9, 165.5]	[−138.1, 162.7]
90% CI	[−176.4, 251.5]	[−176.0, 224.4]	[−178.8, 216.5]	[−178.1, 210.3]	[−181.3, 204.5]
95% CI	[−214.8, 308.1]	[−213.9, 267.1]	[−216.2, 255.0]	[−215.5, 251.5]	[−215.7, 242, 4]
99% CI	[−284.6, 454.2]	[−295.6, 354.1]	[−276.7, 340.9]	[−290.3, 333.7]	[−289.7, 318.3]

80, we estimate the model using the MSC method and predict Columbus' counterfactual sales from weeks 81 to 110. Since there was no showroom during $t = 81$ to 90, there should be no significant differences between y_{1t} and $\hat{y}^0_{1t}$ for $81 \leq t \leq 90$. From Figure 4, we see that for the periods 81 to 90, the predicted sales trace the actual sales closely. The ATE percentage increase for these 10 periods is 0.561% which is close to no effect as expected while the ATE for $t = 91$ to 110 is 68.6% which is very close to the original ATE estimate of 67%. We also compute the 80%, 90%, 95%, and 99% confidence intervals (CIs) for Δ_1 based on the estimated $\hat{\Delta}_1$ using data from $t = 81$ to 90 with 10,000 subsampling simulations. The results are given in Table 7. All confidence intervals contain zero. Therefore, we cannot reject the null hypothesis that there is no treatment effect during the period of $81 \leq t \leq 90$ at any conventional levels and this robustness check supports the MSC estimation result.

7. Conclusion

The SC method is a popular and powerful way of estimating ATE. This article provides the inference theory for the (modified) SC method under long panels with large pre and post-treatment periods. We derive the asymptotic distribution of the SC ATE estimator using projection theory. Because the asymptotic distribution is nonnormal and nonstandard, standard bootstrapping does not work. We resolve the difficulty by proposing a carefully designed and easy-to-implement subsampling method and establish the validity of subsampling method for inference. Our work complements the case of long pretreatment and short post-treatment data where end of sample instability tests are applied (Ferman and Pinto 2016, 2018; Andrews 2003) and permutation tests proposed in Chernozhukov, Wuthrich, and Zhu (2017) for conducting inference. Although theoretically we only consider the case of large T_1 and T_2 with a fixed N, simulations show that the MSC method works reasonably well even when N is larger than T_2 and comparable to T_1. Allowing N to grow with T_1 is left as a future research topic.

We also prove that when the pretreatment sample size is larger than the number of control units ($T_1 \geq N$), the SC estimator, as a constrained minimization problem, has a unique solution under a mild condition that the T_1 by N data matrix has a full column rank. In addition, we show the MSC method can give reliable ATE estimation results even when the "parallel trends" assumption is violated for the conventional SC method. Simulations show that the MSC method performs well in practice. Finally, we apply the SC and MSC method to estimate ATE of opening a showroom by an e-tailer. The empirical application demonstrates that when the conventional SC method fits the data poorly, the MSC method fits the data well and provides reasonable ATE estimation results.

Supplementary Materials

The supplementary materials contain seven appendices. Appendix A provides proofs of the main results in the article. The uniqueness of the SC and MSC estimators are established in Appendix B. Appendix C presents three lemmas: two are used in Appendix A and the third shows our test

is locally uniformly stable. Appendix D provides asymptotic analyses of nonstationary data cases. Additional simulation results are reported in Appendix E. Appendix F explains why subsampling methods work well for a wide range of subsample sizes. Finally, more robustness checks can be found in Appendix G.

Acknowledgments

This article is based on an essay of my PhD dissertation at The Wharton School of the University of Pennsylvania. I would like to thank my co-advisers David R. Bell and Christophe Van den Bulte for invaluable guidance and my committee members, Eric T. Bradlow and Dylan S. Small for helpful comments. Insightful comments from three referees and an associate editor substantially improved the article. The current version of the article also benefited tremendously from invaluable comments from the participants of the 2019 NBER Synthetic Control Conference hosted at MIT and the Marketing Department seminar participants at University of Pennsylvania, University of Michigan, Syracuse University, Oklahoma State University, University of Houston, University of Colorado Boulder, University of Virginia, University of Notre Dame, University of Chicago, Northwestern University, Texas A&M University, Dartmouth College, University of Texas at Austin, UCLA, Washington University St. Louis, UCSD, Southern Methodist University and Stanford University.

References

Abadie, A. (2005), "Semiparametric Difference-in-Differences Estimators," *The Review of Economic Studies*, 72, 1–19. [2069]

Abadie, A., Diamond, A., and Hainmueller, J. (2010), "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493–505. [2068,2069,2070,2071]

Abadie, A., and Gardeazabal, J. (2003), "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93, 113–132. [2068,2070]

Andrews, D. W. K. (2000), "Inconsistency of the Bootstrap When a Parameter Is on the Boundary of the Parameter Space," *Econometrica*, 68, 399–405. [2069,2074]

——— (2003), "End-of-Sample Instability Tests," *Econometrica*, 71, 1661–1694. [2069,2082]

Ashenfelter, O. (1978), "Estimating the Effect of Training Programs on Earnings," *The Review of Economics and Statistics*, 60, 47–57. [2068]

Ashenfelter, O., and Card, D. (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *The Review of Economics and Statistics*, 67, 648–60. [2068]

Athey, S., and Imbens, G.W. (2017), "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives*, 31, 3–32. [2068]

Bickel, P. J., and Sakov, A. (2008), "On the Choice of m in the m Out of n Bootstrap and Confidence Bounds for Extrema," *Statistica Sinica*, 18, 967–985. [2075]

Bronnenberg, B. J., Dubé, J.-P. H., and Gentzkow, M. (2012), "The Evolution of Brand Preferences: Evidence From Consumer Migration," *American Economic Review*, 102, 2472–2508. [2068]

Busse, M., Silva-Risso, J., and Zettelmeyer, F. (2006), "$1,000 Cash Back: The Pass-Through of Auto Manufacturer Promotions," *American Economic Review*, 96, 1253–1270. [2068]

Card, D., and Krueger, A.B. (1994), "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772–784. [2068]

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2017), "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls," arXiv no. 1712.09089. [2069,2082]

Conley, T. G., and Taber, C.R. (2011), "Inference With 'Difference in Differences' With a Small Number of Policy Changes," *The Review of Economics and Statistics*, 93, 113–125. [2069]

Doudchenko, N., and Imbens, G.W. (2016), "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis," Discussion Paper, National Bureau of Economic Research. [2069,2070,2071]

Du, Z., and Zhang, L. (2015), "Home-Purchase Restriction, Property Tax and Housing Price in China: A Counterfactual Analysis," *Journal of Econometrics*, 188, 558–568. [2076]

Fang, Z., and Santos, A. (2018), "Inference on Directionally Differentiable Functions," *The Review of Economic Studies*, 86, 377–412. [2069,2074]

Ferman, B., and Pinto, C. (2016), "Revisiting the Synthetic Control Estimator," Discussion Paper. [2069,2082]

——— (2018), "Inference in Differences-in-Differences With Few Treated Groups and Heteroskedasticity," *Review of Economics and Statistics*, 101, 452–467. [2069,2082]

Forman, C., Ghose, A., and Goldfarb, A. (2009), "Competition Between Local and Electronic Markets: How the Benefit of Buying Online Depends on Where You Live," *Management Science*, 55, 47–57. [2068]

Goldfarb, A., and Tucker, C.E. (2011), "Privacy Regulation and Online Advertising," *Management Science*, 57, 57–71. [2068]

Hahn, J., and Shi, R. (2017), "Synthetic Control and Inference," *Econometrics*, 5, 1–12. [2069]

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton, NJ: Princeton University Press. [2072]

Hong, H., and Li, J. (2018), "The Numerical Delta Method and Bootstrap," *Journal of Econometrics*, 206, 379–394. [2074]

Hsiao, C., Ching, S., and Wan, K.S. (2012), "A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Economic Integration of Hong Kong With Mainland China," *Journal of Applied Econometrics*, 27, 705–740. [2069,2071,2074,2075]

Imbens, G. W., and Wooldridge, J. M. (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86. [2068]

Li, K. T., and Bell, D. R. (2017), "Estimation of Average Treatment Effects With Panel Data: Asymptotic Theory and Implementation," *Journal of Econometrics*, 197, 65–75. [2069,2072]

Mantin, B., and Rubin, E. (2016), "Fare Prediction Websites and Transaction Prices: Empirical Evidence From the Airline Industry," *Marketing Science*, 35, 640–655. [2068]

Ozturk, O. C., Venkataraman, S., and Chintagunta, P. K. (2016), "Price Reactions to Rivals' Local Channel Exits," *Marketing Science*, 35, 588–604. [2068]

Pischke, J.-S. (2007), "The Impact of Length of the School Year on Student Performance and Earnings: Evidence From the German Short School Years," *The Economic Journal*, 117, 1216–1242. [2068]

Politis, D. N., Romano, J. P., and Wolf, M. (1999), *Subsampling*, Springer Series in Statistics, Berlin: Springer. [2075]

Portnoy, S. (1985), "Asymptotic Behavior of M-Estimators of p Regression Parameters When $p^2/2$ Is Large; II. Normal Approximation," *The Annals of Statistics*, 13, 1403–1417. [2074,2078]

Wang, K., and Goldfarb, A. (2017), "Can Offline Stores Drive Online Sales?," *Journal of Marketing Research*, 54, 706–719. [2068]

Zarantonello, E. H. (1971), "Projections on Convex Sets in Hilbert Space and Spectral Theory: Part I. Projections on Convex Sets: Part II. Spectral Theory," in *Contributions to Nonlinear Functional Analysis*, pp. 237–424. New York: Elsevier. [2069]