2018

Three Essays On The Estimation Of Average Treatment Effects In Quasi-Experimental Panel Data

Kathleen Tina Li
University of Pennsylvania, kathyli89@gmail.com

Three Essays On The Estimation Of Average Treatment Effects In Quasi-Experimental Panel Data

Abstract

Identifying average treatment effects (ATE) from quasi-experimental panel data has become one of the most important yet challenging endeavors for social scientists. The difficulty lies in accurately estimating the counterfactual outcomes for the potentially treated units in the absence of treatment. Perhaps the most popular method to estimate average treatment effects is the Difference-in-Differences (DID) method. The key assumption of the DID method is that outcomes of the treated units would have followed a path parallel to the control units in the absence of treatment and violation of this ``parallel lines" assumption will result in biased estimates. This dissertation consists of three essays, which either build on existing methods (essay 1 and 3) or propose a new method (essay 2) that can be used even when the ``parallel lines" assumption of DID does not hold. In essay 1, we derive the asymptotic distribution of the HCW method, which is computationally simple as it only involves least squares regressions. However, in cases where treatment and control units are positively correlated, the HCW method may have less predictive efficiency than other methods such as the synthetic control and modified synthetic control method, which impose the restriction that weights are non-negative. The popular synthetic control method additionally imposes the restriction that the weights sum to one, which can be a helpful regularization condition when there are many control units. In essay 3, we provide the inference theory for both the synthetic control and modified synthetic control method through projection theory and propose a computational algorithm using subsampling to compute the confidence intervals. In order to apply the HCW method, synthetic control method and modified synthetic control method, the number of control units needs to be smaller than the pre-treatment sample size. In essay 2, we propose the augmented DID method, which can be used where there are many treatment and control units, but is less flexible than the three aforementioned methods. In short, this dissertation provides several methods and their inference procedures to identify average treatment effects. Which method should be used when depends on the structure of the data.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Marketing

First Advisor

David R. Bell

Second Advisor

Christophe Van den Bulte

Keywords

average treatment effects, Difference-in-Differences, quasi-experiments, synthetic control methods

Subject Categories

Advertising and Promotion Management | Marketing | Statistics and Probability

THREE ESSAYS ON THE ESTIMATION OF AVERAGE TREATMENT EFFECTS IN QUASI-EXPERIMENTAL PANEL DATA

Kathleen T. Li

A DISSERTATION

in

Marketing

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

2018

Supervisor of Dissertation

_____

David R. Bell
Xinmei Zhang and Yonge Dai Professor
Professor of Marketing

Co-Supervisor of Dissertation

_____

Christophe Van den Bulte
Gayfryd Steinberg Professor
Professor of Marketing

Graduate Group Chairperson

_____

Catherine Schrand, Celia Z. Moh Professor, Professor of Accounting

Dissertation Committee:

Eric T. Bradlow, The K.P. Chao Professor, Professor of Marketing, Professor of Economics, Professor of Education, Professor of Statistics
Dylan S. Small, Class of 1965 Wharton Professor of Statistics, Professor of Statistics

THREE ESSAYS ON THE ESTIMATION OF AVERAGE TREATMENT EFFECTS IN QUASI-EXPERIMENTAL PANEL DATA

ACKNOWLEDGEMENT

My deepest gratitude goes to my dedicated advisers David R. Bell and Christophe Van den Bulte for their guidance, patience, support and mentorship. They have gone above and beyond and helped me to see the bigger picture as a researcher. I am also greatly indebted to my committee members, Eric T. Bradlow and Dylan S. Small, for their invaluable insights, kindness, and encouragement.

I would also like the thank the entire Marketing Department at the Wharton School and especially Jagmohan Raju, Jehoshua Eliashberg, Peter Fader, Robert Meyer, Raghuram Iyengar, Pinar Yildirim, and Qiaowei Shen for their support and encouragement over the years and Barbara E. Kahn and Deborah Small for their service as PhD coordinator.

My friends and family that have carried me through this journey and I am incredibly grateful for their love and support. To Mom and Dad, your unconditional love and continual belief in me have given me great strength and I am so lucky to be your daughter. To my favorite brother, Kevin, I could not have asked for a more amazing buddy to grow up with. To my grandparents, your selfless devotion, guidance and care have made me a better person. Finally, to my wonderful husband, Yuhang Alan Zhou, you are the most amazing partner and make every day a special adventure.

ABSTRACT

THREE ESSAYS ON THE ESTIMATION OF AVERAGE TREATMENT EFFECTS IN QUASI-EXPERIMENTAL PANEL DATA

Kathleen T. Li

David R. Bell

Christophe Van den Bulte

Identifying average treatment effects (ATE) from quasi-experimental panel data has become one of the most important yet challenging endeavors for social scientists. The difficulty lies in accurately estimating the counterfactual outcomes for the potentially treated units in the absence of treatment. Perhaps the most popular method to estimate average treatment effects is the Difference-in-Differences (DID) method. The key assumption of the DID method is that outcomes of the treated units would have followed a path parallel to the control units in the absence of treatment and violation of this "parallel lines" assumption will result in biased estimates. This dissertation consists of three essays, which either build on existing methods (essay 1 and 3) or propose a new method (essay 2) that can be used even when the "parallel lines" assumption of DID does not hold. In essay 1, we derive the asymptotic distribution of the HCW method, which is computationally simple as it only involves least squares regressions. However, in cases where treatment and control units are positively correlated, the HCW method may have less predictive efficiency than other methods such as the synthetic control and modified synthetic control method, which impose the restriction that weights are non-negative. The popular synthetic control method additionally imposes the restriction that the weights sum to one, which can be a helpful regularization condition when there are many control units. In essay 3, we provide the inference theory for both the synthetic control and modified synthetic control method through projection theory and propose a computational algorithm using subsampling to compute

the confidence intervals. In order to apply the HCW method, synthetic control method and modified synthetic control method, the number of control units needs to be smaller than the pre-treatment sample size. In essay 2, we propose the augmented DID method, which can be used where there are many treatment and control units, but is less flexible than the three aforementioned methods. In short, this dissertation provides several methods and their inference procedures to identify average treatment effects. Which method should be used when depends on the structure of the data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 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 potentially 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 et al., 2006), effect of school term length on student performance (Pischke, 2007), price reactions to rivals' local channel exits (Ozturk et al., 2016), offline bookstore openings' effect on sales at Amazon (Forman et al., 2009), effect of consumer relocation on brand preferences (Bronnenberg et al., 2012), effect of privacy regulation on advertising effectiveness (Goldfarb and Tucker, 2011), effect of 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.

We discuss the advantages and disadvantages of various ATE estimation methods and how they are related to each other. Perhaps the most popular method to estimate average treatment effects is the Difference-in-Differences (DID) method. The key advantage of DID is that it is very simple and easy to implement. This method is especially effective when there are large number of treatment and control units over short time periods. One crucial assumption of the DID method is that outcomes of the treated units would have followed a path parallel to the control units in the absence of treatment. Violation of this "parallel lines" assumption will result in biased estimates. In some cases, propensity score matching paired with DID can help achieve the "parallel lines" assumption. However, this only works if there are covariates available and if those covariates are the ones that matter in terms of

the potential treatment assignment.

For panel data with a relatively large number of time series observations, alternative methods may be better suited than DID for estimating counterfactual outcomes. The synthetic control method proposed by Abadie and Gardeazabal (2003), and Abadie et al. (2010) can be used to estimate average treatment effects (ATE). This method has two attractive features. First, it is more general than the conventional difference-in-differences method because it allows for different control units to have different weights when estimating the counterfactual outcome of the treated unit. Second, the synthetic control method restricts the weights assigned to the units in the control group to be non-negative and therefore may lead to better extrapolation than an estimator without the non-negativity restriction when outcome variables are are positively correlated. In fact, Athey and Imbens (2017) describe the synthetic control method as "arguably the most important innovation in the evaluation literature in the last 15 years". 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 paths over time in the absence of treatment. The condition that the weights sum up to one can be restrictive and lead to poor fit when this synthetic control version of the "parallel lines" assumption is violated.

An even more flexible method is the modified synthetic control method (MSCM) proposed by Doudchenko and Imbens (2016). Their modifications include adding an intercept and dropping the sum-to-one restriction in a standard synthetic control model. Dropping the sum-to-one restriction makes the modified method applicable to a wider range of data settings. Specifically, the modified synthetic control method can handle the case of heterogenous treatment and control units better because it can account for cases where the treatment unit's path is outside the range or convex hull of the control units' paths.

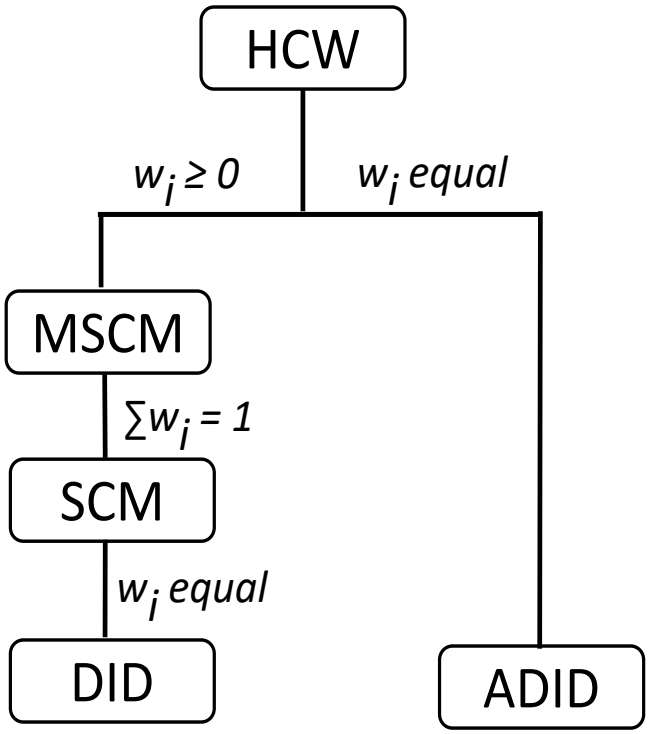
Another method that can be used when treated and control units are heterogenous is proposed by Hsiao, Ching and Wan (2012). The two advantages of the HCW method are that

it does not require that the treatment units and the control units follow parallel paths in the absence of treatment and that it is computationally simple as it only involves least squares regressions. However, in certain cases, compared to the modified synthetic control method, it may have less predictive efficiency if treatment and control units are positively correlated.

In order to apply the synthetic control method, modified synthetic control method and HCW method, the number of control units needs to be smaller than the pre-treatment sample size. In contrast, the DID and augmented DID (ADID) method proposed in essay 2 can be used where there are many treatment and control units. The augmented DID method retains the advantage of simplicity of the DID method but still solves the problem of parallel paths by allowing the average of the control units to be scaled by a factor.

This dissertation consists of three essays, which either build on existing methods or propose a new method for estimating average treatment effects in quasi-experimental panel data. Figure 1 shows the relationship between all these methods. At the top is the HCW method, which is the least restrictive. In essay 1 (Chapter 2), we derive the asymptotic distribution of the HCW method. By imposing the restriction that the weights are non-negative, we arrive at the modified synthetic control method. Additionally imposing that the weights sum to one, we arrive at the synthetic control method. In essay 3 (Chapter 4), we provide the inference theory for both the synthetic control and modified synthetic control method through projection theory and propose a computational algorithm using subsampling to compute the confidence intervals. The popular, widely used DID method is a special case of SCM when all the weights are equal. Finally, augmented DID, which is proposed in essay 2 (Chapter 3), simply has the restriction that the weights are equal. The equal weight restriction for DID and ADID is what lends it to simplicity and ease of use.

Figure 1: Relationship Between Methods



CHAPTER 2 : Asymptotic Theory for HCW Method

2.1. Introduction

Social scientists are often interested in examining average treatment effects (ATE) of some policy interventions. Difference-in-differences (DID) methodology is a popular approach used to estimate average treatment effects (Ashenfelter, 1978; Ashenfelter and Card, 1985; Chevalier and Mayzlin, 2006; Goldfarb and Tucker, 2011). One main advantage of the DID method is that if the underlying assumptions under DID hold, it can consistently estimate the ATE even when the number of time periods is small provided that the number of control units and the number of treatment units are large. While the DID method is relatively straightforward to implement, there are two important limitations to this approach: (i) it assumes that there is no sample selection (bias) effect, i.e., the treatment dummy is considered as strictly exogenous, and (ii) it also assumes that the average outcomes for the treatment and controls units follow parallel paths over time in the absence of treatment (Assumption 3.1 in Abadie (2005)). Recently, Athey and Imbens (2006) generalized the conventional DID method, which is based on linear models to general nonlinear models. They propose an estimation method that can be used to recover the entire counterfactual outcome distribution. However, one crucial assumption made in Athey and Imbens (2006) is that, conditional on some of the individual's unobservable characteristics and in the absence of treatment, the outcome of an individual is the same in a given time period regardless of whether the individual is in the treatment or the control group (see assumption 3.1 and eq. (3) in Athey and Imbens (2006)). Violations to Abadie's (2005) assumption 3.1 or toAthey and Imbens (2006) assumption 3.1 can lead to severely biased estimation results for the DID method.

Recently, Hsiao et al. (2012) proposed a novel method to estimate the average treatment effect (ATE) using panel data. The three advantages of HCW's approach are: (i) it does not need the assumption of no sample selection effect, i.e., it bypasses the issue of correlation

between the treatment dummy and the outcome; (ii) it does not require that the treatment units and the control units follow parallel paths over time in the absence of treatment; (iii) it is computationally simple as it only involves least squares regressions. In short, the HCW method does not require either Abadie (2005) assumption 3.1 or Athey and Imbens (2006) assumption 3.1. For example, even when the treatment units and the control units exhibit substantial individual heterogeneity such as having non-parallel sample paths, under the assumption that the treatment effects are covariance stationary (or trend-stationary), the HCW method can be applied to consistently estimate ATE as long as that the number of time periods before the treatment and the number of time periods after the treatment are large. The HCW method could be considered as an alternative to the popular DID method. In this essay we show that HCW's method can work with less restrictive assumptions than they assumed. Specifically, we relax their linear functional form assumption and remove one of their identification conditions. We also provide a theoretical complement to Hsiao et al. (2012) by deriving the asymptotic distribution of the HCW estimator which facilitates inference.

Another contribution of our essay is that we propose using a better criterion to select the control units. For the HCW method, when faced with a dataset with a large number of control units, one must decide which control units to use. HCW suggest using Akaike information criterion (AIC) or the corrected Akaike information criterion (AICC) model selection criteria to choose the control units. However, the AICC model selection method or other similar methods such as BIC and AIC suffer from the following problems: (i) it cannot be used when the number of control units is larger than the number of time periods before the treatment (pretreatment sample size) because the standard least squares method requires that the number of regressors be less than the sample size; (ii) even when the number of control units is smaller than the pretreatment sample size, the computational burden may make AICC and BIC methods prohibitive. In this essay, we propose using the 'least absolute shrinkage and selection operator' (LASSO) method to select control units when estimating the average treatment effect. The LASSO method shrinks some coefficients

and sets others (less significant ones) to zero. It can avoid both of the above problems regarding model selection and estimation. It is well established that the LASSO method allows for the number of regressors to be larger than the sample size (e.g., Meinshausen and Yu (2009); Bickel et al. (2009)). The adaptive LASSO method can be used for consistently selecting relevant regressors (Zou, 2006; Huang et al., 2008) even when the number of regressors diverges to infinity as sample size increases. Finally, the LASSO method is known to be computationally efficient (Efron et al., 2004). Using simulations, we show that the computational time for the LASSO method is significantly less than that of AICC or BIC methods. Perhaps surprisingly, we show that the LASSO method also leads to significant reductions in predictive mean squared errors (when estimating ATE) compared to conventional model selection procedures such as AIC, AICC, BIC and the 'leave-many-out' cross validation methods.

In summary, we make three contributions. First, we relax HCW's linear functional form assumption and remove one of their identification conditions. Second, we derive the asymptotic distribution of the HCW estimator which facilitates inference. Third, we propose using the LASSO method to select control units and show via simulations that it dominates many conventional methods.

The remaining parts of the essay are organized as follows. Section 2.2 describes HCW's average treatment effect (ATE) estimation method and shows that some distributional assumptions made in HCW can be removed without affecting the consistency of HCW's estimator. Section 2.3 establishes the asymptotic distribution of HCW's ATE estimator. Section 2.4 considers the case when data is trend-stationary. Section 2.5 proposes using the LASSO method to select control units. Section 2.6 presents simulation results to examine finite sample properties of HCW's estimator under various conditions. Finally, Section 2.7 concludes. The proofs of the main results are presented in Appendices A and B.

2.2. HCW method for estimating ATEs

Hsiao et al. (2012) propose a novel method to estimate the effect of a policy intervention, i.e., average treatment effect, using panel data. Below we first introduce their notation and assumptions. Then we relax some of their assumptions to make their method applicable to a wider range of data generating processes. Consider a panel data $\{y_{it}\}_{i=1,t=1}^{N,T}$, where we observe an outcome variable y_{it} (such as GDP, housing prices, sales revenue, etc.) across two dimensions (such as over geographical units and over time). Then at a time period, T_1+1 $(1 < T_1+1 < T)$, a policy intervention or treatment occurs. Let y_{it}^1 and y_{it}^0 denote unit i's outcome in period t with and without treatment, respectively. The treatment effect to the i^{th} unit at time t is defined as $\Delta_{it} = y_{it}^1 - y_{it}^0$. However, for the same unit i, we do not simultaneously observe y_{it}^0 and y_{it}^1. Thus, 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 i^{th} unit is under the treatment at time t and $d_{it} = 0$ otherwise. We observe y_{1t}^1 or y_{1t}^0 depending on whether or not unit i receives a treatment at time t. The difficulty in estimating ATE is how to estimate the counterfactual outcome y_{it}^0 when the i^{th} unit receives a treatment. We discuss HCW's estimation method in the next subsection.

2.2.1. HCW's factor model approach

HCW motivate their estimation method using a factor model approach. Let f_t be a $K \times 1$ vector of *unobservable* common factors which are the main force that drives all the y_{it} to change over time. In this essay we assume that f_t is a weakly stationary process to simplify the exposition. We consider the case where there is no treatment to y_{it} for all i and for $t = 1, ..., T_1$. At $t = T_1 + 1$, one unit receives a treatment and without loss of generality, we let this be the first unit, y_{1t}. Other units y_{jt}, $j = 2, ..., N$, do not receive any treatment. Following HCW (2012) we consider the following factor model (for pre-treatment period)

$$y_{it}^0 = \alpha_i + b_i' f_t + u_{it}, \qquad i = 1, ..., N; \quad t = 1, ..., T_1, \tag{2.2.1}$$

where α_i is i's individual specific intercept, b_i is a $K \times 1$ vector of factor loadings, f_t is a $K \times 1$ vector (unobservable) common factors and u_{it} is a zero mean, weakly dependent and weakly stationary error term.

Let $y_t = (y_{1t}, ..., y_{Nt})'$ be an $N \times 1$ vector of the outcome variables at time t. For the periods prior to the treatment, the observed y_t takes the form

$$y_t = y_t^0 = \alpha + Bf_t + u_t \qquad \text{for } t = 1, ..., T_1, \tag{2.2.2}$$

where $\alpha = (\alpha_1, ..., \alpha_N)'$ is an $N \times 1$ vector of individual intercepts, $B = (b_1, b_2, ..., b_N)'$ is an $N \times K$ factor loading matrix and $u_t = (u_{1t}, ..., u_{Nt})'$ is an $N \times 1$ vector of the error terms. At time $T_1 + 1$, a policy intervention occurs to the first unit. Therefore, for the post-treatment periods, we have $y_{1t} = y_{1t}^1 = \alpha_1 + b_1' f_t + \Delta_{1t} + u_{1t}$ for $t = T_1 + 1, ..., T$, where Δ_{1t} is the treatment effect to the first unit at time t. To estimate the average treatment effect, we need to construct the counterfactual outcomes y_{1t}^0 for $t \geq T_1 + 1$. If T_1 and N are large, the methods of Bai and Ng (2002) and Bai et al. (2014b) can be used to identify the number of common factors K and estimate f_t along with B by the maximum likelihood approach. However, when T_1 and N are not sufficiently large, the number of factors may not be determined correctly and a factor model may not be estimated accurately. Hsiao et al. (2012) suggest a novel method to estimate the counterfactual outcome y_{1t}^0 without the need of estimating unobserved factors and factor loadings. HCW propose using $\tilde{y}_t = (y_{2t}, ..., y_{Nt})'$ in lieu of f_t to predict the counterfactuals y_{1t}^0 for post-treatment periods. The validity of HCW's approach seems to depend on a linear conditional expectation functional form assumption. In this essay we show that even if this linear functional form assumption does not hold, the HCW method still leads to consistent estimation of the average treatment effect. In addition, we derive the asymptotic distribution of HCW's ATE estimator. Hence, using the HCW method one can consistently estimate the average treatment effect and the result is robust to any nonlinear functional form.

2.2.2. Assumptions made in HCW

We first discuss HCW's proposed estimation method and the assumptions HCW made when proving the consistency result of their estimator. Recall that the outcomes for all units can be expressed as $y_t = (y_{1t}, \tilde{y}_t')'$, where $\tilde{y}_t' = (y_{2t}, ..., y_{Nt})$. The pre-treatment data y_t^0 is generated by the factor model (2.2.2). The assumptions in HCW are stated below.

Assumption HCW 1. (i) $||b_i|| = c_i \leq M < \infty$ for all $i = 1, ..., N$, where M is a positive constant; (ii) u_t is a weakly dependent process with $E(u_t) = 0$ and $E(u_t u_t') = V$, where V

is an $N \times N$ diagonal matrix; (iii) $E(u_t f_t') = 0$ for all t; (iv) $E(u_{jt}|d_{1t}) = 0$ for all $j \neq 1$.

Assumption HCW 2.

$Rank(\tilde{B}) = K$, where $\tilde{B}$ is an $N-1$ by K matrix obtained from B by removing the first row of B.

Assumption HCW 3.

For any fixed K and N, there exists an $N \times 1$ vector $a = (1, -\gamma')'$ where $\gamma = (\gamma_2, ..., \gamma_N)'$ such that $a'B = 0$, where $\mathcal{N}(B)$ is the null space of B. At the neighborhood of a, $T_1^{-1} \sum_{t=1}^{T_1} E[(y_{1t}^0 - \delta_1 - \delta' \tilde{y}_t)^2]$ has a unique minimum at $(\delta_{10}, \delta_0')'$.

Using the above assumptions HCW show that $a'y_t \equiv y_{1t} - \gamma' \tilde{y}_t = a'\alpha + a'u_t$ because $a'B = 0$ so that the common factors are dropped out from the right-hand-side of the above equation. Rearranging terms leads to

$$y_{1t} = \gamma_1 + \gamma' \tilde{y}_t + u_{1t}^*, \tag{2.2.3}$$

where $\gamma_1 = a'\alpha$ and $u_{1t}^* = a'u_t = u_{1t} - \gamma' \tilde{u}_t = u_{1t} - \gamma_2 u_{2t} - ... - \gamma_N u_{Nt}$ (recall that $\tilde{u}_t = (u_{2t}, ..., u_{Nt})'$). Because u_{1t}^* depends on all $u_{1t}, ..., u_{Nt}$, it is easy to see that $\tilde{y}_t$ is correlated with u_{1t}^*. Define $\eta_{1t} = u_{1t}^* - E(u_{1t}^*|\tilde{y}_t)$, then $E(\eta_{1t}|\tilde{y}_t) = 0$. HCW assume that

Assumption HCW 4.

$E(u_{1t}^*|\tilde{y}_t) = c_1 + c'\tilde{y}_t$ (linear conditional mean functional form assumption).

Assumption HCW 1 is quite standard and reasonable. Given that in most applications $N-1$ is (much) larger than K, Assumption HCW 2 is quite weak and easily satisfied. Below we show that when assumption HCW 2 is violated, the HCW method may not lead to consistent estimation of ATE. Assumption HCW 3 is less intuitive and we show later that this assumption can be dropped. Assumption HCW 4 is a strong assumption which may or may not hold true in practice. We also show in Section 2.3 that assumption HCW 4 can be removed without affecting the consistency of HCW's proposed estimator.

Writing $u_{1t}^* = E(u_{1t}^*|\tilde{y}_t) + \eta_{1t} = c_1 + c'\tilde{y}_t + \eta_t$ and using assumption HCW 4, (2.2.3) can be

written as

$$y_{1t} = \delta_1 + \delta' \tilde{y}_t + \eta_{1t}, \tag{2.2.4}$$

where $\delta_1 = \gamma_1 + c_1$ and $\delta = \gamma + c$. One can estimate δ_1 and δ by the least squares method regressing y_{1t} on $(1, \tilde{y}_t')$ using the pre-treatment data $t = 1, ..., T_1$. Let $\hat{\delta}_1$ and $\hat{\delta}$ denote the least squares estimator of δ_1 and δ, respectively. The counterfactual outcome y_{1t}^0 is estimated by $\hat{y}_{1t}^0 = \hat{\delta}_1 + \hat{\delta}' \tilde{y}_t$, for $t = T_1 + 1, ..., T$. The treatment effect at period s, for $s \geq T_1 + 1$, is estimated by $\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^0$. And the average treatment effect is estimated by averaging $\hat{\Delta}_{1t}$ over the post-treatment periods:

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

where $T_2 = T - T_1$. Here we would like to emphasize that since there is only one unit (unit 1) that receives the treatment, the ATE estimator is obtained by averaging over the post-treatment periods $t = T_1 + 1, ..., T$ (time series averaging) for unit 1. This differs from the usual DID method in which one often has a larger number of units receiving treatments and the average is usually done over many treatment units (cross sectional averaging).

Under some conditions including HCW 1 to HCW 4, Hsiao et al. (2012) show $\hat{\Delta}_1 - \bar{\Delta}_1 \xrightarrow{p} 0$ as $T_1, T_2 \to \infty$, where $\bar{\Delta}_1 = T_2^{-1} \sum_{t=T_1+1}^{T} \Delta_{1t}$. With some additional assumptions it can also be shown that $\hat{\Delta}_1 \xrightarrow{p} \Delta_1 = E(\Delta_{1t})$. Additional assumptions include that Δ_{1t} is a weakly dependent and weakly stationary process so that a law of large numbers holds: $T_2^{-1} \sum_{t=T_1+1}^{T} \Delta_{1t} \xrightarrow{p} \Delta_1$. We establish the consistency result of $\hat{\Delta}_1$ for Δ_1 under conditions weaker than those used in HCW (2012). In particular, while we maintain assumptions HCW 1 and HCW 2, we remove assumptions HCW 3 and HCW 4.

2.2.3. Our weaker assumptions

In this subsection, we only assume that HCW 1 and HCW 2 hold and remove the other assumptions in HCW. We derive (2.2.4) under weak regularity conditions. Formally, we make the following assumptions:

Assumption 1. Our assumption 1 is the same as assumption HCW 1.

Assumption 2. (i) Let $x_t = (1, \tilde{y}_t')'$. Then, $\{x_t\}_{t=1}^{T}$ is a weakly dependent and weakly stationary process, $T_1^{-1}\sum_{t=1}^{T_1} x_t x_t' \xrightarrow{p} E(x_t x_t')$ as $T_1 \to \infty$, and $[E(x_t x_t')]$ is invertible. (ii) $Rank(\tilde{B}) = K$.

Assumption 2 (i) is not restrictive. If $E(x_t x_t')$ is not invertible, we can remove the linearly dependent regressors and redefine x_t as a subset of $(1, \tilde{y}_t')'$ such that assumption 2 (i) holds. Under assumption 2 (ii) $\tilde{B}'\tilde{B}$ is a $K \times K$ invertible matrix. Let $\tilde{y}_t$, $\tilde{\alpha}$ and $\tilde{u}_t$ be $(N-1) \times 1$ vectors obtained by removing the first rows of y_t, α and u_t, respectively. Then we have $\tilde{y}_t = \tilde{\alpha} + \tilde{B} f_t + \tilde{u}_t$. Multiplying this by $\tilde{B}'$ and then solving for f_t, we obtain

$$f_t = (\tilde{B}'\tilde{B})^{-1}\tilde{B}'(\tilde{y}_t - \tilde{\alpha} - \tilde{u}_t). \tag{2.2.5}$$

Substituting (2.2.5) into (2.2.1) for $i = 1$ we get

$$y_{1t} = \alpha_1 + b_1'(\tilde{B}'\tilde{B})^{-1}\tilde{B}'(\tilde{y}_t - \tilde{\alpha} - \tilde{u}_t) + u_{1t}. \tag{2.2.6}$$

Re-arranging terms in (2.2.6) we obtain

$$y_{1t} = \gamma_1 + \gamma'\tilde{y}_t + \epsilon_{1t}, \tag{2.2.7}$$

where $\gamma_1 = \alpha_1 - b_1'(\tilde{B}'\tilde{B})^{-1}\tilde{B}'\tilde{\alpha}$, $\gamma' = b_1'(\tilde{B}'\tilde{B})^{-1}\tilde{B}'$ and $\epsilon_{1t} = u_{1t} - b_1'(\tilde{B}'\tilde{B})^{-1}\tilde{B}'\tilde{u}_t$.

Note that equation (2.2.7) implies that one vector a satisfying $a'B = 0$ is given by $a' = (1, -b_1'(\tilde{B}'\tilde{B})^{-1}\tilde{B}')$. Indeed using $B = (b_1,\ \tilde{B}')'$ it is easy to check that

$$a'B = b_1' - b_1'(\tilde{B}'\tilde{B})^{-1}\tilde{B}'\tilde{B} = b_1' - b_1' = \mathbf{0}_K', \tag{2.2.8}$$

where $\mathbf{0}_K$ is a K by 1 vector of zeros. Equations (2.2.7) and (2.2.8) show that assumption HCW 3 holds true. Therefore, we have shown that our assumption 2 (ii) implies assumption HCW 3, and there is no need to make assumption HCW 3 as an additional assumption.

Next, we show that HCW's estimator remains to be a consistent estimator for ATE without assumption HCW 4. When the linear conditional mean assumption HCW 4 is violated,

instead of using the conditional mean projection decomposition, we use a linear projection decomposition. First, we give a formal definition of linear projection of ϵ_{1t} on $(1, \tilde{y}_t')$. We assume that $(y_t, \epsilon_{1t})_{t=1}^{T_1}$ is a weakly stationary process and define c_1 and c to be the minimizers of $\min_{c_1,c} E[(\epsilon_{1t} - c_1 - c'\tilde{y}_t)^2]$, where $c = (c_2, ..., c_N)'$. Then we call $c_1 + c'\tilde{y}_t$ the linear projection of ϵ_{1t} onto $(1, \tilde{y}_t')$ and use $L(\epsilon_{1t}|\tilde{y}_t)$ to denote it. Hence, we decompose ϵ_{1t} into $\epsilon_{1t} = L(\epsilon_{1t}|\tilde{y}_t) + v_{1t}$, where $v_{1t} = \epsilon_{1t} - L(\epsilon_{1t}|\tilde{y}_t)$. We re-write (2.2.7) as

$$\begin{aligned} y_{1t} &= \gamma_1 + \gamma'\tilde{y}_t + L(\epsilon_{1t}|\tilde{y}_t) + v_{1t} \\ &= \gamma_1 + \gamma'\tilde{y}_t + c_1 + c'\tilde{y}_t + v_{1t} \\ &= \delta_1 + \delta'\tilde{y}_t + v_{1t} \end{aligned} \tag{2.2.9}$$

where $\delta_1 = \gamma_1 + c_1$ and $\delta = \gamma + c$. Because $L(v_{1t}|\tilde{y}_t) = 0$, using pre-treatment data $t = 1, ..., T_1$, least squares regression of y_{1t} on $(1, \tilde{y}_t')$ gives consistent estimators for δ_1 and δ.

In practice when N is large, one may use an $m \times 1$ sub-vector of $\tilde{y}_t$ (with $m \geq K$) to replace the $K \times 1$ unobservable common factor f_t. We address this model selection issue in Section 2.5. Note that v_{1t} in (2.2.9) only satisfies $L(v_{1t}|\tilde{y}_t) = 0$. Its conditional mean, conditional on $\tilde{y}_t$, may not be 0. We show that for consistent estimation of the average treatment effect Δ_1, we do not need the condition that $E(v_{1t}|\tilde{y}_t) = 0$. Instead, we only need $E(v_{1t}) = 0$ and $E(\tilde{y}_t v_{1t}) = 0$ (zero unconditional moments) which are implied by $L(v_{1t}|\tilde{y}_t) = 0$.

Define $\beta = (\delta_1, \delta')'$ and let $\hat{\beta}$ be the least squares estimator of β. Under the assumptions that y_t is a weakly dependent and weakly stationary process, and that $E(x_t x_t')$ is invertible,[1] where $x_t = (1, \tilde{y}_t')'$, and using a law of large numbers argument, we show in Appendix A that $\hat{\beta} \xrightarrow{p} \beta_0$, where $\beta_0 = [E(x_t x_t')]^{-1} E[x_t y_{1t}]$. Note that some components of the $(N-1) \times 1$ vector β_0 can be zero, but from assumption 2 we know that β_0 has at least K non-zero components. With the above defined β_0 we can re-write (2.2.9) as

$$y_{1t} = x_t'\beta_0 + v_{1t}. \tag{2.2.10}$$

Note that even though we may have $E(v_{1t}|\tilde{y}_t) \neq 0$, we always have $L(v_{1t}|\tilde{y}_t) = 0$, i.e., v_{1t}

[1] If $E(x_t x_t')$ is not invertible, we can always remove the redundant components of x_t to make it invertible.

in (2.2.9) satisfies the following conditions $E(v_{1t}) = 0$ and $E(\tilde{y}_t v_{1t}) = 0$. This is the reason why the least squares method can consistently estimate $\beta_0 = (\delta_1, \delta')'$ in (2.2.10).

2.2.4. Average treatment effect estimator

Let $\hat{\delta}_1$ and $\hat{\delta}$ denote the least squares estimators of δ_1 and δ from (2.2.9). Then, we can estimate the counterfactual y_{1t}^0 by

$$\hat{y}_{1t}^0 = \hat{\delta}_1 + \hat{\delta}'\tilde{y}_t \tag{2.2.11}$$

for $t = T_1 + 1, ..., T$. The treatment effect at time t is estimated by $\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^0$ $(t = T_1 + 1, ..., T)$ and the average treatment effect is estimated by

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

Note that the above estimation procedure is identical to that of HCW (2012) except that we derived it using assumptions HCW 1 and HCW 2 only (removing assumptions HCW 3 and HCW 4).

For post treatment periods we add a treatment effect term to obtain

$$y_{1t} = \delta_1 + \delta'\tilde{y}_t + \Delta_{1t} + v_{1t}, \qquad \text{for } t = T_1 + 1, ..., T, \tag{2.2.12}$$

where Δ_{1t} is the treatment effect for unit 1 at time t. We allow for Δ_{1t} to be random.

2.3. Consistency and asymptotic distribution of HCW estimator

2.3.1. Consistency

Let $\Delta_1 = E(\Delta_{1t})$ be the average treatment effect for the first unit. In this section, we show that $\hat{\Delta}_1 - \Delta_1 = O_p(T_1^{-1/2} + T_2^{-1/2})$, which implies that as T_1, $T_2 \to \infty$, $\hat{\Delta}_1 - \Delta_1 \xrightarrow{p} 0$. We first make an additional assumption.

Assumption 3. (i) $\hat{\delta}_1 - \delta_1 = O_p(T_1^{-1/2})$ and $\hat{\delta} - \delta = O_p(T_1^{-1/2})$, (ii) $Var(T_2^{-1} \sum_{t=T_1+1}^{T} y_t) = O(T_2^{-1})$; (iii) $Var(T_2^{-1} \sum_{t=T_1+1}^{T} \Delta_{1t}) = O(T_2^{-1})$.

Assumption 3 is quite weak and it holds if $\{y_t\}_{t=1}^{T}$ and $\{\Delta_{1t}\}_{t=T_1+1}^{T}$ are weakly dependent processes such as some ARMA processes. Note that assumption 3 (ii) implies that $T_2^{-1}\sum_{t=T_1+1}^{T}(\tilde{y}_t - E(\tilde{y}_t)) = O_p(T_2^{-1/2})$ and $T_2^{-1}\sum_{t=T_1+1}^{T} v_{1t} = O_p(T_2^{-1/2})$ (by noting that $v_{1t} = y_{1t} - \delta_1 - \delta'\tilde{y}_t$ and $E(v_{1t}) = 0$).

We present the consistency result (with convergence rate) in the following proposition.

Proposition 2.3.1 *Under assumptions 1 to 3, we have*

$$\hat{\Delta}_1 - \Delta_1 = O_p(T_1^{-1/2} + T_2^{-1/2}). \tag{2.3.1}$$

The proof of proposition 2.3.1 is given in Appendix A.

From the proof presented in Appendix A, we know that estimation errors come from two parts. The first part comes from estimating δ_1 and δ using the pre-treatment data, which is of order $O_p(T_1^{-1/2})$ because the pre-treatment sample size is T_1. The second part comes from the average of v_{1t} for $t = T_1 + 1, ..., T$, which has an order $O_p(T_2^{-1/2})$. Therefore, a consistent estimation result requires both T_1 and T_2 to be large. Equation (2.3.1) also implies that when both T_1 and T_2 are doubled, the estimation mean squared error of $\hat{\Delta}_1 - \Delta_1$ will be halved. This prediction is confirmed by simulations.

Finally, we comment on the above consistency proof. We used assumptions such as $\hat{\delta}_1 - \delta_1 = O_p(T_1^{-1/2})$ and $\hat{\delta} - \delta = O_p(T_1^{-1/2})$. These assumptions rule out the case that y_t have time trend components. However, for many empirical data sets used for estimating the average treatment effect, outcome variables have an upward trend (e.g., housing price in Bai et al. (2014a); Du and Zhang (2015), and economic stimulus package in Ouyang and Peng (2015)). It can be shown that the consistency result still holds when $\{y_t\}_{t=1}^{T_1}$ and $\{y_t\}_{t=T_1+1}^{T}$ are trend-stationary processes. We study the trend-stationary data case in Section 2.4.

2.3.2. Asymptotic result with stationary data

To derive the asymptotic distribution of $\hat{\Delta}_1 - \Delta_1$, we need to make some additional assumptions.

Assumption 4. Let $\hat{\beta} = (\hat{\delta}_1, \hat{\delta}')'$. The least squares estimator $\hat{\beta}$ is asymptotically normally

distributed: $\sqrt{T_1}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V)$, where $\beta = (\delta_1, \delta')'$ and V is a finite, positive definite matrix.

Assumption 5. For $t = T_1 + 1, ..., T$, both v_{1t} and Δ_{1ts} are weakly dependent and weakly stationary processes so that central limit theorems apply to their partial sums, i.e., $T_2^{-1/2} \sum_{t=T_1+1}^{T} (\Delta_{1t} - E(\Delta_{1t}) + v_{1t}) \xrightarrow{d} N(0, \Sigma_2)$, where Σ_2 is the asymptotic variance of $T_2^{-1/2} \sum_{t=T_1+1}^{T} (\Delta_{1t} - E(\Delta_{1t}) + v_{1t})$. When v_{1t} and Δ_{1t} are serially uncorrelated, we have $\Sigma_2 = Var(\Delta_{1t} + v_{1t})$.

Assumption 6. Let $w_t \overset{def}{=} (\tilde{y}_t', v_{1t})'$ for $t = 1, ..., T$. We assume that w_t is a weakly stationary ρ-mixing process with ρ-mixing coefficients $\rho(\tau) = O(\lambda^{\tau})$ for some constant $0 < \lambda < 1$.

Assumption 7. Let $\eta = \lim_{T_1, T_2 \to \infty} T_2/T_1$. We assume that $\eta \geq 0$ is a finite non-negative constant.

Assumptions 4 and 5 are quite weak. Laws of large numbers and central limit theorems are known to hold for many weakly dependent and weakly stationary processes. Assumption 6 deals with the full data set, $t = 1, ..., T$ and requires that the data is a weakly stationary ρ-mixing process with an exponential decay rate. Assumption 6 can be relaxed to $\rho(\tau) = O(\tau^{-c})$ for some constant $c > 4$ with a much lengthier proof. Many weakly dependent processes, including some strictly stationary ARMA processes, are known to be ρ-mixing processes with exponential decay rates (e.g., Carrasco and Chen (2002)). Assumption 7 implies that T_2 has an order smaller or equal to that of T_1. This assumption is quite reasonable because it only rules out the cases where $T_2/T_1 \to \infty$ as $T_1, T_2 \to \infty$, which corresponds to a case with a relatively small number of pre-treatment observations and a much larger number of post-treatment observations. In such a case, the least squares estimator is not reliable due to the relatively small T_1, and the long range extrapolation (due to large T_2) may not yield reliable out-of-sample forecasting results.

We present the asymptotic distribution result of $\hat{\Delta}_1 - \Delta_1$ in the following theorem.

Theorem 2.3.2 *Under assumptions 1 to 7, we have*

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

where $\Sigma = \Sigma_1 + \Sigma_2$, $\Sigma_1 = \eta E(x_t)' V E(x_t)$, $\eta = \lim_{T_1,T_2\to\infty} T_2/T_1$, *and* Σ_2 *is defined in assumption 5.*

We can see from Theorem 2.3.2 that Σ_1 comes from estimation error using pre-treatment data (since V = asymptotic $var(\sqrt{T_1}\hat{\beta})$) and Σ_2 comes from the averaging over $\Delta_{1t} - \Delta_1 + v_{1t}$ for $t > T_1$. These estimation errors arise because we can only estimate $\delta_1 + \delta'\tilde{y}_t$, the systematic part of y_{1t}^0. Therefore, we can only consistently estimate the ATE, but not the treatment effect at each period because $\hat{y}_{1t}^0 - y_{1t}^0 = v_{1t} + o_p(1) \neq o_p(1)$.[2] Note that if T_1 is much larger than T_2, then Σ_1 is negligible because $T_2/T_1 \approx 0$ and the asymptotic variance reduces to $\Sigma = \Sigma_2$ in this case. This result is quite intuitive because when T_1 is much larger than T_2, the first stage estimation error becomes negligible compared to the post-treatment averaging over v_{1t}. Therefore, Σ_1 is negligible compared with Σ_2.

A consistent estimator of $\Sigma_1 = Avar(A_1)$ is given by $\hat{\Sigma}_1 = (T_2/T_1)\hat{E}(x_t)'\hat{V}\hat{E}(x_t)$, where $\hat{E}(x_t) = (1, \hat{E}(\tilde{y}_t'))'$, $\hat{E}(\tilde{y}_t) = T_1^{-1}\sum_{t=1}^{T_1}\tilde{y}_t$ and $\hat{V}$ is a consistent estimator of $Var(\sqrt{T_1}\hat{\beta})$. Hence, we estimate Σ_1/T_2 by

$$\hat{\Sigma}_1/T_2 \quad = \quad (T_2/T_1)\hat{E}(x_t)'(\hat{V}/T_2)\hat{E}(x_t), \tag{2.3.2}$$

where $\hat{V}/T_2$ is an estimator of $(T_1/T_2)Var(\hat{\beta})$. Since we allow for v_{1t} and Δ_{1t} to be serially correlated processes, we suggest using some autocorrelation robust estimator to estimate Σ_2:

$$\hat{\Sigma}_2 = \frac{1}{T_2}\sum_{t=T_1+1}^{T}\sum_{s=T_1+1,|t-s|\leq l}^{T}\left[\hat{\Delta}_{1t} - \hat{\Delta}_1\right]\left[\hat{\Delta}_{1s} - \hat{\Delta}_1\right], \tag{2.3.3}$$

where $\hat{\Delta}_1 = T_2^{-1}\sum_{t=T_1+1}^{T}\hat{\Delta}_{1t}$, $l \to \infty$ and $l/T_2 \to 0$ as $T_2 \to \infty$. For example, one may choose $l = O(T_2^{1/4})$ (see Newey and West (1987) and White (1984)) or use a faster rate for l (see Andrews (1991)). If both Δ_{1t} and v_{1t} are serially uncorrelated, then Σ_2 can be consistently estimated by

$$\tilde{\Sigma}_2 = \frac{1}{T_2}\sum_{t=T_1+1}^{T}\left[\hat{\Delta}_{1t} - \hat{\Delta}_1\right]^2 \qquad \text{and} \qquad \tilde{\Sigma}_2/T_2 = \frac{1}{T_2^2}\sum_{t=T_1+1}^{T}\left[\hat{\Delta}_{1t} - \hat{\Delta}_1\right]^2. \tag{2.3.4}$$

[2]Recall that $y_{1t}^0 = \delta_1 + \delta'\tilde{y}_t + v_{1t}$. Hence, $\hat{y}_{1t}^0 - y_{1t}^0 = \hat{\delta}_1 - \delta_1 + (\hat{\delta} - \delta)'\tilde{y}_t + v_{1t} = v_{1t} + O_p(T_1^{-1/2}) \neq o_p(1)$ due to $v_{1t} \neq o_p(1)$. We need to average v_{1t} over $t = T_1+1, ..., T$ to obtain $T_2^{-1}\sum_{t=T_1+1}^{T} v_{1t} = O_p(T_2^{-1/2}) = o_p(1)$.

Following the same arguments as in Newey and West (1987), one can show that $\hat{\Sigma}_2$ defined in (A.8) is a consistent estimator for Σ_2, i.e., $\hat{\Sigma}_2 = \Sigma_2 + o_p(1)$.

The above asymptotic result can be used to test the null hypothesis of no (average) treatment effect. Let the null hypothesis be H_0: $\Delta_1 = 0$. We can test H_0 versus the two-sided alternative H_1: $\Delta_1 \neq 0$ or a one-sided hypothesis H_1: $\Delta_1 > (<) 0$. Then our test statistic is given by

$$T.S. \stackrel{def}{=} \frac{\sqrt{T_2}\,\hat{\Delta}_1}{\sqrt{\hat{\Sigma}}} \equiv \frac{\hat{\Delta}_1}{\sqrt{\hat{\Sigma}/T_2}} \quad \stackrel{H_0}{\to} \quad N(0,1), \tag{2.3.5}$$

where $\hat{\Sigma} = \hat{\Sigma}_1 + \hat{\Sigma}_2$, $\hat{\Sigma}_1$ is defined in (A.7) and $\hat{\Sigma}_2$ is defined in (A.8). If the data is serially uncorrelated, one can replace $\hat{\Sigma}_2$ by $\tilde{\Sigma}_2$ defined in (A.9). Equation (2.3.5) implies that we reject H_0 at the 5% level if $|T.S.| > 1.96$ ($T.S. > 1.645$ or $T.S. < -1.645$ for a one-sided test), and we do not reject H_0 otherwise.

2.4. The trend-stationary data case

Many empirical data, such as the online eyeglasses sales data we use in Chapter 3 and 4, exhibit upward trends. In this section, we extend the result of Section 2.3 to the trend-stationary data case. Following HCW (2012) we assume that $y_t^0 = (y_{1t}^0, y_{2t}^0, ..., y_{Nt}^0)'$ is generated via a factor model

$$y_t^0 = \delta_0 + Bf_t + u_t, \tag{2.4.1}$$

where $\delta_0 = (\delta_{01}, ..., \delta_{0N})'$ is an $N \times 1$ vector of intercepts, B is an $N \times k$ factor loading matrix, $f_t = (f_{1t}, ..., f_{kt})'$ is a $k \times 1$ vector of common factor, $u_t = (u_{1t}, ..., u_{Nt})'$ is an $N \times 1$ vector of idiosyncratic errors. We assume that $f_{1t} = t$ and all other factors are stationary variables. Also, u_t is a zero mean, weakly dependent process with finite fourth moment. Hence, y_t^0 follows a trend-stationary process.

We have shown that one can replace the unobservable factor f_t by $\tilde{y}_t = (y_{2t}, ..., y_{Nt})'$ to estimate the counterfactual outcome y_{1t}^0. Specifically, one can estimate the following regression model

$$y_{1t} = \delta_1 + \tilde{y}_t'\delta + v_{1t}, \qquad (t = 1, ..., T_1), \tag{2.4.2}$$

where $\tilde{y}_t = (y_{2t}, ..., y_{Nt})'$ and $\delta = (\delta_2, ..., \delta_N)'$. Let $\hat{\delta}_1$ and $\hat{\delta}$ be the least squares estimators of δ_1 and δ, respectively. Then one estimate y_{1t}^0 by $\hat{y}_{1t}^0 = \hat{\delta}_1 + \tilde{y}_t'\hat{\delta}$ for $t = T_1 + 1, ..., T$.

To facilitate the asymptotic analysis, we consider the time trend component explicitly. We write $y_{jt} = c_{0,j} + c_{1,j}t + y^*_{jt}$, where y^*_{jt} is a weakly dependent stationary process (detrended from y_{jt}) for $j = 2, ..., N$. In vector notation, we have $\tilde{y}_t = \tilde{c}_0 + \tilde{c}_1 t + \tilde{y}^*_t$, where $\tilde{y}_t = (y_{2t}, ..., y_{Nt})'$, $\tilde{c}_0 = (c_{0,2}, ..., c_{0,N})'$, $\tilde{c}_1 = (c_{1,2}, ..., c_{1,N})'$ and $\tilde{y}^*_t = (\tilde{y}^*_{2t}, ..., \tilde{y}^*_{Nt})'$. Then we can write $\delta'\tilde{y}_t = \delta'(\tilde{c}_0 + \tilde{c}_1 t + \tilde{y}^*_t)$. Hence, we can re-write (2.4.2) as

$$\begin{aligned} y_{1t} &= \delta_1 + \delta'\tilde{y}_t + v_{1t} \\ &= \alpha t + \beta_1 + \delta'\tilde{y}^*_t + v_{1t} \\ &= \alpha t + z'_t\beta + v_{1t} \qquad\qquad t = 1, ..., T_1, \end{aligned} \tag{2.4.3}$$

where $\alpha = \delta'\tilde{c}_1$, $\beta_1 = \delta_1 + \delta'\tilde{c}_0$, $\beta = (\beta_1, \delta')'$ and $z_t = (1, \tilde{y}^{*'}_t)' \equiv (1, y^*_{2t}, ..., y^*_{Nt})'$.

Let $\hat{\alpha}$ and $\hat{\beta}$ be the least squares estimators of α and β obtained from estimating (2.4.3) using the pre-treatment data. We estimate y^0_{1t} by $\hat{y}^0_{1t} = \hat{\alpha}t + z'_t\hat{\beta}$ and estimate the ATE

$$\hat{\Delta}_1 = \frac{1}{T_2}\sum_{t=T_1+1}^{T} \hat{\Delta}_{1t}, \tag{2.4.4}$$

where $\hat{\Delta}_{1t} = y_{1t} - \hat{y}^0_{1t}$.

2.4.1. Asymptotic theory with trend stationary data

In this section, we derive the asymptotic distribution of the ATE estimator $\hat{\Delta}_1$ defined in (2.4.4). For the post-treatment, we have $y^1_{1t} = y^0_{1t} + \Delta_{1t}$. Hence, we have for $t = 1, ..., T$,

$$y_{1t} = \alpha t + z'_t\beta + d_t\Delta_{1t} + v_{1t}, \tag{2.4.5}$$

where $d_t = 0$ for $t \leq T_1$ and $d_t = 1$ for $t \geq T_1 + 1$.

Let $\hat{\alpha}$ and $\hat{\beta}$ be the least squares estimators of α and β based on (2.4.3). Then it is well established that $\hat{\alpha} - \alpha = O_p(T_1^{-3/2})$ and $\hat{\beta} - \beta = O_p(T_1^{-1/2})$ (e.g., Hamilton (1994), Chapter

16). Thus, using (2.4.4) and (2.4.5) we have

$$\begin{aligned}
\hat{\Delta}_1 - \Delta_1 &= \frac{1}{T_2}\sum_{t=T_1+1}^{T}\left[y_{1t} - \hat{y}_{1t}^0\right] - \Delta_1 \\
&= \frac{1}{T_2}\sum_{t=T_1+1}^{T}\left[(\alpha-\hat{\alpha})t - z_t'(\hat{\beta}-\beta) + \Delta_{1t} - \Delta_1 + v_{1t}\right] \\
&= -\left[\frac{2T_1+T_2+1}{2}\right](\hat{\alpha}-\alpha) - [E(z_t') + o_p(1)](\hat{\beta}-\beta) + \frac{1}{T_2}\sum_{t=T_1+1}^{T}\epsilon_{1t}, \quad (2.4.6)
\end{aligned}$$

where we used $\sum_{t=T_1+1}^{T} t = (T_1+1+T)T_2/2 = (2T_1+T_2+1)T_2/2$, $\epsilon_{1t} = \Delta_{1t} - \Delta_1 + v_{1t}$.

Hence,

$$\begin{aligned}
\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) &= -\sqrt{T_2/T_1}\left[\frac{2+T_2/T_1}{2}\right]\sqrt{T_1^3}(\hat{\alpha}-\alpha) \quad (2.4.7)\\
&\quad -\sqrt{T_2/T_1}E(z_t')\sqrt{T_1}(\hat{\beta}-\beta) + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T}\epsilon_{1t} + o_p(1) \\
&= -\left(\sqrt{T_2/T_1}(2+T_2/T_1)/2\,,\ \sqrt{T_2/T_1}E(z_t')\right)\begin{pmatrix}\sqrt{T_1^3}(\hat{\alpha}-\alpha)\\ \sqrt{T_1}(\hat{\beta}-\beta)\end{pmatrix} \quad (2.4.8)\\
&\quad +\frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T}\epsilon_{1t} + o_p(1) \\
&= -c'D_{T_1}(\hat{\gamma}-\gamma) + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T}\epsilon_{1t} + o_p(1), \quad (2.4.9)
\end{aligned}$$

where $c = (\sqrt{\eta}(2+\eta)/2, \sqrt{\eta}E(z_t'))'$, $\eta = \lim_{T_1,T_2\to\infty} T_2/T_1$, $\hat{\gamma} = (\hat{\alpha}, \hat{\beta}')'$ and $\gamma = (\alpha, \beta')'$, $D_{T_1} = \sqrt{T_1}diag(T_1, 1, ..., 1)$ which is an $(N+1)\times(N+1)$ diagonal matrix with the first diagonal element equals to $T_1^{3/2}$ and all other diagonal elements equal to $\sqrt{T_1}$.

To establish the asymptotic distribution of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$, we make the following assumptions.

Assumption 8. Let $z_t = (1, y_{2t}^{*'}, ..., y_{Nt}^{*})'$. We assume that (i) $\{z_t\}_{t=1}^{T}$ is a weakly dependent and weakly stationary process, $T_1^{-1}\sum_{t=1}^{T_1} z_t z_t' \xrightarrow{p} E(z_t z_t')$ as $T_1 \to \infty$, and $[E(z_t z_t')]$ is invertible; (ii) $D_{T_1}(\hat{\gamma}-\gamma) \xrightarrow{d} N(0, \Omega)$, where Ω is a positive definite matrix.

Assumption 9. Let $\epsilon_{1t} = \Delta_{1t} - \Delta_1 + v_{1t}$. Then $T_2^{-1/2} \sum_{t=T_1+1}^{T} \epsilon_{1t} \xrightarrow{d} N(0, \Sigma_2)$ as $T_2 \to \infty$, where $\Sigma_2 = \lim_{T_2 \to \infty} T_2^{-1} \sum_{t=T_1+1}^{T} \sum_{s=T_1+1}^{T} E(\epsilon_{1t}\epsilon_{1s})$ is the asymptotic variance of $T_2^{-1/2} \sum_{t=T_1+1}^{T} \epsilon_{1t}$.

Assumption 10. Let $w_t = (\epsilon_{1t}, y_{2t}^*, ..., y_{Nt}^*)'$. We assume that w_t is a ρ-mixing process with the mixing coefficient $\rho(\tau)$ satisfies the condition: $\rho(\tau) \leq C\,\lambda^{\tau}$ for some finite positive constants $C > 0$ and $0 < \lambda < 1$, where $\rho(\tau) = \max_{1 \leq i,j \leq N} \frac{|Cov(w_{it}, w_{j,t+\tau})|}{\sqrt{Var(w_{it}) Var(w_{j,t+\tau})}}$, and w_{it} is the i^{th} component of w_t for $i = 1, ..., N$.

Assumptions 8 and 9 are not restrictive. They require that (z_t, ϵ_{1t}) to be a weakly dependent stationary process so that the law of large numbers and the central limit theorem hold for their (partial) sums. If $E(z_t z_t')$ is not invertible, we can remove the linearly dependent regressors and redefine z_t as a subset of $(1, y_{2t}^*, ..., y_{Nt}^*)'$ such that assumption 8 holds. Assumption 10 further imposes an exponential decay rate for the ρ-mixing processes. Many ARMA processes are known to be ρ-mixing with exponential decay rate.

By Assumption 10 and the proof of Theorem 2.3.2, we know that $\hat{\gamma} - \gamma$ is asymptotically independent with $T_2^{-1/2} \sum_{t=T_1+1}^{T} \epsilon_{1t}$. Therefore, from (2.4.7) we immediately have the following result.

Theorem 2.4.1 *Under assumptions A8 to A10 we have*

$$\begin{aligned} \sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) &= -c' D_{T_1}(\hat{\gamma} - \gamma) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} \epsilon_{1t} + o_p(1) \\ &\xrightarrow{d} N(0, \Sigma), \end{aligned} \tag{2.4.10}$$

where $\Sigma = \Sigma_1 + \Sigma_2$ *with* $\Sigma_1 = c'\Omega c$ *and* $\Sigma_2 = \lim_{T_2 \to \infty} T_2^{-1} \sum_{t=T_1+1}^{T} \sum_{s=T_1+1}^{T} E(\epsilon_{1t}\epsilon_{1s})$.

Note that if y_t is a stationary process, then $\alpha = 0$ and we remove $\hat{\alpha}$ from γ so that $\hat{\gamma}$ becomes $\hat{\beta}$. Also, A and B become $E(z_t z_t')$ and B_{22}, respectively. The result reduces to that of Section 2.3.

Note that $\Sigma_1 = \lim_{T_1 \to \infty} Var(-c' D_{T_1}(\hat{\gamma} - \gamma)) = \lim_{T_1 \to \infty} c' D_{T_1} Var(\hat{\gamma}) D_{T_1} c$. Hence, we can estimate Σ_1 by

$$\hat{\Sigma}_1 = \hat{c}' D_{T_1} \widehat{Var}(\hat{\gamma}) D_{T_1} \hat{c}, \tag{2.4.11}$$

where $\widehat{Var}(\hat{\gamma})$ is an estimator of $Var(\hat{\gamma})$, for example, one can use the Newey and West (1987) autocorrelation and heteroskedasticity robust variance-covariance estimator for $\widehat{Var}(\hat{\gamma})$, $\hat{c} = (\sqrt{T_2/T_1}(T_2/T_1+2)/2, \sqrt{T_2/T_1}\hat{E}(z_t'))'$, $\hat{E}(z_t) = T_1^{-1}\sum_{t=1}^{T_1}\hat{z}_t$ with $\hat{z}_t = (1, \hat{y}_{2t}^*, ..., \hat{y}_{Nt}^*)'$ and $\hat{y}_{jt}^*$ is the de-trended version of y_{jt} for $j = 2, ..., N$.[3]

Σ_2 can be consistently estimated by

$$\hat{\Sigma}_2 = T_2^{-1}\sum_{t=T_1+1}^{T}\sum_{s=T_1+1,|s-t|\leq l}^{T}\hat{\epsilon}_{1t}\hat{\epsilon}_{1s}, \tag{2.4.12}$$

where $\hat{\epsilon}_{1t} = \hat{\Delta}_{1t} - \hat{\Delta}_1$, $l = O(T_2^{1/4})$ as suggested by Newey and West (1987).

Remark 2.4.2 *Note that as long as t and y_{jt}'s ($j = 2, ..., N$) are not perfectly collinear, whether one uses $\tilde{y}_t$, or uses de-trended data $\tilde{y}_t^*$ in (2.4.3), to estimate ATE, one obtains exactly the same numerical value for $\hat{\Delta}_1$. However, when one estimates the asymptotic variance Σ, one must use de-trended data. For example, when one estimates $E(z_t z_t')$, which is finite by assumption, one must use the de-trended data y_{jt}^* for $j = 2, ..., N$ in $\hat{E}(z_t z_t') = T_1^{-1}\sum_{t=1}^{T_1}\hat{z}_t\hat{z}_t'$, where $\hat{z}_t = (1, \hat{y}_{2t}^*, ..., \hat{y}_{Nt}^*)'$, $\hat{y}_{jt}^* = y_{jt} - \hat{c}_{0,j} - \hat{c}_{1,j}t$, where $\hat{c}_{0,j}$ and $\hat{c}_{1,j}$ are the least squares estimators of $c_{0,j}$ and $c_{1,j}$ in $y_{jt} = c_{0,j} + c_{1,j}t + y_{jt}^*$ for $j = 2, ..., N$. If one uses $\tilde{y}_t$ in computing $\hat{E}(z_t z_t')$, the estimator will explode to ∞ asymptotically, which leads to significantly overestimated value of $E(z_t z_t')$ in finite sample applications.*

2.4.2. Explicit expression for Ω

Recall that Ω is the asymptotic variance of $D_{T_1}(\hat{\gamma} - \gamma)$. Similar to the analysis in Hamilton (1994, Chapter 16) one can show that

$$D_{T_1}(\hat{\gamma} - \gamma) \xrightarrow{d} N(0, \Omega), \tag{2.4.13}$$

[3]For $j = 2, ..., N$, we estimate the de-trended variable y_{jt}^* via the regression model $y_{jt} = c_j t + y_{jt}^*$, then $\hat{y}_{jt}^* = y_{jt} - \hat{c}_{0,j} - \hat{c}_{1,j}t$, where $\hat{c}_{0,j}$ and $\hat{c}_{1,j}$ are the least squares estimators of $c_{0,j}$ and $c_{1,j}$, respectively.

where $\Omega = A^{-1}BA^{-1}$,

$$A = \begin{pmatrix} 1/3 & (1/2)E(z_t') \\ (1/2)E(z_t) & E(z_t z_t') \end{pmatrix}$$
$$B = \begin{pmatrix} B_{11} & B_{12}' \\ B_{12} & B_{22} \end{pmatrix}$$

with

$$\begin{aligned} B_{11} &= \lim_{T_1\to\infty} T_1^{-3} \sum_{t=1}^{T_1}\sum_{s=1}^{T_1} t\, sE(v_{1t}v_{1s}), \\ B_{12} &= \lim_{T_1\to\infty} T_1^{-2} \sum_{t=1}^{T_1}\sum_{s=1}^{T_1} tE(v_{1t}v_{1s}z_s), \\ B_{22} &= \lim_{T_1\to\infty} T_1^{-1} \sum_{t=1}^{T_1}\sum_{s=1}^{T_1} E(v_{1t}v_{1s}z_t z_s'). \end{aligned}$$

If v_{1t} is serially uncorrelated, then B simplifies to

$$\begin{aligned} B_{11} &= (1/3)E(v_{1t}^2), \\ B_{12} &= (1/2)E(v_{1t}^2 z_t), \\ B_{22} &= E(v_{1t}^2 z_t z_t'). \end{aligned}$$

We outline the proof of (2.4.13) below. Let $m_t \equiv (t, z_t')' = (t, 1, y_{2t}^*, ..., y_{Nt}^*)'$. Then (2.4.3) can be written as

$$y_{1t} = m_t'\gamma + v_{1t}, \tag{2.4.14}$$

where $\gamma = (\alpha, \beta')'$.

Denote by $Y_1 = (y_{11}, ..., y_{1T_1})'$, $v_1 = (v_{11}, ..., v_{1T_1})'$ and let M be the $T_1 \times (N+1)$ matrix with its t^{th} row given by $m_t' = (t, 1, y_{2t}^*, ..., y_{Nt}^*)$. Then in matrix notation (2.4.14) can be written as

$$Y_1 = M\gamma + v_1.$$

Hence, $\hat{\gamma} = (M'M)^{-1}M'Y_1 = \gamma + (M'M)^{-1}M'v_1$. Thus, we have

$$\begin{aligned} D_{T_1}(\hat{\gamma} - \gamma) &= D_{T_1}(M'M)^{-1}D_{T_1}D_{T_1}^{-1}M'v_1 \\ &= (D_{T_1}^{-1}M'MD_{T_1}^{-1})^{-1}D_{T_1}^{-1}M'v_1 \\ &\xrightarrow{d} A^{-1}N(0,B) = N(0, A^{-1}BA^{-1}), \end{aligned} \tag{2.4.15}$$

because by noting that $m_t m_t' = \begin{pmatrix} t \\ z_t \end{pmatrix}(t, z_t') = \begin{pmatrix} t^2 & tz_t' \\ tz_t & z_t z_t' \end{pmatrix}$ and $m_t v_{1t} = \begin{pmatrix} tv_{1t} \\ z_t v_{1t} \end{pmatrix}$, we have

$$D_{T_1}^{-1}M'MD_{T_1}^{-1} = \begin{pmatrix} T_1^{-3}\sum_{t=1}^{T_1} t^2 & T_1^{-2}\sum_{t=1}^{T_1} tz_t' \\ T_1^{-2}\sum_{t=1}^{T_1} tz_t & T_1^{-1}\sum_{t=1}^{T_1} z_t z_t' \end{pmatrix} \xrightarrow{p} A$$

by a law of large numbers argument and

$$D_{T_1}^{-1}M'v_1 = \begin{pmatrix} T_1^{-3/2}\sum_{t=1}^{T_1} tv_{1t} \\ T_1^{-1/2}\sum_{t=1}^{T_1} z_t v_{1t} \end{pmatrix} \xrightarrow{p} N(0,B)$$

by a central limit theorem argument as in Hamilton (1994).

We estimate Σ_1 by $\hat{\Sigma}_1 = \hat{c}'\hat{\Omega}\hat{c}$, $\hat{\Omega} = \hat{A}^{-1}\hat{B}\hat{A}^{-1}$, $\hat{c}$ was defined earlier and

$$\hat{A} = \begin{pmatrix} 1/3 & (1/2)\hat{E}(z_t') \\ (1/2)\hat{E}(z_t) & \hat{E}(z_t z_t') \end{pmatrix},$$

with $\hat{E}(z_t) = T_1^{-1}\sum_{t=1}^{T_1}\hat{z}_t$, $\hat{E}(z_t z_t') = T_1^{-1}\sum_{t=1}^{T_1}\hat{z}_t\hat{z}_t'$, $\hat{z}_t = (1, \hat{y}_{2t}^*, ..., \hat{y}_{Nt}^*)'$, $\hat{y}_{jt}^*$ is the de-trend

version of y_{jt} for $j = 2, ..., N$. And we estimate B by

$$\begin{aligned}
\hat{B}_{11} &= \frac{1}{T_1^3} \sum_{t=1}^{T_1} \sum_{s=1,|s-t|\leq l}^{T_1} t\, s\, \hat{z}_t \hat{z}_s', \\
\hat{B}_{12} &= \frac{1}{T_1^2} \sum_{t=1}^{T_1} \sum_{s=1,|s-t|\leq l}^{T_1} t\, \hat{v}_{1t} \hat{v}_{1s} \hat{z}_s, \\
\hat{B}_{22} &= \frac{1}{T_1} \sum_{t=1}^{T_1} \sum_{s=1,|s-t|\leq l}^{T_1} \hat{v}_{1t} \hat{v}_{1s} \hat{z}_t \hat{z}_s'.
\end{aligned}$$

If v_{1t} is serially uncorrelated, $\hat{B}_{11}$, $\hat{B}_{12}$ and $\hat{B}_{22}$ can be simplified by changing s to t and removing the summation $\sum_{s=1.|s-t|\leq l}^{T_1}$, i.e., $\hat{B}_{11} = T_1^{-3} \sum_{t=1}^{T_1} t^2 \hat{z}_t \hat{z}_t'$, $\hat{B}_{12} = T_1^{-2} \sum_{t=1}^{T_1} t\, \hat{v}_{1t}^2\, \hat{z}_t$ and $\hat{B}_{22} = T_1^{-1} \sum_{t=1}^{T_1} \hat{v}_{1t}^2 \hat{z}_t \hat{z}_t'$.

2.5. Selecting significant control units by LASSO

When there is a large number of control units, using all of them to estimate ATE may not be the best choice because a large number of regressors usually leads to large estimation variance, which in turn leads to poor out-of-sample predictions. HCW suggest using AIC type approach to select control units. Du and Zhang (2015) propose using a 'leave-many-out' cross validation method (Shao, 1993) to choose control units and show via simulations that the cross-validation method gives smaller out-of-sample prediction results than AICC/BIC approach. In this essay, we recommend using the LASSO method to select important control variables. There are several advantages of using the LASSO method to select control units and to estimate ATE. First, there may be more control units than the number of pre-treatment time periods ($N > T_1$). In such a case, conventional model selection methods such as BIC, AIC and AICC all break down and some modifications are needed in order to use some modified BIC, AIC and AICC methods to select control units when $N > T_1$ (see Section 2.6.3 for a detailed discussion on a modified AICC method). In contrast, the LASSO method allows for $N > T_1$. In fact, it even allows for N to be of a larger magnitude than T_1, i.e., $N/T_1 \to \infty$ as $T_1 \to \infty$. Second, the LASSO method is known to be computationally efficient. We compare the computation costs (time) of various methods in the next (simulation) section. Finally and perhaps most surprisingly, we show that the LASSO method has smaller out-of-sample prediction errors than (modified) AICC, BIC,

AIC and leave-many-out cross validation methods.

2.5.1. The LASSO method

It is well known that linear regression has generally low bias (or zero bias, when the true model is linear) but can have high variance especially when there is a large number of regressors, which may lead to poor predictions. Modern statistical methods introduce some bias but significantly reduce the variance, leading to better predictive accuracy. LASSO is one such method and it is an effective way to select significant variables with high dimensional data. Consider a linear regression model

$$y_{1t} = x_t'\beta + v_{1t}, \qquad t = 1, ..., T_1,$$

where $x_t' = (1, \tilde{y}_t')$ and $\beta = (\delta_1, \delta')'$ is an $N \times 1$ vector of unknown parameters. The number of regressors $N = N_{T_1}$ may be of the same magnitude as T_1 or of a larger magnitude than that of T_1, i.e., $N_{T_1}/T_1 \to \infty$ as $T_1 \to \infty$. In either case, some sparsity assumptions are needed to estimate the model. Usually, one assumes that only a subset of m components of β is non-zero with either m being a finite number or $m = m_{T_1}$, where $m_{T_1} \to \infty$ and $m_{T_1}/T_1 \to 0$ as $T_1 \to \infty$.

The LASSO method selects β to minimize

$$\sum_{t=1}^{T_1} [y_{1t} - x_t'\beta]^2 + \lambda \sum_{j=1}^{N} |\beta_j|, \tag{2.5.1}$$

where $\lambda \geq 0$ is a tuning parameter. The larger the value of λ, the more penalty is imposed on non-zero β_j. Hence, LASSO shrinks β_j toward zero for all j. This introduces some bias, but reduces the variance significantly. Two extreme cases are $\lambda = 0$ and $\lambda = \infty$. The former does not put any constraint on β while the latter shrinks all β_j to 0. In order to obtain a LASSO estimator of β, we need to select a value of λ. Suppose that we search for λ over a discrete set $\Lambda_L = \{\lambda_1, \lambda_2, ..., \lambda_L\}$. Common practices include using BIC or the 'leave-one-out' cross validation methods to select the tuning parameter λ. We discuss 'leave-one-out' cross validation method in the next subsection.

2.5.2. Selection of the tuning parameter: The 'leave-one-out' cross validation (CV) method

For each $\lambda \in \Lambda_L \equiv \{\lambda_1, ..., \lambda_L\}$, and for each $t = 1, ..., T_1$, we compute β by minimizing the following 'leave-one-out' objective function

$$\sum_{s=1, s\neq t}^{T_1} (y_{1s} - x_s'\beta)^2 + \lambda \sum_{l=1}^{N} |\beta_l|. \tag{2.5.2}$$

Let the minimizer to (2.5.2) be $\hat{\beta}_{-t,\lambda}$. We then calculate the error on the validation point t: $e_t(\lambda) = y_{1t} - x_t'\hat{\beta}_{-t,\lambda}$. For each tuning parameter value λ, we compute the average squared error over all T_1 observations:

$$CV(\lambda) = \frac{1}{T_1}\sum_{t=1}^{T_1} e_t(\lambda)^2 = \frac{1}{T_1}\sum_{t=1}^{T_1}(y_{1t} - x_t'\hat{\beta}_{-t,\lambda})^2, \tag{2.5.3}$$

where $\hat{\beta}_{-t,\lambda}$ is the leave-one-out (leave the t^{th} observation out) estimate of β that minimizes (2.5.2). We select $\lambda \in \Lambda_L$ that minimizes $CV(\lambda)$ defined in (2.5.3).

2.6. Simulation results

In this section, we report simulation results to examine the finite sample performance of the HCW average treatment effect estimator. We examine the average treatment effect estimator based on both HCW's assumptions and our weaker assumptions. We show that indeed the HCW estimator works well under our weaker assumptions. Therefore, the simulation results strongly support our theoretical result.

2.6.1. Examining nonlinearity of $E(\epsilon_{1t}|\tilde{y}_t)$

In this subsection we choose $N = 3$ (one treatment and two control units) and $K = 1$ (one common factor) in our simulations. We use the following data generating process (without treatment). The unobservable common factor f_t is generated by an AR(1) process $f_t = 0.5f_{t-1} + v_t$, where v_t is iid $N(0, 1)$. Let y_t^0 denote the $N \times 1$ vector of outcome variables without treatment. Then it is generated via $y_t^0 = a + bf_t + u_t$, $t = 1, ..., T$, where $y_t^0 = (y_{1t}^0, y_{2t}^0, y_{3t}^0)'$, a and b are 3×1 vectors with $a = (1, 1, 1)'$, $b = (1, 1, 1)'$ and

$u_t = (u_{1t}, u_{2t}, u_{3t})'$. For the distribution of u_t, we consider two cases. The first case has a linear conditional mean function (assumption HCW 4 is satisfied), while the second case does not have a linear conditional mean function (assumption HCW4 is violated). Specifically, for $j = 1, 2, 3$, we generate

$$\begin{aligned} DGP1: &\quad u_{jt} \text{ as iid } N(0,1); \\ DGP2: &\quad u_{jt} \text{ is iid uniform}[-2,2]. \end{aligned}$$

It is easy to see that for DGP1, $E(u_{jt}|\tilde{y}_t)$ is linear in $\tilde{y}_t$ while for DGP2, $E(u_{jt}|\tilde{y}_t)$ is *not* linear in $\tilde{y}_t$.[4] For $t \geq T_1+1$, the first unit receives a treatment Δ_{1t} at t. Therefore, $y_{1t} = y_{1t}^1$ is generated by $y_{1t} = y_{1t}^0 + \Delta_{1t}$, where Δ_{1t} is the treatment unit 1 at time t and is generated by $\Delta_{1t} = exp(z_t)/[1 + exp(z_t)] + 1$ for $t = T_1 + 1, ..., T$, where $z_t = 0.5z_{t-1} + \epsilon_t$, ϵ_t is iid $N(0, 0.25)$. The error ϵ_s, $s = T_1 + 1, ..., T$, is independent of $(\tilde{y}_t, u_{jt})$ for all $t = 1, ..., T$ and $j = 1, 2, 3$. We choose $T_1 = 50$, 100, 200 or 400 and $T_2 = 20$, 40, 80, 160 or 320. To assess the finite sample performance of the HCW estimator, we compute the mean squared error which is defined as

$$MSE = \frac{1}{M}\sum_{j=1}^{M}[\hat{\Delta}_{1,j} - \bar{\Delta}_{1,j}]^2,$$

where $M = 10,000$ is the number of replications and the subscript j denotes the estimation result for the j^{th} replication. Also, $\hat{\Delta}_{1,j} = T_2^{-1}\sum_{t=T_1+1}^{T}\hat{\Delta}_{1t,j}$ and $\bar{\Delta}_{1,j} = T_2^{-1}\sum_{t=T_1+1}^{T}\Delta_{1t,j}$ for $j = 1, ..., M$.

2.6.2. Estimation results for DGP1 and DGP2 ($N = 3$, $K = 1$)

Simulation results for DGP1 and DGP2 for using $\tilde{y}_t$ to replace f_t are reported in Table 1. Table 1 reveals the following: (i) The MSE decreases as T_1 (T_2) increases while holding T_2 (T_1) fixed; (ii) We know that when both T_1 and T_2 are doubled, the MSE is expected to be halved. Indeed Table 1 results strongly support our theoretical analysis that the MSE is halved when both T_1 and T_2 are doubled; (iii) The results for DGP1 and DGP2 are very similar and this strongly supports our analysis that the consistent estimation result does

[4] In the Appendix B we consider a simple case of $N = 2$ with $y_{jt}^0 = f_t + u_{jt}$, $f_t \sim N(0, a_t)$, $u_{jt} \sim$ Uniform$[-c, c]$, $j = 1, 2$. For DGP2 we show that $E(u_{2t}|y_{2t}) = y_{2t} + \sqrt{a_t}[e^{-(y_{2t}+c)^2/(2a_t)} - e^{-(y_{2t}-c)^2/(2a_t)}]/\left\{\sqrt{2\pi}\left[\Phi\left(\frac{c-y_{2t}}{\sqrt{a_t}}\right) - \Phi\left(\frac{-(c+y_{2t})}{\sqrt{a_t}}\right)\right]\right\}$, where $\Phi(\cdot)$ is the cdf of a standard normal random variable, which is obviously a nonlinear function of y_{2t}.

not rely on a linear conditional mean function; (iv) For large values of T_1, MSE reduces to about half when T_2 is doubled. This is as expected since when T_1 is large, we can estimate the unknown parameter δ_1 and δ quite accurately and the MSE becomes proportional to $1/T_2$; (v) When T_1 is not large, MSE still decreases as T_2 rises, but it reduces less than 50% when T_2 is doubled. This is because when T_1 is not large enough, the error in estimating δ_1 and δ is not negligible. Therefore, it affects the second stage MSE.

Table 1: MSE of $\hat{\Delta}_1$ using $\tilde{y}_t = (y_{1t}, y_{2t})'$ in place of f_t

	DGP1					DGP2				
$T_1 \setminus T_2$	20	40	80	160	320	20	40	80	160	320
50	0.114	0.076	0.054	0.044	0.038	0.157	0.097	0.074	0.059	0.053
100	0.097	0.055	0.037	0.026	0.022	0.124	0.075	0.050	0.035	0.029
200	0.088	0.046	0.028	0.018	0.013	0.111	0.065	0.036	0.024	0.017
400	0.080	0.043	0.023	0.014	0.009	0.108	0.058	0.031	0.019	0.012

2.6.3. LASSO method: A factor model with large N ($N > T_1$)

As we argued earlier, an appealing feature of the LASSO method is that it allows for the number of regressors to be larger than the sample size ($N > T_1$), while the traditional AICC method cannot be used when $N > T_1$. The AICC method can be modified to handle the case of $N > T_1$. For example, if there exists a positive integer $m \geq 2$ such that $(m-1)T_1 \leq N < mT_1$, one can divide the N units into m groups such that each group contains less than T_1 control units. Then one can use AICC method to select variables in each group. The process can be repeated until the number of selected variables is less than T_1. Below we illustrate this procedure for the case of $m = 3$. The procedure consists of the following steps:

(a) Suppose that $2T_1 \leq N < 3T_1$, then one divides the $N-1$ control units into three (non-overlapping) groups and uses AICC to select the best regressors in each group. Let $\tilde{y}_{1t}^*$, $\tilde{y}_{2t}^*$, $\tilde{y}_{3t}^*$ denote the selected regressors for group 1, 2 and 3, respectively.

(b) If the sum of the number of regressors in $(\tilde{y}_{1t}^*, \tilde{y}_{2t}^*, \tilde{y}_{3t}^*)$ is less than T_1, one uses AICC to select the best approximations from $(\tilde{y}_{1t}^*, \tilde{y}_{2t}^*, \tilde{y}_{3t}^*)$. If the sum exceeds T_1, one divides $(\tilde{y}_{1t}^*, \tilde{y}_{2t}^*, \tilde{y}_{3t}^*)$ into two/three groups and repeats step (a) and (b) until the sum of selected regressors is less than T_1. Then one uses AICC to select the final regressors [5].

[5] We thank an anonymous reviewer for this suggestion to modify the AICC method to handle the case of

We use the same data generating process as in Hsiao et al. (2012) and Du and Zhang (2015) to examine the finite sample performances of the above modified AICC method and the LASSO method. We generate model (2.2.2) with $N = 31$, 51 or 61, $T_1 = 25$ and $T = T_1+10$, i.e., $T_2 = 10$. We consider the same three-factor model as in Hsiao et al. (2012) and Du and Zhang (2015).

$$\begin{aligned} f_{1t} &= 0.8f_{1t-1} + v_{1t}, \\ f_{2t} &= -0.68f_{1t-1} + v_{2t} + 0.8v_{2t-1}, \\ f_{3t} &= v_{3t} + 0.9v_{3t-1} + 0.4v_{3t-2}, \end{aligned} \tag{2.6.1}$$

where v_{it} is iid $N(0, 1)$.

Let y_t^0 denote the $N \times 1$ vector of outcome variables without treatment. It is generated via

$$DGP3: \qquad y_t^0 = a + Bf_t + u_t, \qquad\qquad t = 1, ..., T, \tag{2.6.2}$$

where $y_t^0 = (y_{1t}^0, y_{2t}^0, ..., y_{Nt}^0)'$, a and b are two $N \times 1$ vectors with $a = (a_1, a_2, ..., a_N)'$, $B = (b_1, b_2, b_3)$, $b_j = (b_{j1}, ..., b_{jN})'$, $f_t = (f_{1t}, f_{2t}, f_{3t})'$ and $u_t = (u_{1t}, u_{2t}, ..., u_{Nt})'$. We choose $(a_1, a_2, ..., a_N) = (1, 1, ..., 1)$, b_{ji} is iid $N(1, 1)$ and u_{jt} iid $N(0, \sigma^2)$ with $\sigma^2 = 1$, 0.5, 0.1 (for $j = 1, 2, 3$; $i = 1, ..., N$). Following Hsiao et al. (2012) and Du and Zhang (2015) we compare the post-treatment (out-of-sample) predicted mean squared error

$$PMSE = \frac{1}{T_2} \sum_{t=T_1+1}^{T} \left(\hat{y}_{1t}^0 - y_{1t}^0\right)^2.$$

The simulation results for DGP 3 (i.e., the three factor model (2.6.2)) are reported in Table 2. Because the PMSE for AIC are the largest for all cases, we choose not to report the results for the AIC case to save space.

From Table 2 we observe the following: (i) The LASSO method gives smaller PMSE than the AICC method for all cases considered; (ii) As N increases, the LASSO gives stable results for both Ave.# (average numbers of selected control units) and PMSE. By contrast, the AICC method tends to select more regressors and its PMSE deteriorates as N increases.

$N > T_1$.

Table 2: DGP3: The Case of $N > T_1$. Out of sample PMSE

	$\sigma^2 = 1$		$\sigma^2 = 0.5$		$\sigma^2 = 0.1$	
	Lasso	AICC	Lasso	AICC	Lasso	AICC
	$N = 31$, $T_1 = 25$, $T = 35$					
Ave.#	7.015	10.37	6.550	11.12	5.135	11.50
Ave. PMSE	1.771	2.330	0.9616	1.261	0.2162	0.2661
Time (Sec.)	0.053	0.227	0.053	0.228	0.048	0.2318
	Lasso	AICC	Lasso	AICC	Lasso	AICC
	$N = 51$, $T_1 = 25$, $T = 35$					
Ave.#	8.35	13.19	7.49	14.09	5.72	13.94
Ave. PMSE	1.634	2.766	0.845	1.446	0.2182	0.292
Time (Sec.)	0.0689	17.26	0.0574	17.41	0.544	17.42
	Lasso	AICC	Lasso	AICC	Lasso	AICC
	$N = 61$, $T_1 = 25$, $T = 35$					
Ave.#	8.72	15.71	7.89	15.84	5.95	15.98
Ave. PMSE	1.654	3.596	0.824	1.796	0.2054	0.3532
Time (second)	0.0695	179.7	0.7804	182.7	0.0644	183.0

Given that the number of factors is three, we know that one needs at least three control units to replace the unobserved factors in order to consistently estimate the ATE. However, simulation results suggest that one should select the number of regressors to be slightly larger than the number of (unobserved) factors in order to minimize the PMSE. The LASSO method selects 7-9 regressors on average (for $\sigma^2 = 1$ case) which gives very good PMSE results while the AICC method seems to select too many regressors which hurts its out-of-sample prediction results. Moreover, as N increases, the AICC method tends to select even more regressors and its performance deteriorates. In contrast, the LASSO method only selects slightly more regressors as N goes up, and its out-of-sample performance improves (slightly) as N increases.

The rows with time (second) report computation time (in seconds) for each simulation replication (using an Intel i7-class processor running at 3.39GHz and using a matlab code). We see that the computational advantage of the LASSO method over AICC/AIC becomes more pronounced as N increases. The ratio of computation time of AICC to LASSO is about 4.283 for $N = 31$ and it increases to 250.5 when $N = 51$ and further to 2588.5 for $N = 61$. When N is increased from 31 to 61, the LASSO computation time only increases by about 31%, while the AICC computation time is increased by $79,060\%$.

We see that the LASSO method is computationally more efficient compared to other con-

ventional methods. Moreover, a pleasant surprise is that the LASSO method also gives smaller out-of-sample predictive MSE than those of the conventional methods such as AIC, AICC and leave-many-out cross validation (not reported here to save space).

2.7. Conclusion

This essay makes three contributions: (i) We relax some of the distributional assumptions made in HCW (2012) and show that the HCW method works for a much wider range of data generating processes; (ii) We derive the asymptotic distribution of HCW's ATE estimator under weak regularity conditions; (iii) We propose using the LASSO method to select control units, and show that it is computationally more efficient than many of the conventional model selection methods and it gives more accurate out-of-sample prediction result than conventional approaches such as AIC, AICC and the leave-many-out cross validation methods. The LASSO method also has the appealing feature that it works well even when the number of control units is larger than the sample size ($N > T_1$). We hope that the results of this essay will make the HCW method more attractive to applied researchers.

Appendix A: Proofs of the main results

A.1. Linear projection

In this section, we consider a generic nonlinear regression model and study the property of a linear projection. We consider the following regression model

$$y_t = g(x_t) + \epsilon_t, \qquad t = 1, ..., n, \tag{A.1}$$

where y_t is a scalar and $x_t \in \mathcal{R}^d$, ϵ_t satisfies $E(\epsilon_t|x_t) = 0$. Hence, $g(x_t) = E(y_t|x_t)$. We assume that there does *not* exist a $(d+1)\times 1$ vector $(\alpha, \gamma')' \in \mathcal{R}^{d+1}$ such that $g(x) = \alpha + x'\gamma$ for almost all $x \in \mathcal{R}^d$. Therefore, the conditional mean function is an unspecified nonlinear function of x_t. Let $\alpha + x_t'\gamma \equiv z_t'\beta$ be the linear projection of y_t on $(1, x_t')$, where $z_t = (1, x_t')'$ and $\beta = (\alpha, \gamma')'$. We can re-write (A.1) as

$$\begin{aligned} y_t &= z_t'\beta + [g(x_t) - z_t'\beta] + \epsilon_t \\ &\equiv z_t'\beta + v_t, \end{aligned} \tag{A.2}$$

where $v_t = [g(x_t) - z_t'\beta] + \epsilon_t$. Let $\hat{\beta}$ be the least squares estimator of β. Then by a law of large numbers argument, we know that

$$\begin{aligned} \hat{\beta} &\xrightarrow{p} [E(z_t z_t')]^{-1} E[z_t y_t] \\ &= [E(z_t z_t')]^{-1} E[z_t g(x_t)] \\ &\equiv \beta, \end{aligned} \tag{A.3}$$

where we used $y_t = g(x_t) + \epsilon_t$ and $E[z_t \epsilon_t] = 0$.

The linear projection error v_t is defined as $v_t = y_t - z_t'\beta = g(x_t) + \epsilon_t - z_t'\beta$. Using (A.3), we have

$$\begin{aligned} E(z_t v_t) &= E[z_t g(x_t)] - E(z_t z_t')\beta + E(\epsilon_t z_t) \\ &= E[z_t g(x_t)] - E(z_t z_t')[E(z_t z_t')]^{-1} E[z_t g(x_t)] \\ &= 0. \end{aligned}$$

Obviously, $E(z_t v_t) = E[(1, x_t')' v_t] = 0$ means that $E(v_t) = 0$ and $E(x_t v_t) = 0$. That is, the linear projection error is orthogonal to 1 and x_t. Therefore, $L(v_t|z_t) = 0$ is sufficient to ensure that the least squares estimator $\hat{\beta}$ is a consistent estimator of β even though $E(v_t|x_t) = g(x_t) - z_t'\beta \equiv g(x_t) - \alpha - x_t'\gamma \neq 0$.

A.2. Proof of Proposition 2.3.1.

Using (2.2.11) and (2.2.12) we have

$$\begin{aligned} \hat{\Delta}_1 &= \frac{1}{T_2} \sum_{t=T_1+1}^{T} [y_{1t} - \hat{y}_{1t}^0] \\ &= (\delta_1 - \hat{\delta}_1) + (\delta - \hat{\delta})' \frac{1}{T_2} \sum_{t=T_1+1}^{T} \tilde{y}_t + \bar{\Delta}_1 + \frac{1}{T_2} \sum_{t=T_1+1}^{T} v_{1t} \\ &= \bar{\Delta}_1 + (\delta_1 - \hat{\delta}_1) + (\delta - \hat{\delta})' [E(\tilde{y}_t) + o_p(1)] + \frac{1}{T_2} \sum_{t=T_1+1}^{T} v_{1t} \\ &= \bar{\Delta}_1 + O_p(T_1^{-1/2} + T_2^{-1/2}), \end{aligned} \tag{A.4}$$

where $\bar{\Delta}_1 = T_2^{-1}\sum_{s=T_1+1}^{T}\Delta_{1s}$, the first two equalities follow from the definitions of $\hat{\Delta}_1$ and $\hat{y}_{1t}^0$, the third equality follows from assumption 3 (ii) and the last equality follows from assumption 3 (i) and (ii).

Finally, using (A.4) and assumption 3 (iii), we have $\hat{\Delta}_1 - \Delta_1 = \hat{\Delta}_1 - \bar{\Delta}_1 + \bar{\Delta}_1 - \Delta_1 = O_p(T_1^{-1/2} + T_2^{-1/2})$. This completes the proof of Proposition 2.3.1.

A.3. Proof of Theorem 2.3.2

From (A.4) we have $\hat{\Delta}_1 - \Delta_1 = (\delta_1 - \hat{\delta}_1) + (\delta - \hat{\delta})'E(\tilde{y}_t) + \frac{1}{T_2}\sum_{s=T_1+1}^{T}[\Delta_{1s} - \Delta_1 + v_{1s}] + o_p(T_1^{-1/2})$. Therefore,

$$\begin{aligned}
\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) &= \sqrt{T_2}(\delta_1 - \hat{\delta}_1) + E(\tilde{y}_t')\sqrt{T_2}(\delta - \hat{\delta}) + \frac{1}{\sqrt{T_2}}\sum_{s=T_1+1}^{T}[\Delta_{1s} - \Delta_1 + v_{1s}] + o_p(1) \\
&\equiv \sqrt{T_2/T_1}E(x_t')\sqrt{T_1}(\beta - \hat{\beta}) + \frac{1}{\sqrt{T_2}}\sum_{s=T_1+1}^{T}[\Delta_{1s} - \Delta_1 + v_{1s}] + o_p(1) \\
&\equiv A_1 + A_2 + o_p(1), \qquad \text{(A.5)}
\end{aligned}$$

where

$$\begin{aligned}
A_1 &= \sqrt{T_2/T_1}E(x_t')\sqrt{T_1}(\beta - \hat{\beta}) \\
A_2 &= \frac{1}{\sqrt{T_2}}\sum_{s=T_1+1}^{T}[\Delta_{1s} - \Delta_1 + v_{1s}], \qquad \text{(A.6)}
\end{aligned}$$

$\hat{\beta} = (\hat{\delta}_1, \hat{\delta}')'$, $\beta = (\delta_1, \delta')'$ and $x_t = (1, \tilde{y}_t')'$.

In equation (A.13) and Lemma A.1 (see below) together imply that $Cov(A_1, A_2) = O(T_1^{-1})$. Thus, A_1 and A_2 are all (asymptotically) uncorrelated. Hence, the asymptotic variance of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$ is given by $Var(A_1 + A_2) = Var(A_1) + Var(A_2) \to \Sigma_1 + \Sigma_2$, where $\Sigma_j = Avar(A_j)$ for $j = 1, 2$. Here, $Avar(A_j) = \lim_{T_1,T_2\to\infty} Var(A_j)$.

From $A_1 = \sqrt{(T_2/T_1)}E(x_t')\sqrt{T_1}(\beta - \hat{\beta})$, it is easy to see that a consistent estimator of $\Sigma_1 = Avar(A_1)$ is given by $\hat{\Sigma}_1 = (T_2/T_1)\hat{E}(x_t)'\hat{V}\hat{E}(x_t)$, where $\hat{E}(x_t) = (1, \hat{E}(\tilde{y}_t'))'$, $\hat{E}(\tilde{y}_t) =$

$T_1^{-1}\sum_{t=1}^{T_1}\tilde{y}_t$ and $\hat{V}$ is a consistent estimator of $Var(\sqrt{T_1}\hat{\beta})$. Hence, we estimate Σ_1/T_2 by

$$\hat{\Sigma}_1/T_2 \;=\; (T_2/T_1)\hat{E}(x_t)'(\hat{V}/T_2)\hat{E}(x_t), \tag{A.7}$$

where $\hat{V}/T_2$ is an estimator of $(T_1/T_2)Var(\hat{\beta})$. Since we allow for v_{1s} and Δ_{1s} to be serially correlated processes, we suggest using some autocorrelation robust estimator to estimate Σ_2:

$$\tilde{\Sigma}_2 = \frac{1}{T_2}\sum_{t=T_1+1}^{T}\sum_{s=T_1+1,|t-s|\leq l}^{T}\left[\hat{\Delta}_{1t}-\hat{\Delta}_1\right]\left[\hat{\Delta}_{1s}-\hat{\Delta}_1\right], \tag{A.8}$$

where $\hat{\Delta}_1 = T_2^{-1}\sum_{s=T_1+1}^{T}\hat{\Delta}_{1s}$, $l\to\infty$ and $l/T_2\to 0$ as $T_2\to\infty$. For example, one may choose $l = O(T_2^{1/4})$ (see Newey and West (1987) and **?**) or use a faster rate for l (see Andrews (1991)).

Following the same arguments as in Newey and West (1987), one can show that $\tilde{\Sigma}_2$ defined in (A.8) is a consistent estimator for Σ_2, i.e., $\tilde{\Sigma}_2 = \Sigma_2 + o_p(1)$. To save space we provide a simple consistency proof under an additional assumption that both Δ_{1s} and v_{1s} are serially uncorrelated processes. In this case, it is easy to show that a consistent estimator of Σ_2 is given by

$$\hat{\Sigma}_2 = \frac{1}{T_2}\sum_{s=T_1+1}^{T}\left[\hat{\Delta}_{1s}-\hat{\Delta}_1\right]^2 \qquad \text{and} \qquad \hat{\Sigma}_2/T_2 = \frac{1}{T_2^2}\sum_{s=T_1+1}^{T}\left[\hat{\Delta}_{1s}-\hat{\Delta}_1\right]^2. \tag{A.9}$$

We now show that $\hat{\Sigma}_2 = \Sigma_2 + O_p(T_1^{-1/2}+T_2^{-1/2})$. Note that $\hat{\Delta}_{1s} = y_{1s}-\hat{y}_{1s}^0 = x_s'(\beta-\hat{\beta})+\Delta_{1s}+v_{1s} = \Delta_{1s}+v_{1s}+O_p(T_1^{-1/2})$, which leads to $\hat{\Delta}_1 = \bar{x}'(\beta-\hat{\beta})+\bar{\Delta}_1+\bar{v}_1 = \bar{\Delta}_1 + O_p(T_1^{-1/2}+T_2^{-1/2})$ because $(\beta-\hat{\beta}) = O_p(T_1^{-1/2})$ and $\bar{v}_1 = T_2^{-1}\sum_{s=T_1+1}^{T}v_{1s} = O_p(T_2^{-1/2})$.

Hence,

$$\begin{aligned}
\hat{\Sigma}_2 &= \frac{1}{T_2}\sum_{s=T_1+1}^{T}\left[\hat{\Delta}_{1s}-\hat{\Delta}_1\right]^2 \\
&= \frac{1}{T_2}\sum_{s=T_1+1}^{T}\left[\Delta_{1s}+v_{1s}-\hat{\Delta}_1\right]^2 + O_p(T_1^{-1/2}+T_2^{-1/2}) \\
&= \frac{1}{T_2}\sum_{s=T_1+1}^{T}\left[\Delta_{1s}-E(\Delta_{1s})+v_{1s}\right]^2 + O_p(T_1^{-1/2}+T_2^{-1/2}) \\
&\equiv \Sigma_2 + O_p(T_1^{-1/2}+T_2^{-1/2}), \qquad \text{(A.10)}
\end{aligned}$$

where in the above we have used $\bar{\Delta}_1 = \Delta_1 + O_p(T_2^{-1/2})$.

Finally, we show that $Cov(A_1, A_2) = O(T_1^{-1})$. We first need to obtain the leading term of $\hat{\beta}-\beta$. Using a law of large numbers argument, we have

$$\begin{aligned}
\hat{\beta}-\beta &= \left[E(x_t x_t') + o_p(1)\right]^{-1}\frac{1}{T_1}\sum_{t=1}^{T_1} x_t v_{1t} \\
&= [E(x_t x_t')]^{-1}\frac{1}{T_1}\sum_{t=1}^{T_1} x_t v_{1t} + o_p(T_1^{-1/2}), \qquad \text{(A.11)}
\end{aligned}$$

because $T_1^{-1}\sum_{t=1}^{T_1} x_t v_{1t} = O_p(T_1^{-1/2})$. Substituting (A.11) into (A.5) we obtain

$$\begin{aligned}
A_1 &= -\sqrt{T_2/T_1}\left[\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1} v_{1t}x_t'\right][E(x_t x_t')]^{-1}E(x_t) + o_p(1) \\
&\equiv A_{1,1} + o_p(1), \qquad \text{(A.12)}
\end{aligned}$$

where $A_{1,1} = -\sqrt{T_2/T_1}\left[\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1} v_{1t}x_t'\right][E(x_t x_t')]^{-1}E(x_t)$ is the leading term of A_1.

Note that $\Sigma_{1,2} = Acov(A_1, A_2) = Acov(A_{1,1}, A_2)$ because $A_{1,1}$ is the leading term of A_1. Define

$\eta_s = \Delta_{1s} - \Delta_1 + v_{1s}$. Then we have

$$\begin{aligned} Cov(A_1, A_2) &= Cov(A_{1,1}, A_2) + (s.o.) = E(A_{1,1}A_2) + (s.o.) \\ &= -\sqrt{T_2/T_1}[E(x_t x_t')]^{-1} \left\{ \frac{1}{\sqrt{T_1 T_2}} \sum_{t=1}^{T_1} \sum_{s=T_1+1}^{T} E\left[x_t' v_{1t} \eta_s\right] \right\} E(x_t) + (s.o.) \\ &= \sqrt{T_2/T_1} \frac{1}{\sqrt{T_1 T_2}} O(1) \\ &= O(T_1^{-1}) \\ &= o(1) \end{aligned} \tag{A.13}$$

because T_2/T_1 is bounded, $E(x_t x_t')$ is a finite positive definite matrix and $\sum_{t=1}^{T_1} \sum_{s=T_1+1}^{T} E(x_t' v_{1t} \eta_{1s}) = O(1)$ by Lemma A.1 given below. Since $\Sigma = \Sigma_1 + \Sigma_2$, the test statistic for testing H_0: $\Delta_1 = 0$ is given by

$$\frac{\hat{\Delta}_1}{\sqrt{(\hat{\Sigma}_1/T_2 + \hat{\Sigma}_2/T_2)}} \overset{H_0}{\to} N(0,1) \text{ in distribution,} \tag{A.14}$$

where $\hat{\Sigma}_1/T_2$ is given in (A.7) and $\hat{\Sigma}_2/T_2$ is given in (A.9).

Lemma A.1 *Under assumption 6, we have* $\sum_{t=1}^{T_1} \sum_{s=T_1+1}^{T} E\left[x_t' v_{1t} \eta_s\right] = O(1)$.

Proof: Define $d_t = 1$ if $t \geq T_1 + 1$ and $d_t = 0$ if $t \leq T_1$. For $t = 1, ..., T$, let $w_t = (x_t', v_{1t}, \Delta_{1t} d_t)'$ be a $d \times 2$ vector of a weakly stationary ρ-mixing process with the mixing coefficient $\rho(\tau)$ defined by

$$\rho(\tau) = \max_{1 \leq i,j \leq d} \frac{|Cov(w_{it}, w_{j,t+\tau})|}{\sqrt{Var(w_{it}) Var(w_{j,t+\tau})}},$$

where w_{it} is the i^{th} component of w_t, $i = 1, ..., d$. Assumption 6 implies that $\rho(\tau) \leq C_1 \lambda^{\tau}$ for some finite positive constants $C_1 > 0$ and $0 < \lambda < 1$. This requires that $cov(w_t, w_{t+\tau})$ decays at an exponential rate as τ increases. Many stationary ARMA processes are known to have an exponential decay rate (Carrasco and Chen, 2002). For expositional simplicity, we only consider the case that x_t is a scalar. For a vector x_t case, the following proof holds for each component of x_t. Hence, the proof holds true for a vector x_t case.

For $t \in \{1, ..., T_1\}$ and $s \in \{T_1 + 1, ..., T\}$, we have $|E(x_t v_{1t} \eta_s) - E(x_t v_{1t}) E(\eta_s)| \leq \rho(s -$

$t)\sqrt{Var(x_t v_{1t})Var(\eta_s)}$. By noting that $E(x_t v_{1t}) = 0$ and $\sqrt{var(x_t v_{1t})Var(\eta_s)} \leq C_2$ for some positive constant C_2, we have that $(s > t)$ $|E(x_t v_{1t}\eta_s)| \leq C\lambda^{s-t}$, where $C = C_1 C_2$. Using this we obtain

$$|E[\sum_{t=1}^{T_1}\sum_{s=T_1+1}^{T} x_t v_{1t}\eta_s]| \leq \sum_{t=1}^{T_1}\sum_{s=T_1+1}^{T} E[|x_t v_{1t}\eta_s|] \tag{A.15}$$

$$\begin{aligned}
&= \sum_{s=T_1+1}^{T} [E|x_1 v_{1,1}\eta_s| + E|x_2 v_{1,2}\eta_s| + ... + E|x_{T_1} v_{1,T_1}\eta_s|] \\
&\leq C\left[\left(\lambda^{T_1} + \lambda^{T_1+1} + ... + \lambda^{T-1}\right) + \left(\lambda^{T_1-1} + \lambda^{T_1} + ... + \lambda^{T-2}\right) + ... + \left(\lambda + \lambda^2 + ... + \lambda^{T_2}\right)\right] \\
&= C\left[\lambda + \lambda^2 + ... + \lambda^{T_2}\right]\left[1 + \lambda + \lambda^2 + ... + \lambda^{T_1-1}\right] \\
&= C\left(\frac{\lambda - \lambda^{T_2}}{1-\lambda}\right)\left(\frac{1-\lambda^{T_1}}{1-\lambda}\right) \\
&= O(1).
\end{aligned} \tag{A.16}$$

This completes the proof of Lemma A.1.

Appendix B: Derivation of $E(\epsilon_{1t}|\tilde{y}_t)$ for DGP2

For derivational/expositional simplicity, we only consider the case of $N = 2$. Hence, we have only one control unit y_{2t}. Our DGP2 is $f_t = \rho f_{t-1} + v_t$, where v_t is iid $N(0,1)$. $y_{jt} = a_j + b_j f_t + u_{jt} = f_t + u_{jt}$, where u_{jt} is iid uniform$[-c, c]$ for $j = 1, 2$, $\rho = 0.5$ and $c = 2$ in our simulations. Below we assume that $a_j = 0$ and $b_j = 1$ purely for derivational/expositional simplicity. The derivation for $a_j \neq 0$ and $b_j \neq 1$ is similar but much more tedious. We further assume that $f_0 = 0$. Then $f_t = v_t + \rho v_{t-1} + ... + \rho^t v_1 \sim N(0, a_t)$, where $a_t = 1 + \rho^2 + ... + \rho^{2t}$.

We assume that the times series data v_t, u_{1t} and u_{2t} are independent of each other. From the $j = 2$ equation we obtain $f_t = y_{2t} - u_{2t}$ and by substituting this into the $j = 1$ equation, we obtain

$$y_{1t} = y_{2t} - u_{2t} + u_{1t} \equiv y_{2t} + \epsilon_{1t}, \tag{B.1}$$

where $\epsilon_{1t} = u_{1t} - u_{2t}$. We want to derive the functional form of $E(\epsilon_{1t}|y_{2t})$ and show that it is nonlinear in y_{2t}. Note that $E(\epsilon_{1t}|y_{2t}) = E(u_{1t} - u_{2t}|y_{2t}) = -E(u_{2t}|y_{2t})$ because u_{1t} is independent of y_{2t}. Hence, we only need to derive the functional form for $E(u_{2t}|y_{2t})$. From

the definition of $E(u_{2t}|y_{2t})$, we have

$$E(u_{2t}|y_{2t}) = \int u f_{u_{2t}|y_{2t}}(u|y_{2t})du = \frac{\int u f_J(u, y_{2t})du}{f_y(y_{2t})} = \frac{\int u f_{y_{2t}|u_{2t}}(y_{2t}|u) f_u(u)du}{f_y(y_{2t})},$$

where $f_{u_{2t}|y_{2t}}(u|y_{2t})$ is the conditional density function of u_{2t} (conditional on y_{2t}) evaluated at $u_{2t} = u$ and $f_J(u, y_{2t})$ is the joint density of (u_{2t}, y_{2t}) evaluated at (u, y_{2t}). We first used $f_{u_{2t}|y_{2t}}(u|y_{2t}) = f_J(u, y_{2t})/f_y(y_{2t})$ and then we used $f_J(u, y_{2t}) = f_{y_{2t}|u_{2t}}(y_{2t}|u)f_u(u)$, where $f_y(\cdot)$ and $f_u(\cdot)$ are the marginal densities of y_{2t} and u_{2t}, respectively.

From $f_t \sim N(0, a_t)$ with $a_t = \sum_{s=0}^{t} \rho^{2s}$ and $y_{2t} = f_t + u_{2t}$, we know that $y_{2t}|(u_{2t} = u) \sim N(u, a_t)$. Hence, $f_{y_{2t}|u_{2t}}(y_{2t}|u) = \exp(-(y_{2t} - u)^2/(2a_t))/\sqrt{2\pi a_t}$. Also, note that $f_u(u) = \frac{1}{2c}\mathbf{1}(|u| \leq c)$. From these we can derive $\int u f(y_{2t}|u)f_u(u)du$ and $f_y(y_{2t})$.

We first consider $A \equiv \int u f_{y_{2t}|u_{2t}}(y_{2t}|u)f_u(u)du$. Using $f_u(u) = \frac{1}{2c}\mathbf{1}(|u| \leq c)$, we have

$$\begin{aligned} A &= \int_{-\infty}^{\infty} f_{y_{2t}|u_{2t}}(y_{2t}|u)f_u(u)udu \\ &= \frac{1}{2c}\int_{-c}^{c} f_{y_{2t}|u_{2t}}(y_{2t}|u)udu \\ &= \frac{1}{2c\sqrt{2\pi a_t}}\int_{-c}^{c} e^{-(y_{2t}-u)^2/(2a_t)}udu \\ &= \frac{1}{2c\sqrt{2\pi a_t}}\int_{-c}^{c} e^{-(u-y_{2t})^2/(2a_t)}udu \end{aligned} \tag{B.2}$$

because $(u - y_{2t})^2 = (y_{2t} - u)^2$.

Next, we write

$$\begin{aligned} udu &= (u - y_{2t} + y_{2t})du = (u - y_{2t})du + y_{2t}du \\ &= (1/2)d(u - y_{2t})^2 + y_{2t}du \\ &= a_t d[(u - y_{2t})^2/(2a_t)] + y_{2t}du. \end{aligned} \tag{B.3}$$

Substituting (B.3) into (B.2) we obtain

$$\begin{aligned} A &= \frac{1}{2c\sqrt{2\pi a_t}} \int_{-c}^{c} e^{-(u-y_{2t})^2/(2a_t)} \{a_t d[(u-y_{2t})^2/(2a_t)] + y_{2t} du\} \\ &= A_1 + A_2, \end{aligned} \tag{B.4}$$

where (letting $v = (y_{2t}-u)^2/(2a_t)$ below)

$$\begin{aligned} A_1 &= \frac{a_t}{2c\sqrt{2\pi a_t}} \int_{-c}^{c} e^{-(y_{2t}-u)^2/(2a_t)} d[(y_{2t}-u)^2/(2a_t)] \\ &= \frac{\sqrt{a_t}}{2c\sqrt{2\pi}} \int_{(c+y_{2t})^2/(2a_t)}^{(c-y_{2t})^2/(2a_t)} e^{-v} dv \\ &= \frac{\sqrt{a_t}}{2c\sqrt{2\pi}} [e^{-(c+y_{2t})^2/(2a_t)} - e^{-(c-y_{2t})^2/(2a_t)}], \end{aligned} \tag{B.5}$$

$$\begin{aligned} A_2 &= \frac{y_{2t}}{2c\sqrt{2\pi a_t}} \int_{-c}^{c} e^{-(u-y_{2t})^2/(2a_t)} du \\ &= \frac{y_{2t}}{2c\sqrt{2\pi a_t}} \sqrt{a_t} \int_{-c}^{c} e^{-[(u-y_{2t})/\sqrt{a_t}]^2/2} d[(u-y_{2t})/\sqrt{a_t}] \\ &= \frac{y_{2t}}{2c} \int_{(-c-y_{2t})/\sqrt{a_t}}^{(c-y_{2t})/\sqrt{a_t}} \frac{1}{\sqrt{2\pi}} e^{-v^2/2} dv \\ &= \frac{y_{2t}}{2c} \left[\Phi\left(\frac{c-y_{2t}}{\sqrt{a_t}} \right) - \Phi\left(\frac{-(c+y_{2t})}{\sqrt{a_t}} \right) \right], \end{aligned} \tag{B.6}$$

and where $\Phi(\cdot)$ is cdf of a standard normal random variable.

Finally, we need to consider $f(y_{2t})$. From $y_{2t} = f_t + u_{2t}$, $f_t \sim N(0, a_t)$, $u_{2t} \sim$ uniform$[-c, c]$ and the independence of f_t and u_{2t}, we know that (by the convolution formula)

$$\begin{aligned} f_{y_{2t}}(y_{2t}) &= \int_{-\infty}^{\infty} f_{f_t}(y_{2t}-u) f_u(u) du \\ &= \frac{1}{2c} \int_{-c}^{c} f_{f_t}(y_{2t}-u) du \\ &= \frac{1}{2c} \int_{-c}^{c} \frac{1}{\sqrt{2\pi a_t}} e^{-(u-y_{2t})^2/(2a_t)} du \\ &= \frac{1}{2c} \left[\Phi\left(\frac{c-y_{2t}}{\sqrt{a_t}} \right) - \Phi\left(\frac{-(c+y_{2t})}{\sqrt{a_t}} \right) \right]. \end{aligned} \tag{B.7}$$

Summarizing the above, we have shown that

$$E(u_{2t}|y_{2t}) = \frac{A_1 + A_2}{f_{y_{2t}}(y_{2t})} = \frac{\sqrt{a_t}}{\sqrt{2\pi}} \frac{[e^{-(y_{2t}+c)^2/(2a_t)} - e^{-(y_{2t}-c)^2/(2a_t)}]}{\left[\Phi\left(\frac{c-y_{2t}}{\sqrt{a_t}}\right) - \Phi\left(\frac{-(c+y_{2t})}{\sqrt{a_t}}\right)\right]} + y_{2t}, \qquad \text{(B.8)}$$

which is obviously nonlinear in y_{2t}.

CHAPTER 3 : Augmented Difference-in-Differences

3.1. Introduction

Answering important policy questions in economics and the management sciences often relies on our ability to evaluate causal effects of programs and interventions on outcomes of interest. In the most general terms, the fundamental problem of causal inference in quasi-experimental settings is the following: A researcher desires 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). Difference-in-differences (DID) is the standard, and most widely applied, econometric approach for measuring the average treatment effect (ATE) in panel data with pretreatment/posttreatment time periods and treatment/control units. An essential assumption is that the outcomes of the treated and non-treated units follow parallel paths over time, in the *absence* of any treatment. Violation of this 'parallel lines' assumption leads to biased DID estimates (Donald and Lang, 2007; Bertrand et al., 2004). In this paper we develop, derive, and implement a complementary practical and consistent DID estimator, the augmented DID, that is easy to apply and yields robust estimates of the ATE when the essential parallel lines assumption is violated.

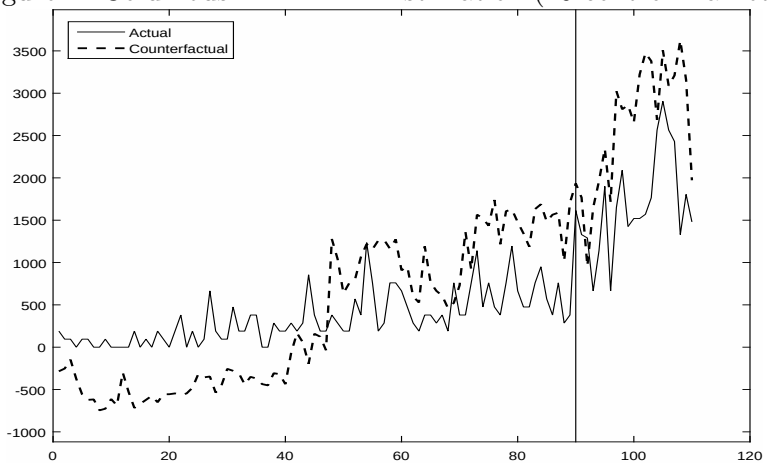
The essence of the augmented DID and our contribution can be better understood via the following motivating example. Consider the case of the 'digital first' eyewear brand Warby Parker, which began life as WarbyParker.com and has subsequently opened showrooms throughout the United States.[1] The rationale for showrooms is that since eyewear is a tactile product, some customers may wish to touch, feel, or try the product before buying. Naturally, management would like to assess whether the 'treatment', i.e., the opening of a showroom in a specific market, impacted overall sales relative to control markets which did not contain showrooms.

[1]We provide more institutional details in section 3.4, but for for now it suffices to note the following. Warby Parker is widely regarded as the exemplar Digitally Native Vertical Brand (DVNB)–a company that initially bypasses wholesale distribution and goes direct to consumers online–but subsequently opens offline sales channels. Other notable examples from the digital economy include DollarShaveClub.com (acquired in July 2016 by Unilever for $1 billion), Casper.com (mattresses), and Harrys.com (razors).

Unsurprisingly, management selected markets in which showrooms were expected to be demand-enhancing, i.e., accretive to overall sales. Management provided us with complete sales data for all markets including six with showrooms and we estimated the ATE in each case. Counterintuitively, for us and for management, in three of the six markets the DID ATE was negative, indicating that sales would have been greater had there not been a showroom opening. Prior research implies that under some conditions, an initial dip in *online* sales might occur, but an overall average decrease in total sales due to opening a showroom is wholly unexpected (Avery et al., 2012; Wang and Goldfarb, 2017). That is, a showroom not only failed to provide a lift in sales, but also was associated with a decrease in total online and offline sales compared to the counterfactual scenario of simply not opening a showroom at all.

How could this be? We provide complete details in Section 3.4, but for now, it suffices to focus on Columbus, Ohio, one market with a negative ATE for the showroom. In Figure 1, the solid line represents Columbus' weekly sales and the dashed line is Columbus predicted sales by the DID method based on the average weekly sales of the control cities. DID requires treatment and controls to follow parallel paths in the absence of treatment, and, as seen in Figure 1, this assumption is clearly violated. Therefore, this is an example of a misapplication of DID.

Figure 2: Columbus: DID ATE Estimation (10 control markets)



As a result, the DID ATE *underestimates* the effect of the treatment. Conversely, our augmented DID method provides a robust and consistent estimate of the ATE and shows that the intervention (showroom) actually *increased* total Columbus sales by around 75%. The method introduced by Hsiao et al. (2012) and extended by essay 1 is also applicable to scenarios where the parallel lines assumption fails to hold. Nevertheless, those innovations carry their own restriction; specifically, the number of control units needs to be much smaller than the number of pre-treatment time periods. In contrast, we show that the augmented DID provides a parsimonious solution which works well irrespective of whether the number of control units is small or large.

In practical management science applications, the most important question for the manager (or 'intervener') is often: "Did the intervention work and have the intended effect?" The answer is of paramount importance as it will dictate the ongoing strategy of the firm, e.g., to open more showrooms or not, and will drive the deployment of capital and business outcomes. The augmented DID estimate of the ATE therefore provides an answer to the question most usually valued by the manager (intervener).

In our application, the important substantive question is whether offline showrooms for online-first retailers, once opened, are demand enhancing. Nuanced versions of this question have been considered by Avery et al. (2012); Bell et al. (2017); Wang and Goldfarb (2017), among others. This general research question has gained prominence among management scientists for at least two reasons. First, in recent years, online retailing has outpaced the traditional retail sector and grown about 10% per annum in the United States (and even more in other markets, including BRIC countries) and now comprises about 8% of total U.S. retail sales. Second, customer behavior requires that new entrants in the online space also develop a physical presence through showrooms, pop-up shops, and even conventional stores.[2]

[2]See, for example, http://www.economist.com/news/business/21694545-why-some-firms-are-opening-shops-no-stock-shops-showrooms, "Shops to Showrooms: Why Some Firms are Opening Shops with no Stock," *The Economist*, May 12, 2016.

Methodologically, the augmented DID ATE is identified using the correlation between outcomes from the treated and control units. We impose no requirement that the sample paths of treated and non-treated units are parallel. Furthermore, we show in simulations and in real data that the augmented DID is robust to the selection of different control units. We elaborate further in Section 3.2.

Our contribution is twofold. First, we propose a new DID estimator that is robust to violation of the key 'parallel lines' assumption as well as to alternative selection schemes for non-treated units; moreover, it is easy to implement. Second, we prove that our estimator is consistent and provide asymptotic analysis to facilitate inference.[3]

Augmented DID is practically useful and easily implemented, yet our method works best under the following data conditions: a moderate or large pre-treatment and post-treatment sample size (larger than required for DID). The Augmented DID, like DID, can handle large number of treated and control units. Therefore, our augmented DID should be viewed as complementary to standard DID.

The remainder of the essay is organized as follows. In Section 3.2 we provide more background on DID and provide detailed estimation steps for conventional DID as well for our new approach. Section 3.3 reports simulation results to examine the finite sample performance of our estimator. In Section 3.4, we present an application. Section 3.5 concludes the essay with a discussion of how our method might best be deployed. Appendices A and B provide the relevant derivations, proofs, and theory for inference, as well as additional empirical results.

3.2. Estimation of ATEs

In this section, we discuss how to implement the conventional DID and our augmented DID as well as limitations of each method.

[3]Note that the inference theory in our paper is not covered by that of essay 1 who derived the asymptotic distribution of an average treatment effects estimator proposed by Hsiao et al. (2012) under the stationary data assumption. In this paper we explicitly allow for the existence of a non-stationary time trend component.

3.2.1. DID Mechanics and Implementation

Let y_{it}^1 and y_{it}^0 denote unit i's outcome in period t with and without treatment, respectively. The treatment intervention effect for the i^{th} observational unit at time t is defined as

$$\Delta_{it} = y_{it}^1 - y_{it}^0. \tag{3.2.1}$$

However, we can observe *either* y_{it}^0 or y_{it}^1, but never both. Thus, the observed data is in the form

$$y_{it} = d_{it}y_{it}^1 + (1 - d_{it})y_{it}^0, \tag{3.2.2}$$

where $d_{it} = 1$ if the i^{th} unit receives a treatment at time t, otherwise $d_{it} = 0$.

In our application, we compute the ATE for each treatment market separately. Therefore, we only need to consider the case where there is one treatment market (i.e., a market where the firm opens a showroom). We use y_{it} to denote our outcome variable of weekly sales (in dollars) of market i in week t. Without loss of generality, we assume that only the first market opened a showroom at time $T_1 + 1$, while the remaining N markets do not have any showroom throughout the sample data period. Therefore, for the treatment market, $y_{1t} = y_{1t}^0$ for $t = 1, ..., T_1$, and $y_{1t} = y_{1t}^1$ for $t \geq T_1 + 1$. We assume that there are N markets that do not have showrooms throughout our sample period. Hence, these N markets serve as the control group. We use y_{jt} for $j = 2, ..., N + 1$ and $t = 1, ..., T$ to denote control markets' weekly sales. We need to estimate y_{1t}^0 for $t \geq T_1 + 1$ in order to estimate the ATE. Let $\hat{y}_{1t}^0$ be a generic estimator of y_{1t}^0. Then the treatment effects at time t can be estimated by $\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^0$ $(t = T_1 + 1, ..., T)$ and the average treatment effects $\Delta_1 = E(y_{it}^1 - y_{it}^0)$ is estimated by

$$\hat{\Delta}_1 = \frac{1}{T_2} \sum_{t=T_1+1}^{T} \hat{\Delta}_{1t}, \tag{3.2.3}$$

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

Here we would like to emphasize that since there is only one unit (unit 1) that receives the treatment, the ATE estimator is obtained by averaging over the post-treatment periods $t = T_1 + 1, ..., T$ (time series averaging) for unit 1. This differs from the usual DID method in which one often has a larger number of units receiving treatments and the average is usually computed over many treatment units (cross sectional averaging). Of course, if we want to calculate the ATE over all treated units, we can first calculate ATE for individual treated units and then average over the treated units.

The difference in average outcomes after and before the treatment date for the treatment market (market 1) can be computed by

$$D_{treatment} = \frac{1}{T_2} \sum_{t=T_1+1}^{T} y_{1t} - \frac{1}{T_1} \sum_{t=1}^{T_1} y_{1t}. \tag{3.2.4}$$

The difference in outcomes for the N control markets after and before T_1 is computed by

$$D_{control} = \frac{1}{N} \sum_{j=2}^{N+1} \left[\frac{1}{T_2} \sum_{t=T_1+1}^{T} y_{jt} - \frac{1}{T_1} \sum_{t=1}^{T_1} y_{jt} \right]. \tag{3.2.5}$$

The conventional difference-in-differences estimate for the average treatment effects is:

$$\begin{aligned} & ATE_{1,DID} = D_{treatment} - D_{control} \\ = & \left(\frac{1}{T_2} \sum_{t=T_1+1}^{T} y_{1t} - \frac{1}{T_1} \sum_{t=1}^{T_1} y_{1t} \right) - \left(\frac{1}{N} \sum_{j=2}^{N+1} \left[\frac{1}{T_2} \sum_{t=T_1+1}^{T} y_{jt} - \frac{1}{T_1} \sum_{t=1}^{T_1} y_{jt} \right] \right). \end{aligned} \tag{3.2.6}$$

Under the 'parallel lines' assumption, it is easy to see that $ATE_{1,DID}$ defined in (3.2.6) is a consistent estimator of the ATE for the treated unit.

It is also possible to use a regression method to estimate ATE. Define the treatment group dummy and the post-treatment time period dummy as follows: $TG_i = 1$ if unit i is a treatment market, and 0 otherwise (we have $TG_1 = 1$ and $TG_j = 0$ for $j = 2, ..., N$), and

$AT_t = 1$ if $t \geq T_1 + 1$ and $AT_t = 0$ otherwise. Then the ATE estimator shown in (3.2.6) is identical to the least squares estimator of β_4 in the following regression model

$$y_{it} = \beta_1 + \beta_2 TG_i + \beta_3 AT_t + \beta_4 (TG_i)(AT_t) + u_{it} \qquad i = 1, ..., N;\ t = 1, ..., T. \quad (3.2.7)$$

To see that β_4 indeed yields the same ATE estimate as in (3.2.6), we obtain from (3.2.7) that

$$\begin{aligned} ATE_{1,DID} &= D_{treatment} - D_{control} \\ &= [(\beta_1 + \beta_2 + \beta_3 + \beta_4) - (\beta_1 + \beta_2)] - [(\beta_1 + \beta_3) - \beta_1] \\ &= \beta_4. \end{aligned} \quad (3.2.8)$$

The intuition behind the conventional DID method is that, if y_{jt}, $j = 1, ..., N+1$, are random draws from a homogenous population, then $\bar{y}_{c,t} = N^{-1} \sum_{j=2}^{N+1} y_{jt}$ may mimic $E(y_{1t})$ well in the absence of treatment. In order to improve the fit of using $\bar{y}_{c,t}$ to approximate y_{1t}, we add an intercept term δ_1 to $\bar{y}_{c,t}$ and use $\delta_1 + \bar{y}_{c,t}$ to approximate y_{1t}. We estimate δ_1 using the pre-treatment data by

$$\hat{\delta}_1 = \bar{y}_1 - \bar{y}_{control} = \frac{1}{T_1} \sum_{t=1}^{T_1} y_{1t} - \frac{1}{T_1} \sum_{t=1}^{T_1} \frac{1}{N} \sum_{j=2}^{N+1} y_{jt}, \quad (3.2.9)$$

where $\hat{\delta}_1$ is the least squares estimator of δ_1 in $y_{1t} - \bar{y}_{c,t} = \delta_1 + error_t$. Therefore, the DID in-sample-fit and the out-of-sample counterfactual estimate is computed by

$$\hat{y}^0_{DID,1t} = \hat{\delta}_1 + \frac{1}{N} \sum_{j=2}^{N+1} y_{jt}, \qquad t = 1, ..., T_1, T_1 + 1, ..., T \quad (3.2.10)$$

where $\hat{\delta}_1$ is given in (3.2.9). For $t = 1, ..., T_1$, (3.2.10) gives the in-sample fitted curve; for $t = T_1 + 1, ..., T$, (3.2.10) gives the out-of-sample counterfactual estimated curve.

To verify that (3.2.10) indeed gives the correct counterfactual estimate of y_{1t}^0, using (3.2.3) and (3.2.10) we obtain that

$$\begin{aligned}
\hat{\Delta}_1 &= \frac{1}{T_2}\sum_{t=T_1+1}^{T}[y_{1t} - \hat{y}_{DID,1t}^0] \\
&= \frac{1}{T_2}\sum_{t=T_1+1}^{T}[y_{1t} - \hat{\delta}_1 - \frac{1}{N}\sum_{j=2}^{N+1} y_{jt}] \\
&= \frac{1}{T_2}\sum_{t=T_1+1}^{T} y_{1t} - \frac{1}{T_1}\sum_{t=1}^{T_1} y_{1t} - \frac{1}{N}\sum_{j=2}^{N+1}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} y_{jt} - \frac{1}{T_1}\sum_{t=1}^{T_1} y_{jt}\right],
\end{aligned} \tag{3.2.11}$$

which identically equals $ATE_{1,DID}$, defined in (3.2.6), as it should. This verifies that (3.2.10) is the correct formula for predicting the the counterfactual outcome y_{1t}^0 for $t = T_1 + 1, ..., T$.

3.2.2. Factor Model Motivation

Similar to Hsiao et al. (2012) and essay 1, we motivate our method using a factor model. The main idea is that there are some common factors that drive all units although we allow for the common factors to affect different units in different ways. For example, in our application, common factors that affect Warby Parker's sales (outcome) in markets could include media coverage of the company, national advertising, and general economic conditions. However, a given factor may have a greater effect in some markets versus others. In the model, this is taken care of by allowing the coefficients of each factor to vary by market.

Following prior research (e.g. Forni and Reichlin (1998); Gregory and Head (1999); Hsiao et al. (2012)), the factor model for pre-treatment period is:

$$y_{it}^0 = \alpha_i + b_i' f_t + u_{it}, \qquad i = 1, ..., N; \quad t = 1, ..., T_1, \tag{3.2.12}$$

where α_i is unit i's individual specific intercept, b_i is a $K \times 1$ vector of coefficients for the

factors (factor loadings), f_t is a $K \times 1$ vector of unobservable factors common to treatment and control units and u_{it} is the error term.

Thus, the treatment unit's outcome, y_{1t}, and the average control units' outcomes, $\bar{y}_{c,t}$, are correlated through these common factors. The correlation between the treatment unit, y_{1t}, and the control units, $\bar{y}_{c,t}$, is what we are exploiting to create the counterfactual in the post-treatment period (what the outcome in treatment unit would have been had there not been an intervention). This is because we assume that had there not been an intervention, the correlation structure between the treatment and control units would remain the same as in the pretreatment period. In fact, this is our identification assumption. In the next section we show that under this assumption we can consistently estimate the counterfactual outcome for the treated unit; after creating the counterfactual, we can then estimate the ATE for the intervention.

3.2.3. Augmented DID method

The conventional DID method elaborated above rests on the assumption that the sample paths of y_{1t} and $\bar{y}_{c,t} = N^{-1} \sum_{j=2}^{N+1} y_{jt}$ are parallel in the absence of treatment. However, when there is heterogeneity in treatment and control groups, this assumption is unlikely to hold in practice. We propose an augmented DID method which is robust to non-parallel paths of the treated and the control units. We derive an estimator to address the question of interest to the practitioner: "Was the intervention a success?" (e.g., did demand go up, costs decline, and so on). We are able to derive an estimator that is consistent and delivers valid inference.

To accomplish this, we introduce two modifications to the conventional DID method. First, we add a time trend regressor to ameliorate the estimation bias coming from the non-parallel (linear) path problem. This results in the following regression model

$$y_{1t} - \bar{y}_{c,t} = \delta_1 + \delta_3 t + e_{1t}, \qquad t = 1, ..., T_1. \tag{3.2.13}$$

Let $\tilde{\delta}_1$ and $\tilde{\delta}_3$ denote the least squares estimates of δ_1 and δ_3 based on (3.2.13). Then we can estimate the counterfactual y_{1t}^0 using $\tilde{\delta}_1 + \tilde{\delta}_3 t + \bar{y}_{c,t}$ for $t = T_1 + 1, ..., T$. Although model (3.2.13) can improve the fit significantly over the conventional DID method, it may still fit data poorly because (i) it only adjusts for a linear trend but outcome variables often co-move in a nonlinear pattern; (ii) the variation of the treatment unit's outcome can differ greatly from that of the average of control units' outcomes. To overcome these problems, we introduce a second modification. We multiply $\bar{y}_{c,t}$ by a constant (δ_2), which leads to the following regression model:

$$y_{1t} = \delta_1 + \delta_2 \bar{y}_{c,t} + \delta_3 t + e_{1t}, \qquad t = 1, ..., T_1. \tag{3.2.14}$$

Let $(\hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3)$ denote the least squares estimator of $(\delta_1, \delta_2, \delta_3)$. Then, we estimate y_{1t}^0 by

$$\hat{y}_{1t}^0 = \hat{\delta}_1 + \hat{\delta}_2 \bar{y}_{c,t} + \hat{\delta}_3 t, \qquad t = 1, ..., T_1, T_1 + 1, ...T, \tag{3.2.15}$$

where for $t \leq T_1$, $\hat{y}_{1t}^0$ is the in-sample fitted value of y_{1t}^0; and for $t \geq T_1 + 1$, $\hat{y}_{1t}^0$ is the out-of-sample estimator for the counterfactual outcome y_{1t}^0. Therefore, using our augmented DID method the ATE estimate is given by

$$\hat{\Delta}_1 = \frac{1}{T_2} \sum_{t=T_1+1}^{T} (y_{1t} - \hat{y}_{1t}^0), \tag{3.2.16}$$

where $\hat{y}_{1t}^0$ is defined in (3.2.15).

Note that the ATE estimator $\hat{\Delta}_1$ defined in (3.2.16) nests the conventional DID estimator as a special case. To see this, we would replace $\hat{\delta}_2$ by 1 and $\hat{\delta}_3$ by 0, and then the estimates δ_1 and $\hat{\delta}_1$ will be the same as defined in (3.2.9). It follows that (3.2.16) becomes identical to (3.2.11), the conventional DID estimator of Δ_1.

Here we give a heuristic argument showing that $\hat{\Delta}_1$ is indeed a consistent estimator of Δ_1. We only need to show that $T_2^{-1} \sum_{t=T_1+1}^{T} \hat{y}_{1t}^0$ consistently estimates $T_2^{-1} \sum_{t=T_1+1}^{T} y_{1t}^0$.

Notice that if the correlation between y_{1t} and $\bar{y}_{c,t}$ is stable in the absence of treatment (our identifying assumption), then y_{1t} would be generated by $y_{1t}^0 = \delta_1 + \delta_2 \bar{y}_{c,t} + \delta_3 t + e_{1t}$ in the absence of treatment for $t = T_1 + 1, ..., T$. Given that $\hat{\delta}_j$ is a consistent estimator of δ_j for $j = 1, 2, 3$ since T_1 is large, and that the average of e_{1t} over the post-treatment period is small (if T_2 is large), then the average of $\hat{y}_{1t}^0$ and the average of y_{1t}^0 (over the post-treatment period) are close to each other and become closer the greater T_1 and T_2 are. Hence, $\hat{\Delta}_1$ is a consistent estimator of Δ_1.

Introducing the multiplicative scale factor δ_2 is more important than adding a time trend regressor as the former can capture nonlinear co-movement between the treated unit's outcome and the average of control units' outcomes. We will further illustrate this point in Appendix B where we show that augmented DID ATE estimation results do not change much when we drop the time trend regressor in model (3.2.14). However, if we impose $\delta_2 = 1$ in model (3.2.14), the in-sample fit may deteriorate significantly and the estimated ATE can change substantially (See Table B.2 in Appendix B for details).

3.3. Consistency and Simulation Results

3.3.1. Consistency

Our ATE estimator is consistent. That is, the ATE estimated using our method converges to the average change in outcome due to an intervention as long as the pre-treatment and post-treatment time periods are large enough. The ATE answers the question of interest to the manager of whether or not the intervention worked. First, we present the model for the treated unit before and after the intervention and then we use linear projections to rewrite our model to aid the consistency proof.

Before the intervention, the outcome for the treated unit in the pretreatment period is given by

$$y_{1t}^0 = \delta_1 + \delta_2 \bar{y}_{c,t} + \delta_3 t + e_{it}, \qquad t = 1, ..., T_1, \tag{3.3.1}$$

After an intervention occurs at time $t = T_1 + 1$, the outcome for treated unit in the post-treatment period is given by

$$y_{1t}^1 = \delta_1 + \delta_2 \bar{y}_{c,t} + \delta_3 t + \Delta_{1t} + e_{it}, \qquad t = T_1 + 1, ..., T, \tag{3.3.2}$$

As previously discussed, we use a factor model as motivation and exploit the correlation between the treated unit's outcome (y_{1t}) and the average of the control units' outcomes ($\bar{y}_{c,t}$).

To show that we have consistency, we use linear projections. We project the error e_{1t} onto the linear space of the regressors $z_t = (1, \bar{y}_{c,t}, t)'$ and get the linear projection $L(e_{1t}|z_t) = \gamma_1 + \gamma_2 \bar{y}_{ct} + \gamma_3 t$. We can re-write equation (3.3.1) as $y_{1t}^0 = \beta_1 + \beta_2 \bar{y}_{ct} + \beta_3 t + \epsilon_{1t}$, where the relationship between new coefficients and old coefficients are given by $\beta_1 = \delta_1 + \gamma_1$, $\beta_2 = \delta_2 + \gamma_2$, $\beta_3 = \delta_3 + \gamma_3$, and the new error term $\epsilon_{1t} = e_{1t} - L(e_{1t}|z_t)$. Then we have $L(\epsilon_{1t}|z_t) = 0$ by definition, which is equivalent to $Cov(z_t, \epsilon_{1t}) = 0$. Therefore, the least squares method consistently estimates the coefficients $\beta = (\beta_1, \beta_2, \beta_3)'$.

To facilitate the exposition, we will slightly abuse notation and continue to use $\delta = (\delta_1, \delta_2, \delta_3)'$ instead of β when discussing our estimation model. That is, we will use model (3.3.1) and interpret δ as the linear projection coefficients, i.e., $L(y_{1t}^0|z_t) = \delta_1 + \delta_2 \bar{y}_{c,t} + \delta_3 t$. Hence, $L(e_{1t}|z_t) = 0$.

Because $\hat{\delta}_{OLS}$ is a consistent estimator of $\delta = (\delta_1, \delta_2, \delta_3)'$ for large T_1 (essay 1), consistency of the ATE follows if both T_1 and T_2 are large:

$$\begin{aligned}
\hat{\Delta}_1 &= \frac{1}{T_2} \sum_{s=T_1+1}^{T} (y_{1t} - \hat{y}_{1t}^0) \\
&= (\delta_1 - \hat{\delta}_1) + (\delta_2 - \hat{\delta}_2)' \frac{1}{T_2} \sum_{s=T_1+1}^{T} \bar{y}_{c,s} + (\delta_3 - \hat{\delta}_3) \frac{1}{T_2} \sum_{T_1+1}^{T} t + \frac{1}{T_2} \sum_{s=T_1+1}^{T} \Delta_{1t} + \frac{1}{T_2} \sum_{s=T_1+1}^{T} e_{1t} \\
&= \bar{\Delta}_1 + O_p(T_1^{-1/2} + T_2^{-1/2}) \to \Delta_1 \text{ in probability}
\end{aligned}$$

for large T_1 and T_2, where $\bar{\Delta}_1 = T_2^{-1} \sum_{t=T_1+1}^{T} \Delta_{1t}$ and $\Delta_1 = E(\Delta_{1t})$.

In Appendix A we provide derivations for establishing the asymptotic normal distribution of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$. Those results can be used for inferences such as testing the null hypothesis of no significant average treatment effect, i.e., testing $\Delta_1 = 0$ against $\Delta_1 \neq 0$ or $\Delta_1 > 0$.

3.3.2. Simulations

In this section, we show that our augmented DID performs better than conventional DID and that this improvement is more pronounced when there is greater heterogeneity in the data generating process (DGP) for the outcomes in treated and control units. To do this, we examine the case where there is no heterogeneity, i.e., treated and control units follow the same DGP, and compare it to cases where there is heterogeneity, i.e., treated and control units follow different DGPs. We compute and compare the ATE predictive mean squared error for the three different DGPs.

For expositional simplicity, we conduct simulations using a simple factor model as given below. In the absence of treatment, the outcome in unit j at time t is given by:

$$y_{jt}^0 = c_{0,j} + c_{1,j}t + \lambda_j f_t + \eta_{jt}, \qquad j = 1, ..., N+1, t = 1, ...T, \tag{3.3.3}$$

where η_{jt} is a zero mean random variable with finite fourth moments, i.e., $E(\eta_{jt}^4)$ is finite. The factor follows an AR(1) process: $f_t = 0.8f_{t-1} + v_{1t}$ with v_t iid $N(0,1)$. As before, we assume without loss of generality that the first unit ($j = 1$) is the treatment unit and remaining N units ($j = 2, ..., N+1$) are the control units.

We first simulate the case DGP1 where there is no heterogeneity in treatment and control units. In this case, the treatment unit's outcome (y_{1t}^0) has the same distribution as control units' outcomes (y_{jt}^0 for $j = 2, ..., N+1$) so that the 'parallel sample paths' assumption is satisfied. In contrast, in DGP2 and DGP3, we introduce heterogeneity in the treatment and control units and consequently the 'parallel sample paths' assumption is violated. There are two ways to introduce this heterogeneity: we can vary the factor loadings (as we do in DGP2) or we can vary the coefficient on the time trend (as we do in DPG3). In DGP2, the

factor loadings of the treated unit and control units are draws from different distributions and in DGP3, the treated unit and control units have different trend components. Table 3 provides the parameter values for the three DGPs. We conduct simulations for three sets of sample sizes $T_1 \in \{25, 50, 100\}$ ($T_2 = 25$) and for 10 and 30 control units.

Table 3: Parameter Values

	$N = 11$				
	Time trend coefficient		Factor Loading		
	$c_{1,1}$	Control $\{c_{1,j}\}_{j=2}^{11}$	λ_1	Control $\{\lambda_j\}_{j=2}^{11}$	
DGP1	0.25	0.25	Unif$[1,2]$	Unif$[1,2]$	
DGP2	0.25	0.25	Unif$[0,1]$	Unif$[1,2]$	
DGP3	0.25	0.30	Unif$[1,2]$	Unif$[1,2]$	
	$N = 31$				
	Time trend coefficient			Factor Loading		
	$c_{1,1}$	Control $\{c_{1,j}\}_{j=2}^{16}$	Control $\{c_{1,j}\}_{j=17}^{31}$	λ_1	Control $\{\lambda_j\}_{j=2}^{16}$	Control $\{\lambda_j\}_{j=17}^{31}$
DGP1	0.25	0.25	0.25	Unif$[1,2]$	Unif$[1,2]$	Unif[1,2]
DGP2	0.25	0.25	0.25	Unif$[0,1]$	Unif$[1,2]$	Unif[2,3]
DGP3	0.25	0.30	0.35	Unif$[1,2]$	Unif$[1,2]$	Unif[1,2]

In all three DGP cases, we assume that the first unit receives a treatment Δ_{1t} at time $t = T_1 + 1$ of the form

$$\Delta_{1t} = \frac{exp(w_t)}{1 + \exp(w_t)} + 1 \qquad \text{for } t = T_1 + 1, ..., T,$$

where $w_t = 0.5w_{t-1} + \epsilon_t$ and ϵ_t is iid $N(0, 0.25)$. Hence, the outcome for the treated unit is $y_{1t}^1 = y_{1t}^0 + \Delta_{1t}$ for the post-treatment periods $t = T_1 + 1, ..., T$.

In all cases, we compute the post-treatment (out-of-sample) predictive mean squared error:

$$PMSE(\hat{\Delta}_1) = \frac{1}{M} \sum_{j=1}^{M} \left[\hat{\Delta}_{1,j} - \bar{\Delta}_{1,j} \right]^2,$$

where $M = 10,000$ is the number of simulation replications, the subscripts j denotes the estimation result for the j^{th} replication, $\hat{\Delta}_{1,j} = T_2^{-1} \sum_{t=T_1+1}^{T} \hat{\Delta}_{1t,j}$ is the estimated ATE, $\Delta_{1t,j} = y_{1t,j} - \hat{y}_{1t,j}^0$ is the true treatment effect and $\bar{\Delta}_{1,j} = T_2^{-1} \sum_{t=T_1+1}^{T} \Delta_{1t,j}$ is the true ATE for $j = 1, ..., M$. The results are reported in Tables 4 and 5.

We first examine the case of using $N-1=10$ control units (Table 4). In this instance, when there is no heterogeneity (DGP1), the 'the parallel sample paths' assumption is satisfied and both the conventional DID and our augmented DID method perform well. In addition, the PMSEs decrease as T_1 increases, which suggests that the ATE estimators are consistent. Next, we examine the cases (DGP2 and DGP3) where there is heterogeneity and thus the 'parallel sample paths' is violated. For DGP2 where the factor loadings of the treated unit and the control units are drawn from two different distributions, the conventional DID method exhibits a larger PMSE than our augmented DID method. For $T_1 = 25$, the PMSE ratio for the DID estimator to the augmented DID estimator is $1.544/0.7807 = 1.978$, and this ratio increases markedly to $1.185/0.1066 = 11.12$ for $T_1 = 100$. Finally, for DGP3 where the treated unit and the control units have different time trends, the conventional DID method suffers from large estimation bias and its performance deteriorates as T_1 increases.

This pattern of results is very similar to those in the Introduction and in the detailed empirical application which follows shortly. Specifically, when the treated unit and the average of control units have different trends, the conventional DID method can substantially over (or under) estimate counterfactual outcomes, which leads to large errors in estimating ATE. Note, however, that even under DGP3 our augmented DID estimator still performs very well. The simulation results show that our augmented DID estimator is robust to different types of 'non-parallel sample paths' and the that PMSE is remarkably similar for very different DGPs. More important, its estimated PMSE decreases as sample size increases, indicating consistency of our proposed augmented DID estimator.

Table 5 reports results for 30 control units instead of 10 control units; for all three DGP the results are similar to those in Table 4. Specifically, when the treated and control units are random draws from a common population (DGP1), both the conventional and our augmented methods perform well. However, conventional DID has lower PMSE than augmented DID because by imposing a correct restriction of equal weights on the control units, DID is more efficient. When the 'parallel sample paths' assumption is violated (DGP2 and

DGP3), the conventional DID method has large PMSEs. Moreover, since the heterogeneity is more pronounced for the 30 control unit case compared to the 10 control units case, the estimated PMSEs for standard DID using 30 control units are larger than PMSEs using 10 control units for DGP2 and DGP3. In contrast, the augmented DID method continues to do very well, which demonstrates two critical points. First, the robust performance of our augmented DID estimator shows that it is not sensitive to the degree of heterogeneity between the treated and the control units. Second, the fact that our augmented DID estimator does so well for both the 10 and 30 control units cases shows that it is robust the the selection of control units.

Table 4: PMSE ($N = 11$, $T_2 = 25$)

	DID			A-DID		
	$T_1 = 25$	$T_1 = 50$	$T_1 = 100$	$T_1 = 25$	$T_1 = 50$	$T_1 = 100$
DGP1	.2155	.1697	.1506	.8506	.2546	.1177
DGP2	1.544	1.305	1.185	.7807	.2346	.1066
DGP3	1.765	3.686	9.921	.8511	.2603	.1191

Table 5: PMSE ($N = 31$, $T_2 = 25$)

	DID			A-DID		
	$T_1 = 25$	$T_1 = 50$	$T_1 = 100$	$T_1 = 25$	$T_1 = 50$	$T_1 = 100$
DGP1	.2802	.1782	.1445	1.017	.2523	.1102
DGP2	2.625	1.598	1.195	.9991	.2455	.1069
DGP3	3.8673	8.039	22.10	1.011	.2521	.1073

3.4. Empirical Application

3.4.1. Institutional Setting and Data

WarbyParker.com is an online-first eyewear brand providing high quality eyeglasses at a lower price point ($95) than that typically encountered in the North American consumer market (upwards of $300). The data we analyze include all transactions during a 110-week period from February 2010 to March 2012 and the variables made available to us are: customer ID, customer ZIP code, item sold, and channel through which sales were made. Warby Parker operates three channels: online, a sampling channel called 'Home Try-On'

(HTO), and showrooms.[4] We aggregate data to the market-week level and the relevant dependent variable is total sales in dollars.

Our focus, of course, is whether the introduction of showrooms–locations in which customers can experience the entire product line and then purchase online–have any impact on total demand in those markets.

The first showroom opened in New York City in February 2010 and later that year showrooms opened in two other markets; six more markets got showrooms in 2011. Table 6 shows the dates when showrooms opened in nine U.S. markets. Of these markets, we examine the showroom treatment effects for those where the number of time periods in the pretreatment data (T_1) exceeds the number of time periods in the post treatment data (T_2), because a small pretreatment sample size can lead to large estimation errors. The six markets that opened showrooms after June 2011 satisfy this criterion, and in order of opening, are: Brooklyn, Boston, Austin, Los Angeles, Columbus, and Philadelphia. For the control group, we used the 10 largest markets by population without showrooms: Chicago, Houston, Portland, Seattle, Denver, Dallas, San Diego, Washington, Atlanta and Minneapolis.

Table 6: Showroom Opening Dates

Showroom Market	Opening Date
New York	2/15/2010
Oklahoma City	10/4/2010
San Francisco	11/9/2010
Brooklyn	7/27/2011
Boston	9/22/2011
Austin	10/6/2011
Los Angeles	11/1/2011
Columbus, OH	11/10/2011
Philadelphia	11/17/2011

We are interested in calculating the effect of opening showrooms on average weekly sales,

[4]In all three channels sales are fulfilled by shipping to a location of the customer's choosing. In the HTO channel customers select five frames (without lens) for a 5-day trial period, and then return them to the firm. HTO orders are said to convert to sales if an HTO customer buys product within two months of initiating the HTO. Showrooms are displays of the Warby Parker product line inside a third party retailer. In March 2013, Warby Parker opened its first company owned and operated stores.

aggregated across channels, in each of these markets (the ATE). Our strategy is to, one by one, examine each market that has a showroom as a unit that experiences a treatment, and then to ask how opening the showroom affects its weekly sales (the treatment effects).

3.4.2. Showroom ATEs for Individual Markets

We begin with the Columbus and Brooklyn markets, where the misapplication of the conventional DID method is especially pronounced. Specifically, the (average of) control units do a poor job of mimicking the treatment unit in the pre-treatment period. We believe that the conventional DID would underestimate the ATE for Columbus and overestimate it for Brooklyn and explain why. Next, we demonstrate that our augmented DID method successfully uses control units to mimic the treatment unit in the pre-treatment period. For completeness, we then briefly present results for the four remaining markets: Boston, Austin, Los Angeles, and Columbus.

Columbus

In the case of Columbus, we refer to Figure 2 sales data for Columbus in dollars (solid line) and the DID estimated curve (dashed line) that was presented in the Introduction. The vertical line shows when the showroom opened (at the 90^{th} week). We denote the dashed curve before the showroom opening week as the "in-sample fitted sales" and the part of the dashed line after the showroom opening as the "out-of-sample predicted counterfactual sales". The counterfactual sales is an estimate of what sales in Columbus *would have been* had there not been a showroom opening. In Section 3.2.1, we provided analytical expressions for the DID in-sample fitted curve and out-of-sample counterfactual predicted curve as shown in Figure 2, but for the moment focus on the intuition.

In Figure 2, from week 1 to week 46, the (average) of the control units with an intercept shift is below the sales in Columbus and from week 47 to week 90, it is above the sales in Columbus. In other words, the control units as used by the DID method do a poor job of mimicking the treatment unit in the pre-treatment period. greatly overestimates the out-

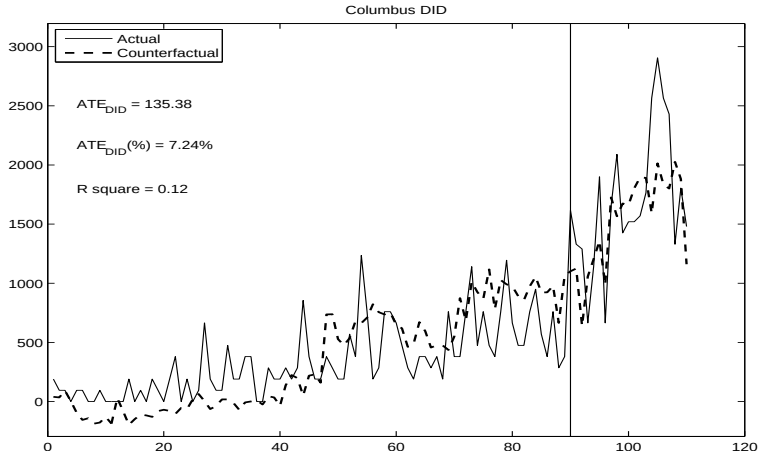
of-sample. As a result, the DID method overestimates the counterfactual and consequently, underestimates the average treatment effects and yields a large negative ATE. Management would therefore conclude (incorrectly), that the showroom depressed overall sales in the Columbus market.

The first problem with the conventional DID method when applied to Columbus's data is that the average of the ten control markets' sales and Columbus's sales exhibit very different upward trends; in other words, the sales of the treatment market (Columbus) and sales of the control group do not follow parallel paths in the absence of treatment, as required by DID. Consequently, the simple average of these control markets' sales does not accurately predict Columbus's counterfactual sales, which leads to an inaccurate estimate for the ATE.

The second problem (illustrated previously via our simulations), is that the DID ATE can be sensitive to the selection of control units, i.e., the choice of which units to include in the control group from among all potential units in the population of units that did not receive the treatment. To demonstrate this, we plot, in Figure 3, estimation results that use as the control group the thirty largest U.S. markets that do not have showrooms, rather than just ten markets as in Figure 2. The estimated ATE now becomes positive (whereas using 10 markets it was negative), but it is obvious that before the showroom opening week the fitted curve (dashed line) has a steeper upward trend than does the solid line of Columbus's actual sales. This suggests that the DID method still overestimates Columbus's counterfactual sales and consequently continues to underestimate the ATE. Hence, merely increasing the pool of non-treated units is insufficient to overcome the problem.

Fortunately, the augmented DID method overcomes both problems and continues to provide a consistent estimate of the ATE. As demonstrated analytically in Section 3.2 and via simulation in Section 3.3, we exploit the correlation between the treated and the control units' outcomes to consistently estimate the ATE. That is, we do not require that treatment and control outcomes follow parallel sample paths. Recall that to implement our augmented DID estimator, we: (1) add a time trend to the conventional DID model in order to allow

Figure 3: Columbus: DID ATE Estimation (30 control markets)



the outcomes of the treatment and the control units to have different linear trends, and (2) refrain from using a simple sample average of the control markets' sales to approximate the treatment market's sales sample path, and instead multiply the sample average of the control markets' sales by a scaling constant. This latter scaling constant is determined, along with the coefficient of the time trend variable, by the least squares method.[5]

Estimation results from applying augmented DID to Columbus's sales using 10 and 30 control markets are plotted in Figures 4 and 5, respectively. The first thing to note is that Figures 4 and 5 for augmented DID are in dramatic and sharp contrast to Figures 2 and 3 for conventional DID. For conventional DID the fitted curve (dashed line before showroom opening) is a highly inaccurate approximation of actual sales. By contrast, the fitted curve in Figures 4 and 5 mimics the trend of Columbus' sales before the showroom opening week very well. Consequently, the estimated out-of-sample predicted counterfactual sales are much more trustworthy than those obtained using conventional DID. Conventional DID would lead one to believe that weekly sales in the Columbus market *declined* by more than 25% due to the showroom, whereas augmented DID shows that they in fact *increased* by

[5]It turns out that, for our empirical application, introducing a multiplicative scale factor is much more important than introducing a time trend regressor. This implies that our WarbyParker.com online sales' data exhibits nonlinear trends that are taken care of by the multiplicative scale factor. In other words, it is not enough to simply add a time trend variable, which just accounts for linear trends. See Table B.2 in Appendix B.2 for our detailed estimation results on this point.

about 75%.

The second thing to note is that, remarkably, augmented DID produces *exactly the same* ATE lift estimate (75.4%, see Tables 7 and 8), when drawing on different markets with different trending behaviors as control groups leads. The only difference is that in using 30 markets instead of 10, the ATE is estimated more precisely and has a higher corresponding t-statistic. Hence, our augmented DID estimation results are robust to the selection of different control units in this case. Conversely, the inaccurate ATE from conventional DID swings from -26.5% (10 control markets) to 7.2% (30 control markets).

Figure 4: Columbus: A-DID ATE Estimation (10 control markets)

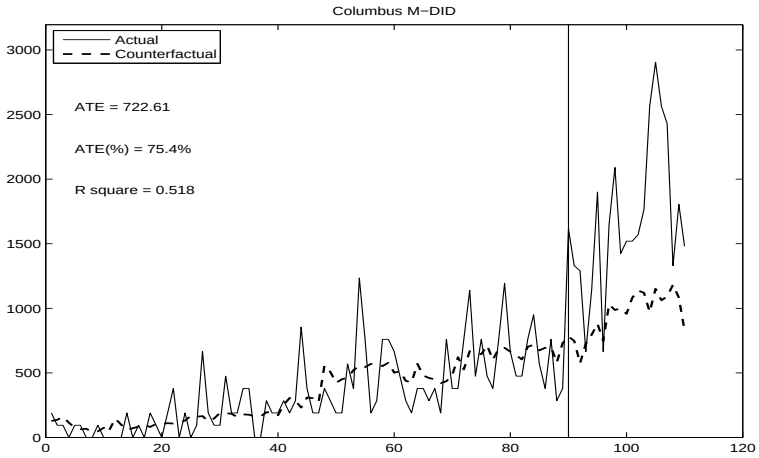
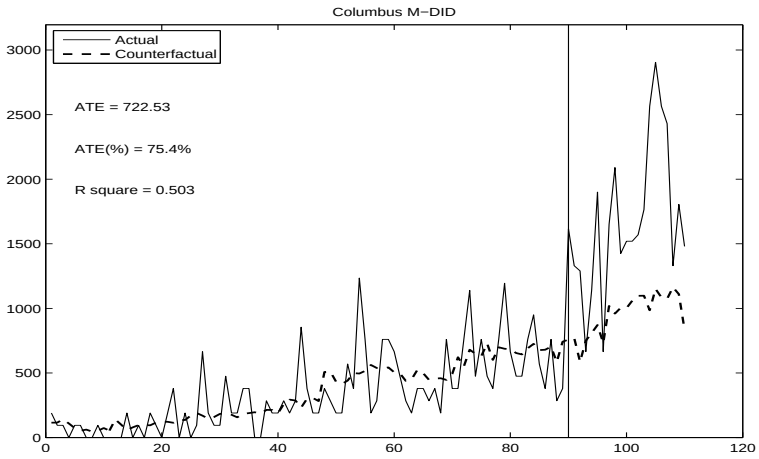


Figure 5: Columbus: A-DID ATE Estimation (30 control markets)



Brooklyn

Brooklyn offers an opposing case to that just presented for Columbus. Figures 6 and 7 plot the conventional DID estimation results, using 10 and 30 control markets, respectively. Again, the control units as used by DID method do not mimic the treatment well in the pre-treatment period. However, here the average sales of control markets has a much flatter slope than the sales for Brooklyn does. We expect that the conventional DID method would underestimate the counterfactual for Brooklyn and as a result, overestimate the effect of opening a showroom. Using the 30 control markets instead of 10 control markets does not help. On the contrary, it increases the estimation bias for Brooklyn.

Figure 6: Brooklyn: DID ATE Estimation (10 control markets)

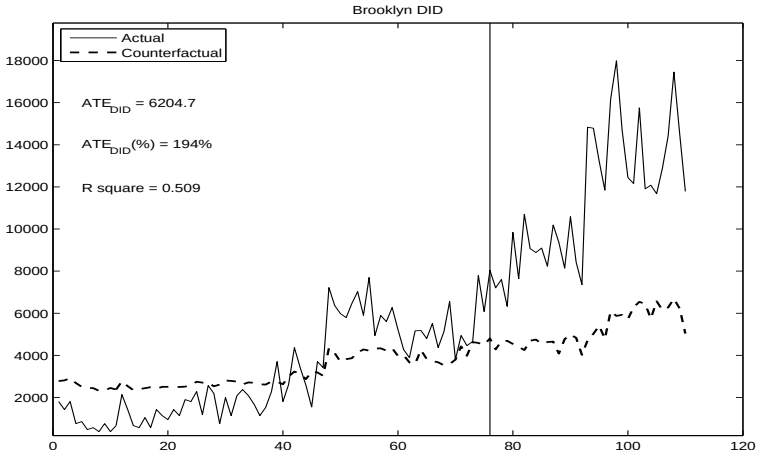


Figure 7: Brooklyn: DID ATE Estimation (30 control markets)

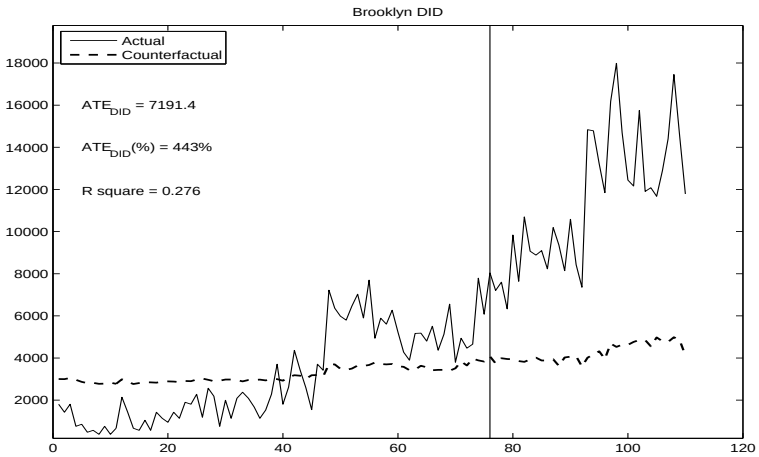


Figure 8: Brooklyn: A-DID ATE Estimation (10 control markets)

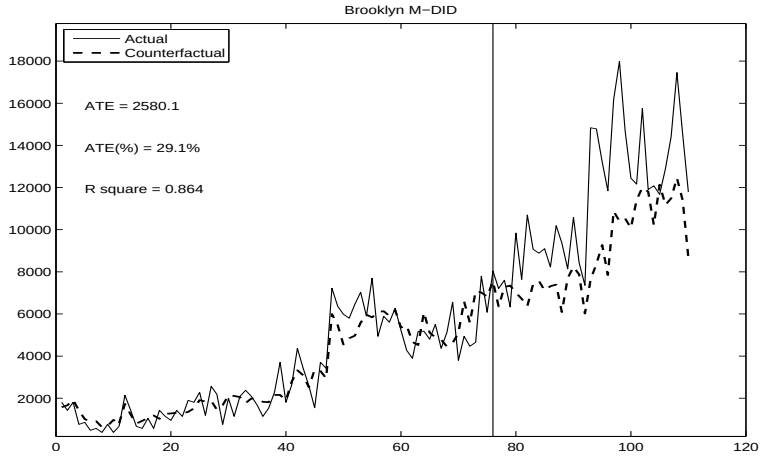
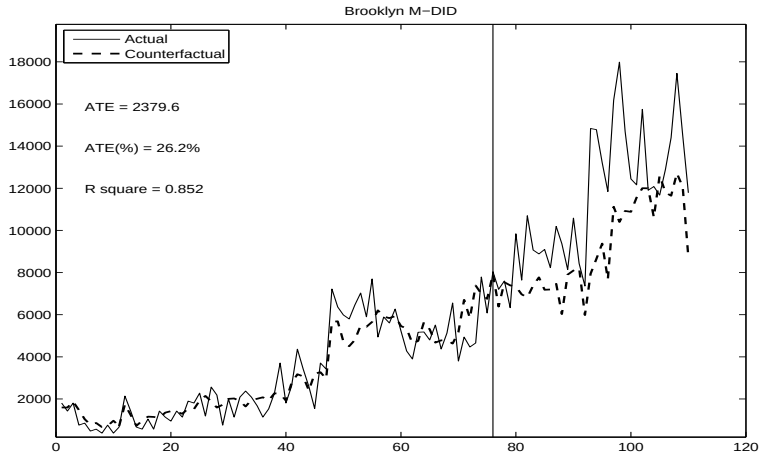


Figure 9: Brooklyn: A-DID ATE Estimation (30 control markets)



Figures 8 and 9 plot the estimated curves for augmented DID with 10 and 30 control markets, respectively. Similar to the case of Columbus, these figures show that our augmented DID method has better in-sample-fit (higher R-squared) than DID and sensible out-of-sample predictions. Once again, the estimation results are robust to the choice of markets in the control group and we find that opening a showroom in Brooklyn increased weekly sales by approximately 26 - 29%.

Remaining Markets: Boston, Austin, Los Angeles, Philadelphia

In the interests of brevity and ease of exposition, the figures for the remaining four markets are not presented here (see Appendix B). We do, however, provide estimates of the ATEs for all six markets and both methods using 10 (see Table 7) and 30 (Table 8) control markets, respectively. The tables provide the average weekly sales changes, percentage weekly changes after showroom openings, and R-squares from 24 models.

Estimation results for our augmented DID are very similar regardless of whether we use 10 or 30 control markets, whereas the conventional DID estimates are not only biased, but also fluctuate wildly. Augmented DID ATEs are significantly different from zero for Boston, Brooklyn, and Columbus ($p < .01$) and these three markets also exhibit the largest percentage increases (from about 30 to 75%). Austin and Los Angeles are significant at $p < .05$ (one sided test) using 30 control markets, and in Philadelphia the small percentage increases are not significantly different from zero.

Table 7: Augmented DID vs DID ATE results (10 Controls)

	A-DID			DID			
Market	ATE	% ATE	R^2	ATE	% ATE	T_1	T_2
Boston	935***	63.8	0.508	-303	-9.0	83	27
Brooklyn	2,580***	29.1	0.864	6205	194	83	27
Austin	832**	24.3	0.777	857	24.9	85	25
Columbus	723***	75.4	0.518	-973	-26.5	90	20
Los Angeles	1,337*	21.2	0.773	2917	79.6	90	20
Philadelphia	165	4.6	0.699	-173	-4.5	92	18

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

3.5. Conclusions

As noted in the Introduction, management and social scientists increasingly look to evaluate causal effects of interventions on outcomes of interest. In this pursuit of casual effects, the popularity of conventional DID methods and diversity of applications are due in large part to the widespread availability of quasi-experimental data, and the ease with which DID is implemented. While this method has much to commend it, it relies on the restrictive

Table 8: Augmented DID vs Conventional DID ATE results (30 Controls)

	A-DID			DID			
Market	ATE	% ATE	R^2	ATE	%ATE	T_1	T_2
Boston	906***	60.6	0.493	727	42.3	83	27
Brooklyn	2,380***	26.2	0.852	7191	443	83	27
Austin	744**	21.2	0.786	1903	109	85	25
Columbus	723***	75.4	0.503	135	7.2	90	20
Los Angeles	1,170**	18.1	0.775	4025	215	90	20
Philadelphia	79	2.1	0.694	994	51.1	92	18

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

'parallel lines' assumption; namely, that outcomes for treated and non-treated units follow parallel sample paths in the absence of any treatment.

The more varied the circumstances for treatment and control contexts, and the greater the pool from which control units can be drawn, the more likely it is that this critical assumption will not hold. These conditions are encountered by researchers in markets that are large and heterogeneous. Our empirical setting, the opening of physical showrooms in diverse locations throughout the United States by a digital-first brand, is case in point, as are those for the majority of articles referenced in the Introduction and throughout the essay. In this essay we propose a suitable method for dealing with this problem, irrespective of whether the number of control units is small or large. Using analytical results, simulations, and an empirical application, we show that our method is practically useful and has desirable theoretical properties. Specifically:

- *Practically Useful.* We proposed and implemented an augmented method that allows for treatment units and control units to be drawn from a heterogenous population, provided that the outcomes of the treatment and control units have a stable correlation relationship in the absence of treatment. In other words, the augmented DID method is robust to the selection of control units. In doing so, we have developed a practical method to address the question that is usually most important to the manager (or, implementer of the intervention): "Did the intervention have the intended effect?"

- *Theoretically Valid.* We proved, analytically, that our estimator is consistent and developed the asymptotic analysis necessary for valid inference. We then deployed simulated data and showed that the greater the heterogeneity in the data generating process, the larger the performance gap between conventional and augmented DID. The performance of conventional DID deteriorates rapidly, whereas that of augmented DID does not.

Finally, while our overall contribution is methodological, the results from our empirical application contribute to the emerging literature on online-offline market interaction (e.g., Forman et al. (2009); Anderson et al. (2010); Bell et al. (2017)). After studying offline showroom openings in six very different US markets by the digital first brand Warby Parker, we find that showrooms are demand accretive overall in the markets in which they are opened.

Appendix A: Inferences for the A-DID estimator

To prove the consistency of $\hat{\Delta}_1$ and derive the asymptotic distribution of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$, we assume that both T_1 and T_2 are large. We further assume that $T_2 \leq cT_1$ for some finite positive constant c. This means that either T_2 has the same magnitude as T_1, or T_2 has a smaller magnitude than T_1. Also, we assume that $\Delta_{1t} - E(\Delta_{1t})$ and e_{1t} are some weakly dependent processes so that laws of large numbers (LLN) and central limit theorems (CLT) apply to their (partial) sums.

We consider the case where y_{jt} is a sum of a trend component and a weakly dependent component in the absence of treatment: $y_{jt} = c_{0j} + c_{1j}t + \lambda_j' f_t + \eta_{jt}$, where f_t is a $r \times 1$ vector of common factor which may or may not be observable, λ_i is a $r \times 1$ vector of factor loadings (coefficients), f_t and η_{jt} are weakly dependent stationary processes (in t) so that LLN and CLT apply to partial sums over t. Note that $c_{0j} + \lambda_j' f_t + \eta_{jt}$ is a de-trended version of y_{jt}.[6] For the asymptotic analysis of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$, we use a de-trended version

[6] We can define $\tilde{\lambda}_j = (c_{1j}, \lambda_j')'$ and $\tilde{f}_t = (t, f_t')'$, then we have $y_{jt} = c_{0j} + \tilde{\lambda}_j' \tilde{f}_t + \eta_{jt}$. However, we explicitly separate stationary component f_t and the non-stationary component t for the convenience of asymptotic

of y_{jt} because, otherwise, the regressor t and y_{jt} will be asymptotically collinear.[7] We emphasize that de-trending is only needed for the theoretical analysis. In applications, whether one de-trends y_{jt} or not yields identical estimation result of $\hat{\Delta}_1$. This is because in finite sample applications, y_{jt} will not be collinear with the time trend regressor t, they do only asymptotically. So only in the asymptotic analysis, we de-trend y_{jt} for $j \geq 2$.

Using the pre-treatment period data $t = 1, ..., T_1$, we estimate the following regression model:

$$\begin{aligned}
y_{1t} &= a_1 + a_2 \bar{y}_{c,t} + a_3 t + e_{1t} \\
&= a_1 + a_2 \left(\frac{1}{N} \sum_{j=2}^{N+1} y_{jt} \right) + a_3 t + e_{1t} \\
&= a_1 + a_2 \left[\frac{1}{N} \sum_{j=2}^{N+1} (c_{0j} + c_{1j} t + \lambda_j' f_t + \eta_{jt}) \right] + a_3 t + e_{1t} \\
&= \left[a_1 + \frac{a_2}{N} \sum_{j=2}^{N+1} c_{0j} \right] + a_2 \left[\frac{1}{N} \sum_{j=2}^{N+1} (\lambda_j' f_t + \eta_{jt}) \right] + \left[a_3 + \frac{a_2}{N} \sum_{j=2}^{N+1} c_{1j} \right] t + e_{1t} \\
&= \delta_1 + \delta_2 \bar{\xi}_{c,t} + \delta_3 t + e_{1t} \\
&= z_t' \delta + e_{1t}, \qquad \text{(A.1)}
\end{aligned}$$

where $z_t = (1, \bar{\xi}_{c,t}, t)'$, $\bar{\xi}_{c,t} = N^{-1} \sum_{j=2}^{N+1} (\lambda_j' f_t + \eta_{jt})$, $\delta = (\delta_1, \delta_2, \delta_3)'$, $\delta_1 = a_1 + (a_2/N) \sum_{j=2}^{N+1} c_{0j}$, $\delta_2 = a_2$, $\delta_3 = a_3 + (a_2/N) \sum_{j=2}^{N+1} c_{1j}$.

For post-treatment period of the treated unit we have

$$\begin{aligned}
y_{1t} &= \delta_1 + \delta_2 \bar{\xi}_{c,t} + \delta_3 t + \Delta_{1t} + e_{1t} \\
&= z_t' \delta + \Delta_{1t} + e_{1t}, \qquad t = T_1 + 1, ..., T, \qquad \text{(A.2)}
\end{aligned}$$

where Δ_{1t} is the change in week t's sales (treatment effects) for the treatment market due

analysis.

[7] The reason for this is that $c_j t$ becomes the dominate component of y_{jt} when t is large, and $c_j t$ is collinear with the time-trend regressor t.

to the showroom opening event.

Substituting (A.2) and $\hat{y}_{1t}^0 = z_t'\hat{\delta}$ into (3.2.16), we obtain

$$\begin{aligned} \hat{\Delta}_1 - \Delta_1 &= \frac{1}{T_2}\sum_{t=T_1+1}^{T}[z_t'(\delta - \hat{\delta}) + v_{1t}] \\ &= A_1 + A_2, \end{aligned} \tag{A.3}$$

where $v_{1t} = e_{1t} + \Delta_{1t} - \Delta_1$, $\Delta_1 = E(\Delta_{1t})$, and

$$A_1 = \left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} z_t'\right](\delta - \hat{\delta}), \tag{A.4}$$

$$A_2 = \frac{1}{T_2}\sum_{t=T_1+1}^{T} v_{1t}. \tag{A.5}$$

Note that from $\delta_j - \hat{\delta}_j = O_p(T_1^{-1/2})$ for $j = 1, 2$ and $\delta_3 - \hat{\delta}_3 = O_p(T_1^{-3/2})$, one can easily show that $A_1 = O_p(T_1^{-1/2})$. Also, $A_2 = O_p(T_2^{-1/2})$ because v_{1t} is a zero mean weakly dependent process. Therefore, we see that for $\hat{\Delta}_1$ to be a consistent estimator of Δ_1, we need both T_1 and T_2 to be large. The large T_1 ensures the estimation error in $\hat{\delta} - \delta$ is small, while a large T_2 guarantees that the average of $v_{1t} = \Delta_{1t} - E(\Delta_{1t}) + e_{1t}$ over the post-treatment period is small.

Under the assumption that v_{1t} is a zero mean weakly dependence process, we show in Appendix B that

$$\sqrt{T_2}A_1 \xrightarrow{d} N(0, \Omega_1), \tag{A.6}$$

where $\Omega_1 = C_1 V_0 C_1'$, $C_1 = \sqrt{\alpha}(1, E(\bar{\xi}_{c,t}), 1 + \alpha/2)$, $\alpha = \lim_{T_1,T_2\to\infty} T_2/T_1 \leq c$, and V_0 is the

asymptotic variance of $D_{T_1}(\hat{\delta}-\delta)$, where $D_{T_1}=\begin{pmatrix}\sqrt{T_1} & 0 & 0\\ 0 & \sqrt{T_1} & 0\\ 0 & 0 & T_1^{3/2}\end{pmatrix}$, i.e.,

$$D_{T_1}(\hat{\delta}-\delta)=\begin{pmatrix}\sqrt{T_1}(\hat{\delta}_1-\delta_1)\\ \sqrt{T_1}(\hat{\delta}_2-\delta_2)\\ T_1^{3/2}(\hat{\delta}_3-\delta_3)\end{pmatrix}\xrightarrow{d}N(0,V_0), \tag{A.7}$$

See Chapter 16 in Hamilton (1994). Obviously, Ω_1 can be consistently estimated by

$$\hat{\Omega}_1=\hat{C}_1\hat{V}_0\hat{C}_1', \tag{A.8}$$

with $\hat{C}_1=\sqrt{T_2/T_1}(1,T_2^{-1}\sum_{t=T_1+1}^{T}\bar{\xi}_{c,t},1+T_2/(2T_1))$,

$$\hat{V}_0=D_{T_1}\hat{\Sigma}D_{T_1}, \tag{A.9}$$

where $\hat{\Sigma}$ is an estimator of $Var(\hat{\delta})$. For example, if the error e_{1t} is serially uncorrelated and conditional homoskedastic, we have $\hat{\Sigma}=\hat{\sigma}_e^2(Z'Z)^{-1}=\hat{\sigma}_e^2(\sum_{t=1}^{T_1}z_tz_t')^{-1}$, $\hat{\sigma}_e^2=(1/T_1)\sum_{t=1}^{T_1}\hat{e}_{1t}^2$ and $\hat{e}_{1t}=y_{1t}-z_t'\hat{\delta}$. If the error is conditional heteroskedastic, we can use conditional heteroskedastic robust estimator $\hat{\Sigma}=(Z'Z)^{-1}(Z'\hat{V}_eZ)(Z'Z)^{-1}$, $\hat{V}_e=diag(\hat{e}_{1t}^2)$ is a $T_1\times T_1$ diagonal matrix with the t^{th} diagonal element given by $\hat{e}_{1t}^2$ (White, 1984). If the error is serially correlated, we can use Newey and West (1987) type heteroskedasticity and serial correlation robust estimator to estimate Σ.

Assuming that Δ_{1t} and e_{1t} are weakly dependent stationary processes so that CLT apply to their partial sums, then we have

$$\sqrt{T_2}A_2=\frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T}v_{1t}\xrightarrow{d}N(0,\Omega_2), \tag{A.10}$$

where

$$\Omega_2 = \lim_{T_2 \to \infty} Var \left(T_2^{-1/2} \sum_{t=T_1+1}^{T} v_{1t} \right). \tag{A.11}$$

Note that when v_{1t} is serially uncorrelated, a consistent estimate of Ω_2 is given by

$$\hat{\Omega}_2 = \frac{1}{T_2} \sum_{t=T_1+1}^{T} (\hat{\Delta}_{1t} - \hat{\Delta}_1)^2. \tag{A.12}$$

If Δ_{1t} or e_{1t} is serially correlated. Then one can use a Newey-West type estimator:

$$\hat{\Omega}_2 = \frac{1}{T_2} \sum_{t=T_1+1}^{T} \sum_{s=T_1+1, |s-t| \leq l}^{T} (\hat{\Delta}_{1t} - \hat{\Delta}_1)(\hat{\Delta}_{1s} - \hat{\Delta}_1), \tag{A.13}$$

where $l \to \infty$ and $l/T_2 \to 0$ as $T_2 \to \infty$. For example, one may choose $l = O(T_2^{1/4})$ (Newey and West, 1987).

In Appendix B we show that $Cov(\sqrt{T_1}A_1, \sqrt{T_2}A_2)$ is negligible when T_1 and T_2 are both large. Summarizing the above, we have shown that

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) = \sqrt{T_2}(A_1 + A_2) \xrightarrow{d} N(0, \Omega), \tag{A.14}$$

where $\Omega = \Omega_1 + \Omega_2$, Ω_1 and Ω_2 are defined in (A.6) and (A.11), respectively. Ω can be consistently estimated by $\hat{\Omega} = \hat{\Omega}_1 + \hat{\Omega}_2$, where $\hat{\Omega}_1$ is defined in (A.8), $\hat{\Omega}_2$ is defined in (A.12) when e_{1t} and Δ_{1t} are serially uncorrelated, and by (A.13) when e_{1t} and Δ_{1t} are general weakly dependent processes.

The inference theory developed in (A.14) can be used to test the null hypothesis of zero treatment effects, i.e., we can test the null hypothesis H_0: $\Delta_1 = 0$ against H_1: $\Delta_1 \neq 0$ (or $\Delta_1 > 0$). The t-statistic is given by

$$\frac{\sqrt{T_2}\ \hat{\Delta}_1}{\sqrt{\hat{\Omega}}} \xrightarrow{d} N(0,1) \qquad \text{under } H_0.$$

Appendix B

B.1. Asymptotic analysis of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$

Multiplying (A.4) by $\sqrt{T_2}$ gives

$$\sqrt{T_2}A_1 \quad = \quad -B_1 D_{T_1}(\hat{\delta} - \delta), \tag{B.1}$$

where

$$\begin{aligned} B_1 &= \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} z_t' D_{T_1}^{-1} \\ &= \sqrt{T_2/T_1} \left(1, \frac{1}{T_2} \sum_{t=T_1+1}^{T} \bar{\xi}_{c,t}, \frac{1}{T_1 T_2} \sum_{t=T_2+1}^{T} t \right) \\ &\xrightarrow{p} \sqrt{\alpha}(1, E(\bar{\xi}_{c,t}), 1 + \alpha/2) \\ &\equiv C_1 \end{aligned} \tag{B.2}$$

where we used $\alpha = \lim_{T_1,T_2\to\infty} T_2/T_1$, $T_2^{-1}\sum_{t=T_1+1}^{T} \bar{\xi}_{c,t} \to E(\bar{\xi}_{c,t})$ (in probability) by a laws of large number argument, $\frac{1}{T_1 T_2}\sum_{t=T_1+1}^{T} t = (2 + T_2/T_1 + 1/T_1)/2 \to 1 + \alpha/2$. Note that $C_1 = \sqrt{\alpha}(1, E(\bar{\xi}_{c,t}), 1 + \alpha/2)$ is a 1×3 vector of constants (it is non-random).

Now, (A.6) follows from (B.1), (B.2) and (A.7), i.e.,

$$\sqrt{T_2}A_2 = \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} \xrightarrow{d} N(0, \Omega_2). \tag{B.3}$$

In lemma B.1 we show that $Cov(\sqrt{T_2}A_1, \sqrt{T_2}A_2) \to 0$ as T_1, T_2 get large. This result, together with (A.6) and (A.10), proves (A.14).

We now show that when Δ_{1t} and e_{1t} are serially uncorrelated, $\hat{\Omega}_2$ defined in (A.12) is a consistent estimator for Ω_2. To show that, let $\bar{w} = T_2^{-1}\sum_{t=T_1+1}^{T} w_t$ (for $w_t = z_t$, Δ_{1t} or

e_{1t}), we have

$$\begin{aligned}\hat{\Omega}_2 &= \frac{1}{T_2}\sum_{t=T_1+1}^{T}\left[(z_t-\bar{z})'(\delta-\hat{\delta})+\Delta_{1t}-\bar{\Delta}_1+e_{1t}-\bar{e}_1\right]^2 \\ &= \frac{1}{T_2}\sum_{t=T_1+1}^{T}E\left\{[\Delta_{1t}-E(\Delta_{1t})+e_{1t}]^2\right\}+O_p(T_1^{-1/2}+T_2^{-1/2}) \\ &\xrightarrow{p} E(v_{1t}^2)=\Omega_2, \end{aligned} \tag{B.4}$$

where we used $(z_t-\bar{z})'(\delta-\hat{\delta})=O_p(T_1^{-1/2})$, $\bar{\Delta}_1=\Delta_1+O_p(T_2^{-1/2})$ and $\bar{e}_1=O_p(T_2^{-1/2})$.

If Δ_{1t} and e_{1t} weakly dependent processes we know that $\hat{\Omega}_2$ defined in (A.13) is a consistent estimator of Ω_2 (Newey and West, 1987).

Before we prove that $Cov(\sqrt{T_2}A_1,\sqrt{T_2}A_2)=o(1)$, we first obtain a leading term of $\sqrt{T_2}A_1$. By inserting an identity matrix $I=D_{T_1}^{-1}D_{T_1}$ in the middle of (B.2) we obtain

$$\begin{aligned}\sqrt{T_2}A_1 &= -\frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T}z_t'D_{T_1}^{-1}D_{T_1}(\hat{\delta}-\delta) \\ &= -[C_1+o_p(1)]D_{T_1}(\hat{\delta}-\delta), \end{aligned} \tag{B.5}$$

where C_1 is defined in (B.2).

Let Z be the $T_1\times 3$ matrix with the t^{th} row given by $(1,\bar{\xi}_{c,t},t)$ and $e_1=(e_{11},...,e_{1T_1})'$. Then

$$\begin{aligned}D_{T_1}(\hat{\delta}-\delta) &= \left[D_{T_1}^{-1}ZZ'D_{T_1}^{-1}\right]^{-1}D_{T_1}Ze_1 \\ &= [M+o_p(1)]^{-1}D_{T_1}Ze_t, \end{aligned} \tag{B.6}$$

where

$$M=\begin{pmatrix}1 & E(\bar{\xi}_{c,t}) & (1/2)\\ E(\bar{\xi}_{c,t}) & E(\bar{\xi}_{c,t}^2) & E(\bar{\xi}_{c,t})/2\\ 1/2 & E(\bar{\xi}_{c,t})/2 & 1/6\end{pmatrix}.$$

Because M^{-1} is finite, in order to show that $Cov(\sqrt{T_2}A_1, \sqrt{T_2}A_2) = o(1)$, it sufficient to show that $Cov(\sqrt{T_2}A_{1,2}, \sqrt{T_2}A_2) = o(1)$, where

$$\sqrt{T_2}A_{1,2} = D_{T_1}Ze_1 = \begin{pmatrix} \frac{1}{\sqrt{T_2}}\sum_{t=1}^{T_1} z_{1t}e_{1t} \\ \frac{1}{T_1^{3/2}}\sum_{t=1}^{T_1} te_{1t} \end{pmatrix} \equiv \begin{pmatrix} A_{1,2,1} \\ A_{1,2,2} \end{pmatrix} \tag{B.7}$$

with $z_{1t} = (1, \bar{\xi}_{c,t})'$. Hence, we only need to show that $Cov(\sqrt{T_2}A_{1,2,1}, \sqrt{T_2}A_2) = o(1)$ and $Cov(\sqrt{T_2}A_{1,2,2}, \sqrt{T_2}A_2) = o(1)$.

We introduce some notation and a definition of a weakly dependent process. For $t = 1, ..., T$, let $\zeta_t = e_{1t}$ or $e_{1t}\bar{\xi}_{c,t}$, we assume that ζ_t and v_{1t} are strictly stationary process satisfying

$$\frac{|Cov(\zeta_t, v_{1,t+\tau})|}{\sqrt{Var(\zeta_t)Var(v_{1,t+\tau})}} \leq C\gamma^{\tau} \tag{B.8}$$

for all $1 \leq t < t + \tau \leq T$, for some finite positive constants C, $0 < \gamma < 1$. In the econometrics/statistical literature, ζ_t and v_{1t} satisfying (B.8) are termed as ρ-mixing processes. Equation (B.8) implies that the correlation coefficient between ζ_t and $v_{1,t+\tau}$ decays to zero at an exponential rate. Many weakly dependent processes, including some strictly stationary ARMA processes, are known to be ρ-mixing processes with exponential decay rates (Carrasco and Chen, 2002).

Lemma B.1 *Assume that ζ_t and v_{1t} ($\zeta_t = e_{1t}$ or $e_{1t}\bar{\xi}_{c,t}$) satisfy (B.8). Then*

$Cov(\sqrt{T_2}A_{1,2,1}, \sqrt{T_2}A_2) = o(1)$ *and* $Cov(\sqrt{T_2}A_{1,2,2}, \sqrt{T_2}A_2) = o(1)$,

where $A_{1,2,1}$ and $A_{1,2,2}$ are defined in (B.7) and A_2 is defined in (A.5), respectively.

Proof: Using (B.8) we have

$$
\begin{aligned}
& |Cov(\sqrt{T_2}A_{1,2,1}, \sqrt{T_2}A_2)| = \frac{1}{\sqrt{T_1}}|\sum_{t=1}^{T_1}\sum_{s=T_1+1}^{T} E(z_{1t}e_{1t}v_{1s})| \\
\leq & \frac{1}{\sqrt{T_1}}\sum_{s=T_1+1}^{T}[\,|E(z_{11}e_{11}v_{1s})| + |E(z_{12}e_{12}v_{1s})| + ... + |E(z_{1T_1}e_{1T_1}v_{1s})|\,] \\
\leq & \frac{C}{\sqrt{T_1}}\left[(\gamma^{T_1} + \gamma^{T_1+1} + ... + \gamma^{T-1}) + (\gamma^{T_1-1} + \gamma^{T_1} + ... + \gamma^{T-2}) + ... + (\gamma + \gamma^2 + ... + \gamma^{T_2})\right] \\
= & \frac{C}{\sqrt{T_1}}[\gamma + \gamma^2 + ... + \gamma^{T_2}][1 + \gamma + \gamma^2 + ... + \gamma^{T_1-1}] \\
= & \frac{C}{\sqrt{T_1}}\left(\frac{\gamma - \gamma^{T_2+1}}{1-\gamma}\right)\left(\frac{1-\gamma^{T_1}}{1-\gamma}\right) \\
= & O\left(\frac{1}{\sqrt{T_1}}\right) \to 0
\end{aligned} \tag{B.9}
$$

as $T_1 \to \infty$.

Similarly, we have

$$
\begin{aligned}
& |Cov(\sqrt{T_2}A_{1,2,2}, \sqrt{T_2}A_2)| = \frac{1}{T_1^{3/2}}|\sum_{t=1}^{T_1}\sum_{s=T_1+1}^{T} tE(e_{1t}v_{1s})| \\
\leq & \frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\sum_{s=T_1+1}^{T}|E(e_{1t}v_{1s})| \\
\leq & \frac{C}{\sqrt{T_1}}\left(\frac{\gamma - \gamma^{T_2+1}}{1-\gamma}\right)\left(\frac{1-\gamma^{T_1}}{1-\gamma}\right) \\
= & O\left(\frac{1}{\sqrt{T_1}}\right) \to 0
\end{aligned} \tag{B.10}
$$

as $T_1 \to \infty$.

B.2. Empirical results using only scale factor or using only trend

In this appendix, we report ATE empirical estimation results using the following two simple models:

$$
y_{1t} = \delta_1 + \delta_2 \bar{y}_{c,t} + e_{1t}, \qquad t = 1, ..., T_1, \tag{B.11}
$$

and

$$y_{1t} - \bar{y}_{c,t} = \delta_1 + \delta_3 t + e_{1t}, \qquad t = 1, ..., T_1. \tag{B.12}$$

Model (B.11) simply modifies the conventional DID estimator by multiplying the average control units' outcome by a scale factor δ_2. Compared with our augmented DID estimation equation defined in (3.2.14), model (B.11) does not have the additional time trend regressor. Model (B.12) has the time trend regressor, but it imposes $\delta_2 = 1$ in model (3.2.14). Both model (B.11) and (B.12) are special cases of model (3.2.14) and both have two parameters to be estimated. In Table B.2, we report estimation results (B.11) and (B.12) using 10 control markets. We also report estimation results for using models (3.2.14) for comparison convenience.

From Table B.2, we observe (i) that the R^2 for model (B.11) are larger than model (B.12) for all six markets, hence, model (B.11) fits the data better than model (B.12) for all cases; (ii) the estimated ATEs (in %) of model (B.11) are quite close to those obtained using the model general model (3.2.14) for all markets, while for model (B.12), only Boston's estimate ATE is relatively close to that obtained using model (3.2.14). These results show that between the model modifications of introducing a multiplicative scale factor (to multiply the average control units' outcome) and by adding a time trend regressor, the former is more important than the latter. This is because the average control units outcome also has a (nonlinear) trend component, just that it may not be parallel to the (nonlinear) trend in the treated unit's outcome, the multiplicative scale factor helps to adjust the two sample paths to be parallel to each other in a nonlinear way. Although model (B.12) can catch the linear trends differences between the average control units' outcome and the treated unit's outcome, it fails to catch any nonlinear trend component differences between the two outcomes. Therefore, the multiplicative scale factor is more important modification than simply adding a time trend regressor to the conventional DID model.

Table 9: ATE results for models (3.2.14), (B.11) and (B.12)

	Model (3.2.14)		Model (B.11)		Model (B.12)	
Market	% ATE	R^2	% ATE	R^2	% ATE	R^2
Boston	63.8	0.508	64.9	0.494	63.5	0.492
Brooklyn	29.1	0.864	29.8	0.857	40.9	0.766
Austin	24.3	0.777	23.7	0.774	30.9	0.759
Columbus	75.4	0.518	72.2	0.513	41.4	0.397
Los Angeles	21.2	0.773	19.1	0.767	33.0	0.736
Philadelphia	4.6	0.699	5.5	0.699	14.1	0.665

B.3. Additional empirical estimation results

Figures 10 to 17 plot estimated curves using our augmented DID method for Austin, Boston, Los Angeles and Philadelphia with 10 and 30 control markets, respectively. The estimation results using 10 and using 30 control markets are quite similar. This once again confirms that our augmented DID method is robust to the selection of control markets because our method does not require that the sales for a treatment market and the average sales for control markets follow parallel paths in the absence of treatment.

Figure 10: Austin: A-DID ATE Estimation (10 control markets)

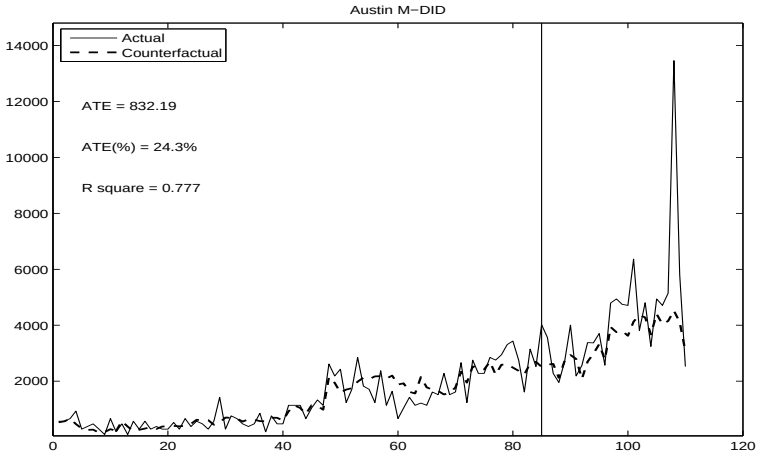


Figure 11: Austin: A-DID ATE Estimation (30 control markets)

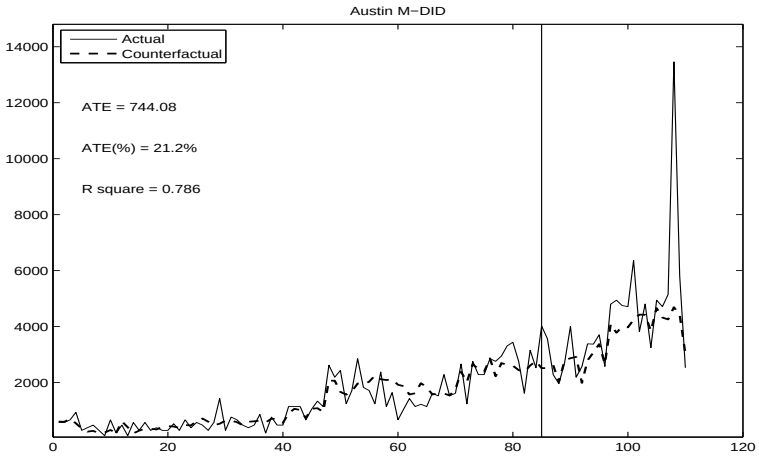


Figure 12: Boston: A-DID ATE Estimation (10 control markets)

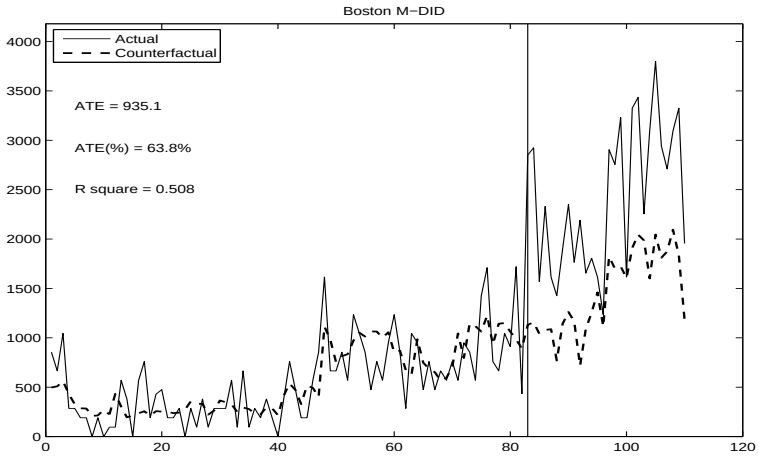


Figure 13: Boston: A-DID ATE Estimation (30 control markets)

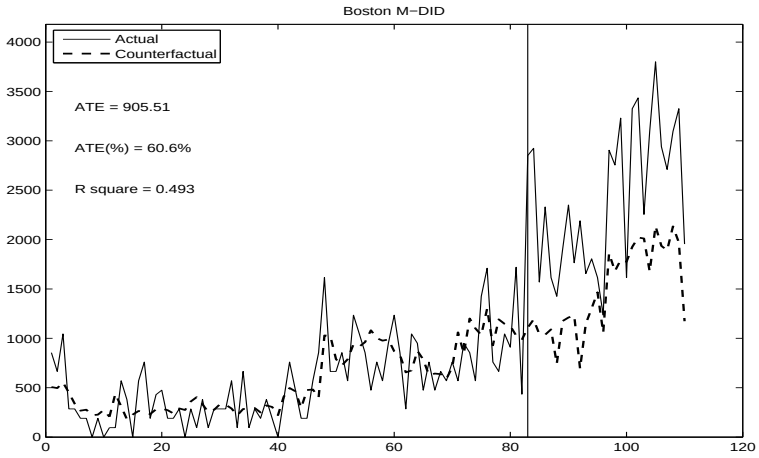


Figure 14: Los Angeles: A-DID ATE Estimation (10 control markets)

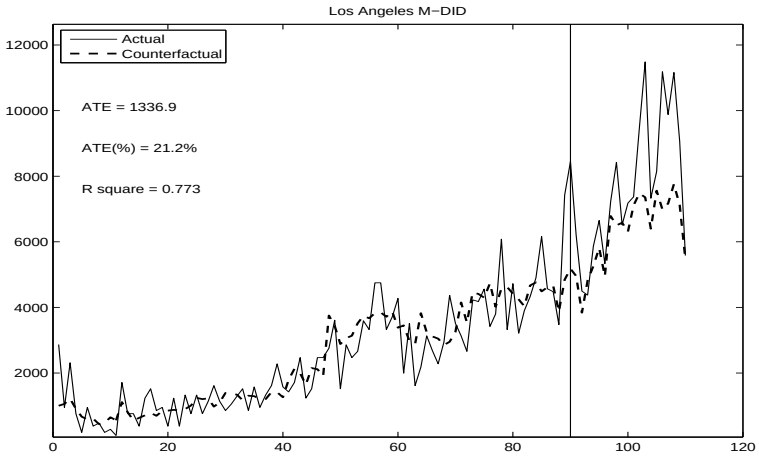


Figure 15: Los Angeles: A-DID ATE Estimation (30 control markets)

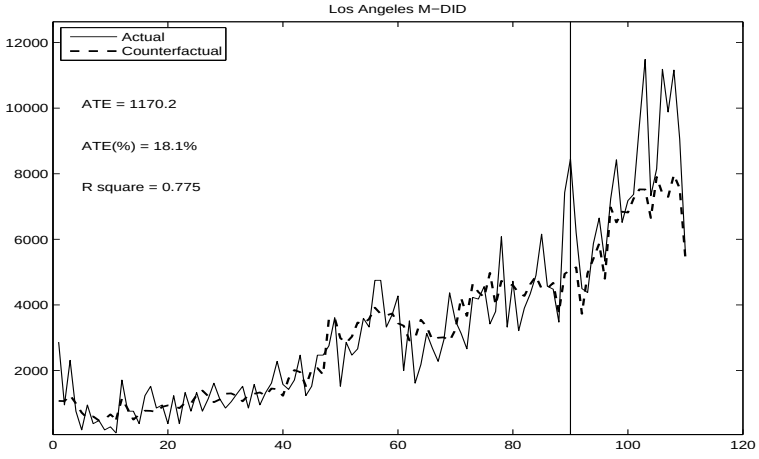


Figure 16: Philadelphia: A-DID ATE Estimation (10 control markets)

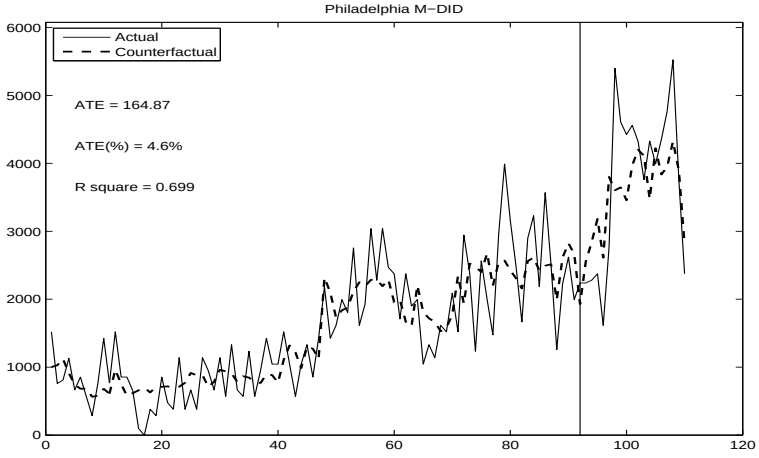
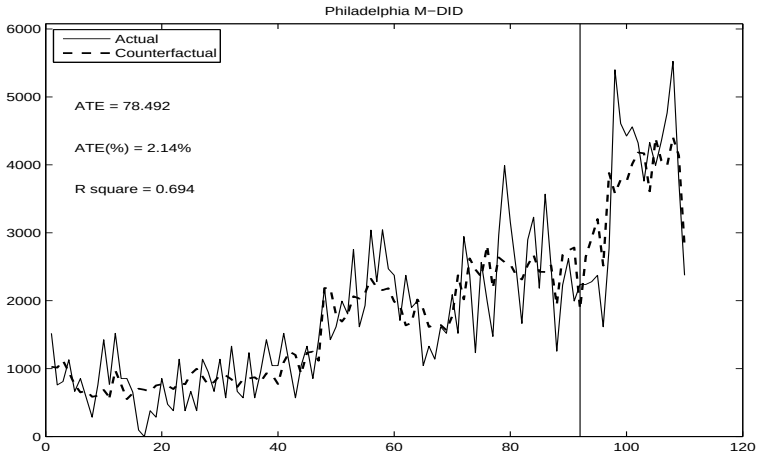


Figure 17: Philadelphia: A-DID ATE Estimation (30 control markets)



CHAPTER 4 : Statistical Inference for the Synthetic Control Method

4.1. 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. 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 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 paths in the absence of treatment. Violation of this parallel lines 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 method proposed by Abadie and Gardeazabal (2003), and Abadie et al. (2010) can be used successfully to estimate average treatment effects (ATE). This method has many attractive features: First, it is more general than the conventional difference-in-differences 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 synthetic control method restricts the weights assigned to the control group to be non-negative and may lead to better extrapolation. In fact, Athey and Imbens (2017) describe the synthetic control method as "arguably the most important innovation in the evaluation literature in the last 15 years".

To date, there has been no formal inference theory for the synthetic control and modified synthetic control ATE estimator with long panels under general conditions. Thus, the main contribution of this essay is to derive the asymptotic distribution of the synthetic control and modified synthetic control ATE estimators with long panels. We do this using projec-

tion theory and show that a properly designed subsampling method can be used to obtain confidence intervals and conduct inference whereas the standard bootstrap cannot. For inference, applications of the synthetic control method 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) show 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 (2015, 2016) propose rigorous inference methods for DID and synthetic control ATE estimators under different conditions. Conley and Taber (2011) assume 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 the random assignment to the treated unit). They show that one can conduct proper inference for the DID ATE estimator by using the control units' information. Their method allows for both the pre and the post-treatment periods to be small. Assuming instead that the pre-treatment period is large and the post-treatment period is small, Ferman and Pinto (2015, 2016) show that Andrew's 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 et al. (2017) recently proposed a general inference procedure for a number of different ATE estimators, including DID, synthetic control, 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 pre-treatment 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 pre-treatment sample. Therefore, for this type of data, inference methods based on small

post-treatment sample size will be invalid.

In this essay 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 use projection theory (Zarantonello, 1971; Fang and Santos, 2016) to derive the asymptotic distributions of the standard and the modified synthetic control ATE estimators with long panels. The asymptotic distributions are non-normal and non-standard. Moreover, it is known that the standard bootstrap does not work (Andrews, 2000; Fang and Santos, 2016). Yet, we show that a carefully designed subsampling method, i.e., applying the subsampling method only to part of the statistic, provides valid inferences. We also apply our new theoretical results to conducting inferences for empirical data. We estimate the effect of an e-tailer's showroom opening on its average weekly sales. For this data, we have $T_1 = 90$ and $T_2 = 20$, where T_1 and T_2 are pre- and post-treatment sample sizes, respectively. Using simulations for this T_1, T_2 combination, the inference based on our proposed subsampling method yields more accurate estimated confidence intervals than the estimates using Andrews' (2003) instability test. The reason is that $T_2 = 20$ is not negligible compared to $T_1 = 90$, rendering Andrews' (2003) test improper for our empirical data.

We make three contributions. First, and most importantly, we derive the formal inference theory for the synthetic control and modified synthetic control method ATE estimator under long panels. The asymptotic distribution is non-normal and non-standard, and standard bootstrapping breaks down. Second, we propose our easy-to-implement subsampling procedure and show that it leads to valid inferences. Finally, we provide a simple sufficient condition under which the synthetic and modified synthetic control estimator is uniquely determined, and we show via simulations and an empirical example that a modified synthetic control method, which is robust to 'non-parallel paths' situations, can greatly enhance the applicability of the synthetic control 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 et al. (2017)).

4.2. 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 i^{th} 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 i^{th} 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 heterogenous distributions (i.e., they 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 ATE (over post-treatment period) for each treated unit separately. In this way, one can obtain ATE for each treated unit. If one also wants to obtain ATE over all the treated units, one can average (possibly with different weights) over all treated units. Hence, in this essay 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: $\Delta_1 = E(\Delta_{1t})$, where $\Delta_{1t} = y_{1t}^1 - y_{1t}^0$. 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.

4.2.1. 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. In order to use unified notation to cover

both the synthetic control and the modified synthetic control methods, we add an intercept to the standard synthetic control method. Therefore, utilizing the correlation between y_{1t} and y_{jt} where $j = 2, ..., N$, one can estimate the synthetic control counterfactual outcome y_{1t}^0 based on the following regression model:

$$y_{1t} = x_t'\beta_0 + u_{1t}, \qquad t = 1, ..., T_1, \tag{4.2.1}$$

where $x_t = (1, y_{2t}, ..., y_{Nt})'$[1] is an $N \times 1$ vector of the control units' outcome variables, $\beta_0 = (\beta_{0,1}, ..., \beta_{0,N})'$ 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.

Abadie and Gardeazabal (2003) and Abadie et al. (2010) propose a synthetic control 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 pre-treatment data, and the weights are non-negative and sum to one. Specifically, one selects $\beta = (\beta_1, ..., \beta_N)'$ via the following constrained minimization problem:

$$\hat{\beta}_{T_1,Syn} = \arg \min_{\beta \in \Lambda_{Syn}} \sum_{t=1}^{T_1} \left[y_{1t} - x_t'\beta \right]^2, \tag{4.2.2}$$

where $\Lambda_{Syn} = \{\beta \in \mathcal{R}^N : \ \beta_j \geq 0 \text{ for } j = 2, ..., N \text{ and } \sum_{j=2}^N \beta_j = 1\}$. With $\hat{\beta}_{T_1,Syn}$ defined as the minimizer to (4.2.2), the synthetic control fitted/predicted curve is

$$\hat{y}_{1t,Syn}^0 = x_t'\hat{\beta}_{T_1,Syn}, \qquad t = 1, ..., T_1, T_1 + 1, ..., T. \tag{4.2.3}$$

Note that $\hat{y}_{1t,Syn}^0$ is the in-sample fitted curve for $t = 1, ..., T_1$, and $\hat{y}_{1t,Syn}^0$ gives the predicted counterfactual outcome of y_{1t}^0 for $t = T_1 + 1, ..., T$. The ATE is estimated by $\hat{\Delta}_{1,Syn} = T_2^{-1} \sum_{t=T_1+1}^{T} (y_{1t} - \hat{y}_{1t,Syn}^0)$.

[1] In order to use unified notation to cover both the synthetic control and the modified synthetic control methods, we add an intercept to the standard synthetic control method.

When the number of control units is larger than the number of pre-treatment time periods, an unique weight vector β that minimizes (4.2.2) may not exist. In such cases, it is necessary to regulate the weights such as imposing non-negativity and sum to one restrictions. The rationale for imposing non-negativity restriction is that in most applications, y_{jt}'s are positively correlated with each other, and therefore they tend to move up or down together. The add-to-one restriction $\sum_{j=2}^{N} \beta_j = 1$ introduced by Abadie et al. (2010) implicitly assumes that a weighted average outcomes for the control units and the treated unit's outcome would have followed parallel paths 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 lines" 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).

Since our main interest is to forecast y_{1t}^0 for $t \geq T_1+1$ rather than in-sample-fit, as long as T_1 is moderately large, we recommend using $N < T_1$ control units in estimating $\hat{y}_{1t}^0$. There are at least two reasons for doing this. The first is than when treated and control outcomes are generated by a fixed number of common factors, using a finite number of control units gives more accurate predicted counterfactual outcomes than using a large number of control units. The reason is that using too many regressors in a forecasting model leads to large prediction variance. The second resason to use $N > T_1$ is that $\hat{\beta}_{T_1,Syn}$ cannot be uniquely determined in general. In practice when one faces a large number of control units, one can use AIC, BIC, LASSO (Efron et al., 2004), or the best subset selection method proposed by Doudchenko and Imbens (2016) to select significant control units. Abadie et al. (2010) also suggest 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 essay, we consider the case without any relevant covariates and discuss how to add relevant covariances in the empirical application in Section 4.6.

4.2.2. The modified synthetic control method

For many quasi-experimental data used in economics, 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 sum to one) of the control units' outcomes may not follow parallel paths in the absent 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 a standard synthetic control model, i.e., we still keep the non-negative constraints: $\beta_j \geq 0$ for $j = 2, ..., N$ 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 add-to-one restriction. Therefore, the modified synthetic control method is the same as (4.2.2) except that the add-to-one restriction on the slope coefficients is removed, i.e., one solves the following (constrained) minimization problem:

$$\hat{\beta}_{T_1,Msyn} = \arg \min_{\beta \in \Lambda_{Msyn}} \sum_{t=1}^{T_1} \left[y_{1t} - x_t' \beta \right]^2, \tag{4.2.4}$$

where $x_t = (1, y_{2t}, ..., y_{Nt})'$, β is an $N \times 1$ vector of parameters, and $\Lambda_{Msyn} = \{\beta \in \mathcal{R}^N : \beta_j \geq 0 \text{ for } j = 2, ..., N\}$. Let X be the $T_1 \times N$ matrix with its t^{th} row given by $x_t' = (1, y_{2t}, ..., y_{Nt})$. We show in the Appendix B that when X has full column rank (which requires that $T_1 \geq N$), the synthetic control minimizers $\hat{\beta}_{T_1,Syn}$ and $\hat{\beta}_{T_1,Msyn}$ are uniquely defined. With $\hat{\beta}_{T_1,Msyn}$ defined in (4.2.4), the counterfactual outcome is estimated by $\hat{y}^0_{1t,Msyn} = x_t' \hat{\beta}_{T_1,Msyn}$ for $t = T_1 + 1, ..., T$, and the ATE is estimated by $\hat{\Delta}_{1,Msyn} = T_2^{-1} \sum_{t=T_1+1}^{T} \left[y_{1t} - \hat{y}^0_{1t,Msyn} \right]$.

4.3. Distribution Theory

4.3.1. A projection of the unconstrained estimator

To study the distribution theory of the synthetic control ATE estimator, we first show that one 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 synthetic control ATE estimator.

Let $\hat{\beta}_{OLS}$ denote the ordinary least squares estimator of β_0 using data $\{y_{1t}, x_t\}_{t=1}^{T_1}$. We show in 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 $\hat{\beta}_{OLS}$ onto the convex set Λ, where $\Lambda = \Lambda_{Syn}$ or $\Lambda = \Lambda_{Msyn}$.

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:

$$\Pi_{\Lambda,T_1}\theta = \arg\min_{\lambda\in\Lambda}(\theta-\lambda)'(X'X/T_1)(\theta-\lambda), \tag{4.3.1}$$

$$\Pi_{\Lambda}\theta = \arg\min_{\lambda\in\Lambda}(\theta-\lambda)'E(x_tx_t')(\theta-\lambda). \tag{4.3.2}$$

Here we use the notation Π_Λ to denote a projection onto the set Λ. Note that 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 non-random norm $\|a\|_E = \sqrt{a'E(x_tx_t')a}$, i.e., $\Pi_{\Lambda,T_1}\theta = \arg\min_{\lambda\in\Lambda}||\lambda-\theta||_X^2$ 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 $\hat{\beta}_{OLS}$, and the second projection relates the limiting distributions of $\sqrt{T_1}(\hat{\beta}_{T_1}-\beta_0)$ and $\sqrt{T_1}(\hat{\beta}_{OLS}-\beta_0)$.

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

$$\hat{\beta}_{T_1} = \arg\min_{\beta\in\Lambda}(\hat{\beta}_{OLS}-\beta)'(X'X/T_1)(\hat{\beta}_{OLS}-\beta) = \Pi_{\Lambda,T_1}\hat{\beta}_{OLS}. \tag{4.3.3}$$

Equation (4.3.3) says 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, then there is a simple closed form solution to the constrained minimization problem (4.3.3). Let $\hat{\beta}_{OLS}$ denote the least squares estimator of β. Then it is easy to see that the closed form solution is $\hat{\beta}_{T_1,j} = \hat{\beta}_{OLS,j}$ if $\hat{\beta}_{OLS,j} \geq 0$; and $\hat{\beta}_{T_1,j} = 0$ if $\hat{\beta}_{OLS,j} < 0$ for $j = 2, ..., N$, i.e., the projection simply keeps the positive component as is and maps the negative component to zero. However, when $X'X/T_1$ is not a diagonal matrix, a simple non-iterative closed form solution does not exist. Nevertheless, we show in Appendix B that when $X'X/T_1$ is positive definite, the objective function is globally convex and there is an unique solution to the constrained minimization problem.

To derive the asymptotic distribution of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ (hence, for $\hat{\Delta}_1$), we first examine the asymptotic (as $T_1 \to \infty$) range of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$. Note that even when both $\hat{\beta}_{T_1}$ and β_0 take values in the constrained set Λ, $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ does not necessarily take values in Λ. The range of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ depends on Λ as well as how many components of the β_0 vector take value 0, i.e., it depends on how many non-negativity constraints are binding. We illustrate this point via a simple example. We use $T_{\Lambda_{Syn},\beta_0}$ to denote the asymptotic range of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$.

4.3.2. An example of $T_{\Lambda_{Msyn},\beta_0}$

We consider an example to illustrate the asymptotic range, $T_{\Lambda_{Msyn},\beta_0}$, for the modified synthetic control method. For expositional simplicity, we consider a simple model with two control units and without an intercept: $y_{1t} = x_t'\beta_0 + u_{1t} = x_{1t}\beta_{0,1} + x_{2t}\beta_{0,2} + u_{1t}$ with $\beta_0 = (\beta_{0,1}, \beta_{0,2})' \in \Lambda_{Syn} = \mathcal{R}_+^2$. To characterize $T_{\Lambda_{Msyn},\beta_0}$ for $\beta_0 \in \mathcal{R}_+^2$, we consider four cases: (i) $\beta_0 = (0,0)'$; (ii) $\beta_0 = (0, \beta_{0,2})$ with $\beta_{0,2} > 0$; (iii) $\beta_0 = (\beta_{0,1}, 0)$ with $\beta_{0,1} > 0$; and (iv) $\beta_0 = (\beta_{0,1}, \beta_{0,2})$ with $\beta_{0,j} > 0$ for $j = 1, 2$. For case (i) we have $\sqrt{T_1}(\hat{\beta}_{T_1,j} - \beta_{0,j}) = \sqrt{T_1}\hat{\beta}_{T_1,j} \in [0, +\infty)$ for $j = 1, 2$. Hence, $T_{\Lambda_{Msyn},\beta_{0,(i)}} = \mathcal{R}_+ \times \mathcal{R}_+$, which is the first quadrant. For case (ii), it is easy to see that $\sqrt{T_1}(\hat{\beta}_{T_1,1} - \beta_{0,1}) = \sqrt{T_1}\hat{\beta}_{T_1,1} \in [0, +\infty)$, and $\sqrt{T_1}(\hat{\beta}_{T_1,2} - \beta_{0,2}) \in \sqrt{T_1}[-\beta_{0,2}, +\infty) \to (-\infty, +\infty) = \mathcal{R}$ as $T_1 \to \infty$. Hence, $T_{\Lambda_{Msyn},\beta_{0,(ii)}} = \mathcal{R}_+ \times \mathcal{R}$, which is the union of the first and the fourth quadrants. Similarly,

it is easy to check that $T_{\Lambda_{Msyn},\beta_0,(iii)} = \mathcal{R} \times \mathcal{R}_+$, the union of the first and the second quadrants. Finally, for case (iv), because $\hat{\beta}_{T_1,j} - \beta_{0,j}$ can be either positive or negative for $j = 1, 2$, $\sqrt{T_1}(\hat{\beta}_{T_1,j} - \beta_{0,j}) \to \mathcal{R}$ as $T_1 \to \infty$. Hence, $T_{\Lambda_{Msyn},\beta_0,(iv)} = \mathcal{R} \times \mathcal{R}$, the whole plane.

Remark 4.3.1 *Through the above example one can see that T_{Λ,β_0} gives the asymptotic range of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$. Hence, it is quite intuitive to expect that the limiting distribution of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ can be represented as a projection of the limiting distribution of $\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0)$ onto T_{Λ,β_0}.*

We show in the next subsection that the intuition stated in remark 4.3.1 is indeed correct.

4.3.3. The asymptotic theory: the stationary data case

We refer to the set T_{Λ,β_0} as the asymptotic range of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ based on intuitive argument. In the projection theory, the set T_{Λ,β_0} is referred to as the tangent cone of Λ at β_0. We give a formal definition of a tangent cone as well as some explanation of the term 'tangent' in Appendix A.

Theorem 4.3.2 *Let Z_1 denote the limiting distribution of $\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0)$. Then under the assumptions 1 to 4 presented in Appendix B, we have*

$$\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) \xrightarrow{d} \Pi_{T_{\Lambda,\beta_0}} Z_1. \tag{4.3.4}$$

Note that Theorem 4.3.2 states that the limiting distribution of the constrained estimator can be represented as a projection of the unconstrained (least squares) estimator onto the tangent cone T_{Λ,β_0}. With the help of Theorem 4.3.2, we derive the asymptotic distribution of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$ as follows.

Theorem 4.3.3 *Under the same conditions as in Theorem 4.3.2, we have*

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) \xrightarrow{d} -\phi E(x_t')\Pi_{T_{\Lambda,\beta_0}} Z_1 + Z_2, \tag{4.3.5}$$

where $\hat{\Delta}_1 = \hat{\Delta}_{1,Syn}$ *or* $\hat{\Delta}_{1,Msyn}$, $\Delta_1 = E(\Delta_{1t})$, $\phi = \lim_{T_1,T_2\to\infty}\sqrt{T_2/T_1}$, Z_1 *is defined in Theorem 4.3.2,* Z_2 *is independent with* Z_1 *and distributed as* $N(0,\Sigma_2)$, $\Sigma_2 = \lim_{T_2\to\infty} T_2^{-1}$ $\sum_{t=T_1+1}^{T}\sum_{s=T_1+1}^{T} E(v_{1t}v_{1s})$, $v_{1t} = \Delta_{1t} - E(\Delta_{1t}) + u_{1t}$, *and* u_{1t} *has zero mean and is defined in (4.2.1).*

The proof of Theorem 4.3.3 is given in Appendix A.

Although one can use projection theory to characterize the asymptotic distribution of $\sqrt{T_1}(\hat{\Delta}_1 - \Delta_1)$, the inference is not straightforward as one has to know β_0 in order to calculate the tangent cone T_{Λ,β_0}. We show in Section 4.4 that a carefully designed subsampling method can be used to conduct valid inference. In particular, one does not need to know β_0 when using the subsampling method for inference.

4.3.4. 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 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

$$y_{1t} = \alpha_0 t + z_t'\beta_0 + u_{1t} \qquad t = 1, ..., T_1, \tag{4.3.6}$$

where $z_t = (1, \eta_{2t}, ..., \eta_{Nt})'$, and η_{jt} is the de-trended data from y_{jt} for $j = 2, ..., N$. Let $\hat{\alpha}_{T_1}$ and $\hat{\beta}_{T_1}$ be the constrained least squares estimators of α_0 and β_0 subject to $\beta_j \geq 0$ for $j = 2, ..., N$ and $\sum_{j=2}^{N}\beta_j = 1$ for the synthetic control estimator (or without the sum to one restriction for the modified synthetic control estimator) using the pre-treatment data. We estimate y_{1t}^0 by $\hat{y}_{1t}^0 = \hat{\alpha}_{T_1} t + z_t'\hat{\beta}_{T_1}$ and estimate the ATE is estimated by

$$\hat{\Delta}_1 = \frac{1}{T_2}\sum_{t=T_1+1}^{T}\hat{\Delta}_{1t}, \tag{4.3.7}$$

where $\hat{\Delta}_{1t} = y_{1t} - \hat{y}^0_{1t}$. 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 $\hat{\gamma}_{OLS}$ denote the ordinary least squares estimator of γ_0. Define $M_{T_1} = \sqrt{T_1} diag(T_1, 1, ..., 1)$, 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))

$$M_{T_1}(\hat{\gamma}_{OLS} - \gamma_0) \xrightarrow{d} N(0, \Omega), \tag{4.3.8}$$

where Ω is a $(N+1) \times (N+1)$ positive definite matrix, the explicit definition of which is presented in 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 modified synthetic control models are $\Lambda_{Syn} = \{\gamma \in \mathcal{R}^{N+1} : \ \gamma_j \geq 0 \text{ for } j = 3, ..., N+1, \ \sum_{j=3}^{N+1} \gamma_j = 1\}$; and $\Lambda_{Msyn} = \{\gamma \in \mathcal{R}^{N+1} : \ \gamma_j \geq 0 \text{ for } j = 3, ..., N+1\}$. Define the synthetic control estimator

$$\hat{\gamma}_{T_1} = \arg\min_{\gamma \in \Lambda} \sum_{t=1}^{T_1} (y_{1t} - w_t'\gamma)^2, \tag{4.3.9}$$

where $w_t = (t, z_t')'$ and $\Lambda = \Lambda_{Syn}$ or Λ_{Msyn}. Then similar to Theorem 4.3.2, we have the next theorem.

Theorem 4.3.4 *Let Z_3 denote the limiting distribution of $\sqrt{T_1}(\hat{\gamma}_{OLS} - \gamma_0)$ as described in (4.3.8). Then under the assumptions D1 to D3 presented in the Appendix D, we have*

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

where T_{Λ,γ_0} is the tangent cone of Λ evaluated at γ_0, and Z_3 is the weak limit of $M_{T_1}(\hat{\gamma}_{OLS} - \gamma_0)$ as described in (4.3.8).

With Theorem 4.3.4 we can derive the asymptotic distribution of $\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$.

Theorem 4.3.5 *Under the same conditions as in Theorem 4.3.4, we have*

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

where $\hat{\Delta}_1$ *is defined in (4.3.7),* $\Delta_1 = E(\Delta_{1t})$*,* $c = (\sqrt{\phi}(2+\phi), \sqrt{\phi}E(z_t'))'$*,* $\phi = \lim_{T_1,T_2\to\infty} \sqrt{T_2/T_1}$*,* Z_3 *is the limiting distribution of* $M_{T_1}(\hat{\gamma}_{OLS} - \gamma_0)$ *as defined in (4.3.8), and* Z_2 *is independent of* Z_3 *and normally distributed with zero mean and variance* Σ_2*.*

We provide the intuition of the proof of Theorem 4.3.5 in the Appendix D.

4.4. Inference Theory

In this section, we discuss inference methods for the ATE estimator $\hat{\Delta}_1$. For ease of exposition, we only discuss inference for the stationary data case. For trend-stationary data, one can first de-trend the data and then use the inference method discussed in this section for the de-trended data. In Section 4.4.1, we consider the case where both T_1 and T_2 are large while in Section 4.4.2, we discuss the case of small T_2.

4.4.1. Subsampling method

As discussed in the above section, the inference theory for the synthetic control estimator is complicated. The asymptotic distribution of $\hat{\beta}_{T_1}$ depends on whether $\beta_{0,j} = 0$ or $\beta_{0,j} > 0$ for $j = 2, ..., N$. When $\beta_{0,j} > 0$ for all $j = 2, ..., N$, asymptotically, the constraints are non-binding 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 $j \in \{2, ..., N\}$, the asymptotic distribution of the constrained estimator is much more complex (e.g., (4.3.4)). The asymptotic distribution of the synthetic control coefficient estimators depends on whether the true parameters take value at the boundary or not. 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 to the boundary of the parameter space, the standard bootstrap method does not work

(Andrews, 2000; Fang and Santos, 2016)). We resolve this difficulty by proposing our easy-to-implement subsampling method. The proposed method works whether constraints are binding, partially binding or non-binding. 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.[2]

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. A 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

$$y_{1t} = x_t'\beta_0 + d_t\Delta_{1t} + u_{1t}, \qquad t = 1, ..., T_1, T_1 + 1, ..., T, \tag{4.4.1}$$

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

Substituting (4.4.1) into the left-hand-side of (4.3.5) we obtain

$$\begin{aligned} \hat{A} &\overset{def}{=} \sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) \\ &= -\sqrt{\frac{T_2}{T_1}}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} x_t'\right]\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) \\ &\quad + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t}, \end{aligned} \tag{4.4.2}$$

where $v_{1t} = \Delta_{1t} - \Delta_1 + u_{1t}$.

[2]Hong and Li (2015) show that numerical differentiation bootstrap method can consistently estimate the limiting distribution in many cases where the conventional bootstrap is known to fail. One can also use Hong and Li (2015) method to conduct inference for the synthetic control estimator. In this essay, we focus on the simple subsampling method.

Now we impose the 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 assumptions is violated, more sophisticated method such as block subsampling methods can be used to deliver valid inferences.

Expression (4.4.2) suggests that we only need to apply the subsampling method to the term $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ because only this term is related to the constrained estimator. We now describe the subsampling steps. In Appendix A, we show that when v_{1t} is serially uncorrelated, one can consistently estimate Σ_2 by $\hat{\Sigma}_2 = \frac{1}{T_2}\sum_{t=T_1+1}^{T} \hat{v}_{1t}^2$, where $\hat{v}_{1t} = \hat{\Delta}_{1t} - \hat{\Delta}_1$. We generate $v_{1t}^* \sim$ iid $N(0, \hat{\Sigma}_2)$ for $t = T_1 + 1, ..., T$. 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 $t = 1, ..., m$, we randomly draw (y_{1t}^*, x_t^*) from $\{y_{1t}, x_t\}_{t=1}^{T_1}$ with replacement (subsampling).[3] Then we use the subsample $\{y_{1t}^*, x_t^*\}_{t=1}^{m}$ to estimate β_0 by the constrained least squares method, i.e., $\hat{\beta}_m^* = \arg\min_{\beta\in\Lambda} \sum_{t=1}^{m}(y_{1t}^* - x_t^{*'}\beta)^2$. The subsampling-bootstrap version of the statistic $\hat{A}$ is given by

$$\hat{A}^* = -\sqrt{\frac{T_2}{T_1}}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} x_t'\right]\sqrt{m}(\hat{\beta}_m^* - \hat{\beta}_{T_1}) + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t}^*. \tag{4.4.3}$$

We repeat the above process for a large number of times, say, 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 $\hat{A}_{(1)}^* \le \hat{A}_{(2)}^* \le ... \le \hat{A}_{(J)}^*$. Then the $1-\alpha$ confidence interval for Δ_1 is given by

$$[\hat{\Delta}_1 - T_2^{-1/2}\hat{A}_{((1-\alpha/2)J)}^*\,,\ \hat{\Delta}_1 - T_2^{-1/2}\hat{A}_{((\alpha/2)J)}^*]. \tag{4.4.4}$$

[3]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 $m = T_1$.

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

Theorem 4.4.1 *Under the same conditions as in Theorem 4.3.3 and assuming that* $m \to \infty$ *and* $m/T_1 \to 0$ *as* $T_1 \to \infty$*, then for any* $\alpha \in (0,1)$*, the* $(1-\alpha)$ *confidence interval of* Δ_1 *can be consistently estimated by* $[\hat{\Delta}_1 - T_2^{-1/2}\hat{A}^*_{((1-\alpha/2)J)}\,,\ \hat{\Delta}_1 - T_2^{-1/2}\hat{A}^*_{((\alpha/2)J)}]$.

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 synthetic control ATE estimator. Politis et al. (1999) provide proofs and arguments showing that 'subsampling method works' under very weak regularity conditions.

Remark 4.4.2 *We apply the subsampling method only to part of the statistic,* A_1*, because* 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,* A_2*. This is important because it is difficult to apply a subsampling method to* A_2 *(which depends on* T_2*) as* T_2 *is usually smaller than* T_1*. Subsampling methods applied to* A_2 *with subsample sizes much smaller than* T_2 *usually do not work well in practice.*

Remark 4.4.3 *Even though we randomly draw* (y_t^*, x_t^*) *from* $\{y_s, x_s\}_{s=1}^{T_1}$ *for* $t = 1, ..., m$*, 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, e.g.* $\{y_{jt}\}_{j=1}^{N}$ *is generated by some serially correlated common factors. We only need that the idiosyncratic error* u_{1t} *in (4.2.1) is serially uncorrelated. This can be easily tested given data. In Section 4.5, 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 one can replace the random subsampling method by block subsampling method. We leave the formal justification of using block subsampling method as a future research topic.*

Remark 4.4.4 *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 $j \in \{2, ..., N\}$, and the statistic*

$$\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) = -\sqrt{\frac{T_2}{T_1}}\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} x_t'\right]\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0) + \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t} \tag{4.4.5}$$

contains a term $T_2^{-1}\sum_{t=T_1+1}^{T} v_{1t}$, which is not related to $\hat{\beta}_{T_1}$, the subsampling method works reasonably well for a wider range of m.[4] *Even for $m = n$ (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 4.5.2 and in the Appendix E.*

4.4.2. Inference theory when T_2 is small

The asymptotic theories presented in Section 4.4.1 assume that both T_1 and T_2 are large. However, in practice, many datasets have T_2 much smaller than T_1. When T_2 is small, Ferman and Pinto (2015) propose using Andrews' (2003) end-of-sample instability test to conduct inference for DID and synthetic control ATE estimators. One can also use the permutation test proposed in Chernozhukov et al. (2017) in this case. In this subsection, we focus on Andrews's (2003) test to illustrate why it works in the large T_1 and small T_2 scenario.

When T_1 is large and T_2 is small, the first term on the right-hand-side of (4.4.2), which has an order $O_p(\sqrt{T_2/T_1}) = O_p(T_1^{-1/2})$, becomes negligible. Then we have

$$\hat{A} = \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T}(y_{1t} - \hat{y}_{1t}^0 - \Delta_1) = \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t} + o_p(1), \tag{4.4.6}$$

[4]Note that $v_{1t} = \Delta_{1t} - \Delta_1 + u_{1t}$

where $v_{1t} = \Delta_{1t} - \Delta_1 + u_{1t}$ has zero mean and finite variance.

One can test the null hypothesis of a constant treatment effect H_0: $\Delta_{1t} = \Delta_{1,0}$ for some pre-specified value $\Delta_{1,0}$ for $t = T_1 + 1, ..., T$ against a one-sided treatment effect such as H_1: $\Delta_1 = E(\Delta_{1t}) > \Delta_{1,0}$ for $t = T_1 + 1, ..., T$. Following Andrews (2003), we can use the following test statistic

$$\hat{B}_{T_2} = \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} (y_{1t} - \hat{y}_{1t}^0 - \Delta_{1,0}) = \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t,0} + o_p(1), \tag{4.4.7}$$

where $v_{1t,0} = \Delta_{1t} - \Delta_{1,0} + u_{1t}$. Under H_0, $v_{1t,0} = u_{1t}$ has zero mean, and under H_1, it has a positive mean.

To conduct inference based on the test statistic $\hat{B}_{T_2}$, we compute the following quantity

$$\hat{B}_{T_2,j} \overset{def}{=} \frac{1}{\sqrt{T_2}} \sum_{t=j}^{T_2+j-1} \hat{u}_{1t} = \frac{1}{\sqrt{T_2}} \sum_{t=j}^{T_2+j-1} u_{1t} + O_p(T_1^{-1/2}), \qquad \text{for } j = 1, ..., T_1 + 1 - T_2, \tag{4.4.8}$$

where for $t = 1, ..., T_1$, $\hat{u}_{1t} = y_{1t} - \hat{y}_{1t}^0 = x_t'(\beta_0 - \hat{\beta}_{T_1}) + u_{1t} = u_{1t} + O_p(T_1^{-1/2})$ because $\hat{\beta}_{T_1} - \beta_0 = O_p(T_1^{-1/2})$. The empirical distribution of $\{\hat{B}_{2,j}\}_{j=1}^{T_1+1-T_2}$ can be used to obtain critical values for the test statistic $\hat{B}_{T_2}$ under the null hypothesis H_0: $\Delta_{1t} = \Delta_{1,0}$ for all $t = T_1 + 1, ..., T$. If $\hat{B}_{T_2}$ is at the tail of this empirical distribution, we reject the null hypothesis and accept the alternative hypothesis.

Remark 4.4.5 *We can only test a constant treatment effect for each post-treatment period using Andrews' (2003) test, i.e., we can only test $\Delta_{1,t} = \Delta_{1,0}$ for all $t = T_1 + 1, ..., T$. We cannot test $\Delta_1 = \Delta_{1,0}$ because under this null hypothesis, $\Delta_{1,t} - \Delta_1 = \Delta_{1t} - \Delta_{1,0}$ has zero mean and finite variance. We cannot use pre-treatment data to estimate the variance of Δ_{1t}. Therefore, Andrews' (2003) method become invalid when treatment effects varies with t.*

Remark 4.4.6 *Andrews' (2003) test will have good estimated sizes for large T_1. However, it is not a consistent test because T_2 is small. The power of the test depends on the strength*

of the treatment effects under H_1. The power of the test should increase with T_2, but when T_2 is large, an estimation error of order $\sqrt{T_2/T_1}$ may not be negligible, rendering Andrews' (2003) test invalid.

4.5. Simulation results

In this section, we first consider the case of large T_1 and T_2 and examine the performance of our subsampling method inferences through simulations. Then we consider the case of large T_1 and small T_2 and examine the performance of Andrews' (2003) end-of-sample instability test.

4.5.1. A three-factor data generating process

We conduct simulation studies using the same data generating process as in Hsiao et al. (2012) and Du and Zhang (2015). We consider the following three-factor data generating process.

$$y_t^0 = a + Bf_t + u_t, \qquad t = 1, ..., T \tag{4.5.1}$$

where $y_t^0 = (y_{1t}^0, y_{2t}^0, ..., y_{Nt}^0)'$, $a = (a_1, a_2, ..., a_N)'$ and $u_t = (u_{1t}, u_{2t}, ..., u_{Nt})'$ are all $N \times 1$ vectors, $B = (b_1, b_2, ..., b_N)'$ 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 where $f_{1t} = 0.8f_{1t-1} + \epsilon_{1t}$, $f_{2t} = -0.6f_{1t-1} + \epsilon_{2t} + 0.8\epsilon_{2t-1}$, and $f_{3t} = \epsilon_{3t} + 0.9\epsilon_{3t-1} + 0.4\epsilon_{3t-2}$, and ϵ_{jt} is iid $N(0, \sigma^2)$ with $\sigma^2 = 0.5$. We choose $(a_1, a_2, ..., a_N) = (1, 1, ..., 1)$.

We use a set up similar to our Warby Parker empirical data by setting $T_1 = 90$, $T_2 = 20$, $T = T_1 + T_2 = 110$ and $N = 11$ (with 10 control units). For factor loadings, we use b_1 to denote the 3×1 vector of loadings for the first unit (the treated unit), $\tilde{b}_2 = (b_2, ..., b_{s+1})$ to denote the $3 \times s$ loading matrix for units $j = 2, ..., s+1$ ($1 \leq s \leq N-2$) and $\tilde{b}_3 = (b_{s+1}, ..., b_N)$ to denote the $3 \times (N - 1 - s)$ loading matrix for the last $N - s - 1$ units. We fix $s = 6$ and

consider the following two sets for factor loadings:

$$DGP1: \quad b_1 = \mathbf{1}_{3\times1}; \quad b_j = \mathbf{1}_{3\times1} \text{ for } j = 2, ..., 7; \quad \text{and} \quad b_j = \mathbf{0}_{3\times1} \text{ for } j = 8, ..., 11,$$

$$DGP2: \quad b_1 = 2(\mathbf{1}_{3\times1}); \quad b_j = \mathbf{1}_{3\times1} \text{ for } j = 2, ..., 7; \quad \text{and} \quad b_j = \mathbf{0}_{3\times1} \text{ for } j = 8, ..., 11,$$

where $\mathbf{1}_{3\times1}$ and $\mathbf{0}_{3\times1}$ denote 3×1 vectors of ones and zeros, respectively.

Note that for both DGP1 and DGP2, 6 out of 10 control units have non-zero loadings and the remaining 4 control units have zero loadings. For DGP1, all non-zero factor loadings are set to be ones so that the treated and the control units (with non-zero 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 non-zero 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}:

$$\Delta_{1t} = \alpha_0 \left[\frac{e^{z_t}}{1 + e^{z_t}} + 1 \right], \qquad t = T_1 + 1, ..., T, \tag{4.5.2}$$

where $z_t = 0.5 z_{t-1} + \eta_t$ and η_t is iid $N(0, 0.5^2)$. Note that 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 (4.5.2). There exist zero or positive treatment effects corresponding to $\alpha_0 = 0$ and $\alpha_0 > 0$, respectively.

4.5.2. Simulations results for coverage probabilities

In this section, we report estimated coverage probabilities. Since we have $N = 11$ parameters in the regression model, we need to select a subsample size $m > N$. We select $m = 20$, 40, 60, 80 and 90. Note that we include the case where the subsample size m equals the full sample size, $m = T_1 = 90$, for the reason discussed in remark 4.4.4. The number of simulations are 1000 for each setup, and 400 subsamples are generated within each simulation.

Table 10 reports estimated coverage probabilities for DGP1. The top panel corresponds to no treatment effects ($\alpha_0 = 0$) while the bottom panel corresponds to the case of a positive treatment effects ($\alpha_0 = 1$ in (4.5.2)). From Table 10 we observe that both the standard synthetic control and the modified synthetic control methods give estimated coverage probabilities that are close to their nominal levels for all values of m including $m = T_1 = 90$. The reason that the subsampling method works well for a wide range of m was discussed in remark 4.4.4. See the supplementary Appendix E for further explanations. Finally, we observe that the estimation results are not sensitive to different values of α_0 (the magnitude of the treatment effects).

Table 11 reports estimated coverage probabilities for DGP2. From Table 11, we observe that the standard synthetic control method gives biased estimation results. The estimated coverage probabilities are much smaller than the corresponding nominal values. The estimation results are biased for DGP2 because the treated and the control units are drawn from different distributions. In contrast to the standard synthetic control approach, the modified synthetic control method has good estimated coverage probabilities. This verifies that the modified synthetic control method is robust to data drawn from heterogeneous distributions. Also, similar to the DGP1 case, the results are not sensitive to different values of α_0.

We conduct additional simulations with large T_1 (100, 200) and small T_2 (3, 5) and use Andrews' (2003) end-of-sample instability test to test the null hypothesis of zero ATE. The results are reported in the Appendix F.

We also conducted simulations of Andrews' (2003) test under DGP1 using $T_1 = 90$ and $T_2 = 20$ (same T_1 and T_2 as in our empirical data). Based on 10,000 simulations with $\alpha_0 = 0$, the estimated sizes are 0.1660 and 0.1964 for nominal levels 5% and 10%, respectively. We see that for $T_2 = 20$, $T_1 = 90$ is not large enough for the test to have good estimated sizes because an error term of order $\sqrt{T_2/T_1}$ is not negligible which causes Andrews' (2003) test to be invalid in our context. Therefore, the end-of-sample stability tests and the subsampling

Table 10: Coverage probabilities for DGP1 (a common distribution)

	DGP1 with $\alpha_0 = 0$ (zero treatment)									
	Synthetic control					Modified synthetic control				
m	20	40	60	80	90	20	40	60	80	90
50%	0.499	0.492	0.462	0.500	0.482	0.517	0.489	0.488	0.507	0.493
80%	0.767	0.786	0.762	0.788	0.778	0.785	0.798	0.786	0.800	0.790
90%	0.883	0.890	0.879	0.889	0.885	0.894	0.879	0.882	0.885	0.883
95%	0.940	0.934	0.940	0.945	0.936	0.942	0.945	0.940	0.945	0.938
	DGP1 with $\alpha_0 = 1$ (positive treatment)									
	Synthetic control					Modified synthetic control				
m	20	40	60	80	90	20	40	60	80	90
50%	0.497	0.510	0.509	0.466	0.483	0.497	0.510	0.509	0.466	0.483
80%	0.805	0.775	0.784	0.778	0.782	0.805	0.775	0.784	0.778	0.782
90%	0.903	0.868	0.891	0.877	0.884	0.903	0.868	0.891	0.877	0.884
95%	0.944	0.931	0.950	0.929	0.934	0.944	0.931	0.950	0.929	0.934

Table 11: Coverage probabilities for DGP2 (a heterogenous distribution)

	DGP2 with $\alpha_0 = 0$ (zero treatment)									
	Synthetic control					Modified synthetic control				
m	20	40	60	80	90	20	40	60	80	90
50%	0.294	0.308	0.314	0.292	0.306	0.474	0.458	0.492	0.474	0.470
80%	0.526	0.534	0.522	0.510	0.540	0.776	0.756	0.770	0.742	0.738
90%	0.658	0.630	0.638	0.632	0.666	0.884	0.854	0.876	0.844	0.866
95%	0.752	0.710	0.720	0.720	0.754	0.936	0.924	0.930	0.908	0.926
	DGP2 with $\alpha_0 = 1$ (positive treatment)									
	Synthetic control					Modified synthetic control				
m	20	40	60	80	90	20	40	60	80	90
50%	0.306	0.278	0.276	0.278	0.286	0.508	0.486	0.468	0.478	0.482
80%	0.522	0.478	0.510	0.472	0.496	0.802	0.764	0.796	0.796	0.770
90%	0.634	0.614	0.620	0.580	0.594	0.888	0.890	0.894	0.894	0.884
95%	0.710	0.716	0.710	0.678	0.668	0.948	0.944	0.940	0.950	0.944

testing procedures are complements to each other. The former can be used when T_2 is small while the later is preferred when T_2 is not small.

4.6. An Empirical Application

In this section, we illustrate the usefulness of the modified synthetic control method in practice. In the application, we calculate the ATE based on the modified synthetic control method and the confidence intervals using the subsampling method.

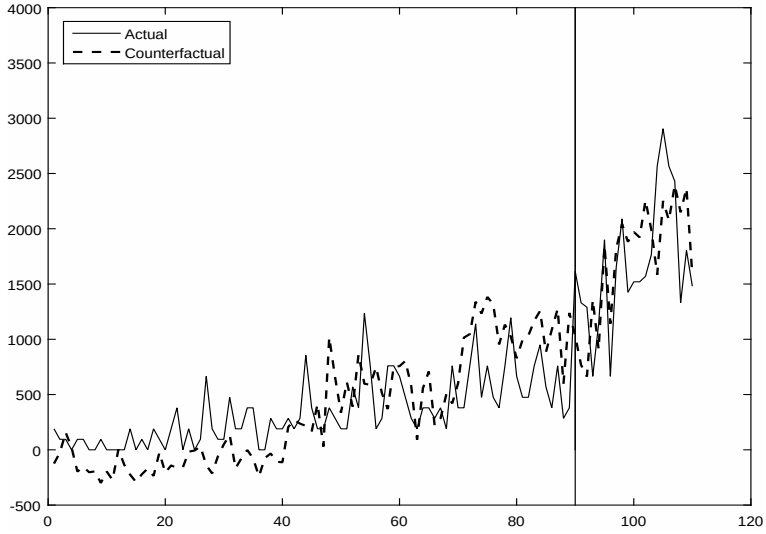
4.6.1. Warby Parker and ATE estimation

We have data from WarbyParker.com, an online provider of high quality eyewere at modest prices ($95 instead of $300+range). In February 2010, WarbyParker.com opened its first physical showroom in New York City. Later, they opened showrooms in several more cities hoping that opening physical showrooms would significantly promote sales. They opened a showroom in Columbus, Ohio on October 11, 2011. In this section, we want to evaluate how the showroom opening in Columbus affected Columbus' average weekly sales (the average treatment effect) in the post-treatment period. As discussed in Section 4.2 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. Hence, we run a constrained regression, i.e., we regress weekly sales of Columbus on sales of control cities to obtain the estimated coefficients under the restriction that the coefficients are non-negative. 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.

Figure 18 plots Columbus' actual sales (solid line) and the in-sample-fit and out-of-sample

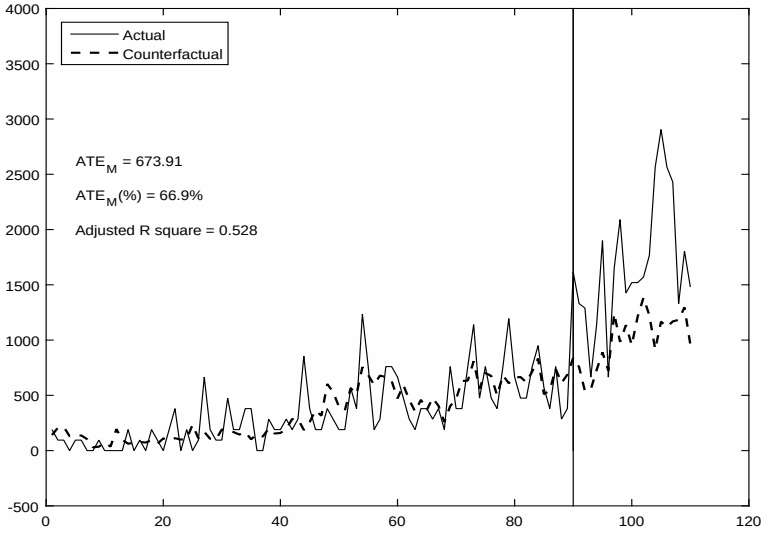
counterfactual forecast (dotted line) curve computed via the synthetic control method (i.e., using (4.2.3)). From Figure 18, we see that the synthetic control method fits in-sample data poorly as it underestimates the actual sales for the first half of the in-sample data and overestimates the actual sales for the second half of the in-sample data. We can also see that if one applies the synthetic control method to this data set, one overestimates the counterfactual outcome which results in an underestimation of ATE. The reason for this is that without the restriction of coefficients adding to one, the sum of the slope coefficients is 0.234. The standard synthetic control method imposes the restriction that the slope coefficients add 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 pre-treatment period data). This leads the fitted curve to be below the actual data for the first half of the pre-treatment time period and above the actual data for the second half of the pre-treatment time period. 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.

Figure 18: Columbus: The synthetic control fitted curve



The above analysis suggests that restricting the slope coefficients adding to one is the reason for a large estimation bias of the standard synthetic control method. Therefore, we relax the

Figure 19: Columbus: Modified synthetic control ATE estimation



weights add-to-one condition, i.e., we only keep the non-negativity of the weights but drop the add-to-one restriction. The estimation results are plotted in Figure 19. The results in Figure 19 show a greatly improved in-sample-fit. Unlike Figure 18, the fitted curve in Figure 19 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 of showroom opening in Columbus is highly statistically significant.

4.6.2. Confidence intervals for the ATE

In this section, we use the subsampling method discussed in Section 4.4 to estimate confidence intervals (CI) for the ATE (Δ_1) estimated in Section 4.6.1. Since our proposed subsampling method requires that the idiosyncratic error u_{1t} defined in (4.2.1) 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 $m = 20, 40, 60, 80, 90$. 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 (4.4)). We then sort these 10,000 statistics to obtain $\alpha/2$ percentile and $(1-\alpha/2)$ percentile for $\alpha = 0.2$, 0.1, 0.05 and 0.01. The results are given in Table 12.

First, from Table 12 we observe that the estimated confidence intervals are quite 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 4.4.4 and further illustrated in the Appendix E.3, the subsampling method works well for a wide range of m values. Next, we notice that the lower bound of these intervals are all positive and far above zero for all m values. 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 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 at Columbus significantly increased WarbyParker.com's eyewear sales.

Table 12: Confidence intervals (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]

4.6.3. Robustness Checks

In this section, 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 to the unconstrained (i.e. least squares or 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.

To save space we only report robustness checks 1 and 2 here and report robust checks 3 and 4 in Appendix F.

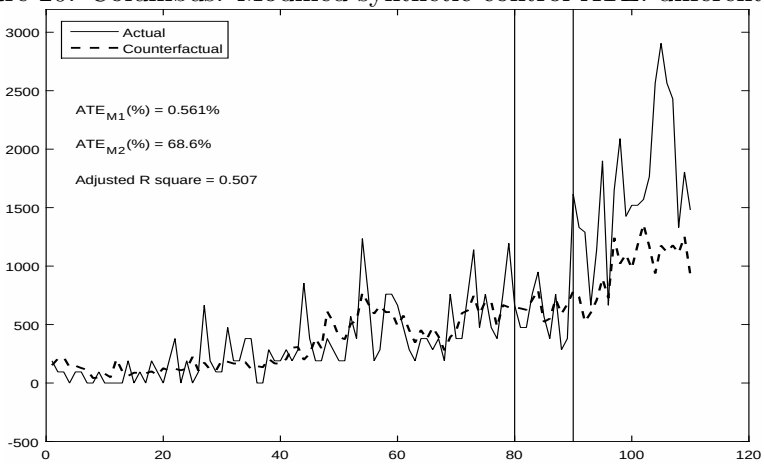
Change the treatment date

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 80, we estimate the model using the modified synthetic control 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 20, we see that for the periods 81 to 90, the predicted sales trace the actual sales quite closely. The ATE percentage increase for these 10 periods is 0.561% which is quite 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 $\hat{\Delta}_1$ estimated using data from $t = 81$ to 90 with 10,000 subsampling simulations. The results are given in Table 13. We see that all confidence intervals contain zero. Hence, we cannot reject the null hypothesis that there is no treatment effects during the period of $81 \leq t \leq 90$ at any conventional levels. Thus, this robust check supports the modified synthetic control estimation result.

Table 13: Confidence intervals (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]

Figure 20: Columbus: Modified synthetic control ATE: different 'T_1'



Comparison with the unconstrained estimator (OLS or HCW)

In this subsection, we consider using the ordinary least squares method[5] to estimate the counterfactual outcome. Let $\hat{\beta}_{OLS}$ denote the least squares estimator of β using the pre-treatment sample. Then the counterfactual outcome is estimated by $\hat{y}_t^0 = x_t'\hat{\beta}_{OLS}$ (e.g., Hsiao et al. (2012)). Applying this method to the Columbus data gives an estimated ATE of \$645.26 increase in weekly sales after the opening of a showroom in Columbus. While this number is close to the ATE estimation result of \$673.91 by the modified synthetic control, we would like to compare the out-of-sample forecasting performances of the two estimation methods in order to judge which method gives a more accurate ATE estimation result.

The difference between the least squares method and our modified synthetic control method is that the synthetic control method imposes a non-negativity restriction on the slope coefficients when estimating the regression model using the pre-treatment data. The rationale for imposing the non-negativity constraints is that outcome variables from treated and control units are driven by some common factors and therefore, they are more likely to move up and down together. Imposing a correct restriction can improve out-of-sample forecast. Therefore, we compare the out-of-sample forecast performances of the modified synthetic

[5]We interchangeably use ordinary least squares, HCW and unconstrainted estimator

control method and the least squares method. We choose a value $T_0 \in (1, T_1) = (1, 90)$ to estimate the regression model. Then we forecast outcome y_{1t} for $t = T_0 + 1, ..., T_1$. Since there is no treatment prior to T_1, we can compare the average prediction squared error over the period $t = T_0 + 1, ..., T_1$. Specifically, we estimate the following model

$$y_t = x_t'\beta + u_{1t}, \qquad t = 1, ..., T_0 \tag{4.6.1}$$

by the modified synthetic control and the least squares method. Let $\hat{\beta}_{T_0}$ and $\hat{\beta}_{OLS}$ denote the resulting estimators using the two methods, respectively. We predict y_{1t}^0 by $\hat{y}_{1t,Msyn}^0 = x_t'\hat{\beta}_{T_0}$ and $\hat{y}_{1t,OLS}^0 = x_t'\hat{\beta}_{OLS}$ for $t = T_0 + 1, ..., T_1$. Then we compute the prediction MSEs by $PMSE_{Msyn} = (T_1 - T_0)^{-1} \sum_{t=T_0+1}^{T_1} (y_{1t} - \hat{y}_{1t,Msyn}^0)^2$ and $PMSE_{OLS} = (T_1 - T_0)^{-1} \sum_{t=T_0+1}^{T_1} (y_{1t} - \hat{y}_{1t,OLS}^0)^2$. In essay 1, we consider the cases where the 'pre-treatment' estimation sample is larger than the 'post-treatment' evaluation sample. We choose six different values for $T_0 = \{60, 65, 70, 75, 80, 85\}$. The corresponding evaluation sample sizes are $T_1 - T_0 = \{30, 25, 20, 15, 10, 5\}$. We report the ratio of PMSE as $PMSE_{OLS}/PMSE_{Msyn}$. The results are reported in Table 14.

Table 14: Out-of-sample Prediction MSE ratio

T_0	60	65	70	75	80	85
$\frac{PMSE_{OLS}}{PMSE_{Msyn}}$	1.680	1.104	1.020	1.273	1.188	1.143

From Table 14 we observe that the least squares method has larger PMSE than the modified synthetic control method for all cases. The PMSE for the former ranges from 2% to 68% larger than the later. Thus, the empirical example shows that, in order to more accurately predict the counterfactual outcomes for the treated unit, it is helpful to impose non-negativity restriction on the slope coefficients when estimating model (4.6.1).

4.7. Conclusion

The synthetic control method is a popular and powerful way for estimating average treatment effects. This essay provides the inference theory of the synthetic control method (and

modified synthetic control method) under long panels with large pre and post-treatment periods. We derive the asymptotic distribution of the synthetic control ATE estimator using projection theory. Because the asymptotic distribution is non-normal and non-standard, 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. This work complements the case of long pre-treatment and short post-treatment data where end of sample instability tests are applied (Ferman and Pinto, 2015, 2016; Andrews, 2003) and permutation tests proposed in Chernozhukov et al. (2017) for conducting inference.

We also prove that, when the pre-treatment sample size is larger than the number of control units (i.e., $T_1 > N - 1$), the synthetic control estimator, as a constrained minimization problem, has an unique solution under a mild condition that the T_1 by N data matrix has a full column rank. In addition, we show the modified synthetic control method can give reliable ATE estimation results even when the "parallel lines" assumption is violated for the standard synthetic control method. Simulations show that the modified synthetic control method performs well in practice. Finally, we apply the synthetic and modified synthetic control method to estimate ATE of opening a showroom by an e-tailer. The empirical application demonstrates that when the standard synthetic control method fits the data poorly, the modified synthetic control method fits the data well and gives reasonable ATE estimation results.

Appendix A: Proofs of Theorems 4.3.2, 4.3.3 and 4.4.1

A.1. Proof of Theorem 4.3.2

The constrained estimator is defined by

$$\hat{\beta}_{T_1} = \arg\min_{\beta \in \Lambda} (\beta - \hat{\beta}_{OLS})'(X'X/T_1)(\beta - \hat{\beta}_{OLS}). \tag{A.1}$$

Thus, $\hat{\beta}_{T_1}$ is the projection of $\hat{\beta}_{OLS}$ onto Λ with respect to the norm $\|a\| = \sqrt{a'(X'X/T_1)a}$ which is random, rendering the theory in **?** not directly applicable. However, since $X'X/T_1 \xrightarrow{p} E(X_t X_t')$, we show that one can replace $X'X/T_1$ by $E(X_t X_t')$ without affecting the asymptotic results. Define the following "infeasible estimator" (it is infeasible because $E(X_t' X_t)$ is unknown in practice):

$$\tilde{\beta}_{T_1} = \arg\min_{\beta \in \Lambda} (\beta - \hat{\beta}_{OLS})' E(X_t X_t')(\beta - \hat{\beta}_{OLS}) = \Pi_\Lambda \hat{\beta}_{OLS} \ , \tag{A.2}$$

where Π_Λ is the projection onto Λ with respect to the norm $\|a\| = \sqrt{a' E(X_t X_t')a}$, i.e., $\Pi_\Lambda \beta = \arg\min_{\lambda \in \Lambda} (\beta - \lambda)' E(X_t X_t')(\beta - \lambda)$. By Lemma 4.6 of Zarantonello (1971) on page 300 and Proposition 4.1 of Fang and Santos (2016), we know that

$$\begin{aligned} \sqrt{T_1}(\tilde{\beta}_{T_1} - \beta_0) &= \sqrt{T_1}(\Pi_\Lambda \hat{\beta}_{OLS} - \Pi_\Lambda \beta_0) \\ &= \sqrt{T_1} \Pi_{T_{\Lambda,\beta_0}} (\hat{\beta}_{OLS} - \beta_0) + o_p(1) \\ &= \Pi_{T_{\Lambda,\beta_0}} \sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0) + o_p(1) \\ &\xrightarrow{d} \Pi_{T_{\Lambda,\beta_0}} Z_1, \end{aligned} \tag{A.3}$$

where the first equality follows from $\tilde{\beta}_{T_1} = \Pi_\Lambda \hat{\beta}_{OLS}$ and $\beta_0 \in \Lambda$ so that $\beta_0 = \Pi_\Lambda \beta_0$.

We give some explanations of the above derivations. Hilbert Space projection onto convex sets was studied by Zarantonello (1971) and extended to general econometric model settings by Fang and Santos (2016). The projection operator Π_Λ: $\mathcal{R}^N \to \Lambda$ (Λ is a convex subset in $\mathcal{R}^N$) can be viewed as a functional mapping. Zarantonello (1971) showed that Π_Λ is (Hadamard) directional differentiable, and its directional derivative at $\beta_0 \in \Lambda$ is $\Pi_{T_{\Lambda,\beta_0}}$, the projection onto the tangent cone of Λ at β_0. Hence, the second equality of (A.3) follows from a functional Taylor expansion, the third equality follows from that T_{Λ,β_0} is positive homogenous of degree one.[6] The last line follows from $\sqrt{T_1}(\hat{\beta}_{OLS} - \Lambda\beta_0) \xrightarrow{d} Z_1$ and the continuous mapping theorem because projection is a continuous mapping.

[6]The Projection T_{Λ,β_0} is not a linear operator. However, for $\alpha \geq 0$, we have $\alpha\, T_{\Lambda,\beta_0}\, \theta = T_{\Lambda,\beta_0}\, \alpha\, \theta$ for all $\theta \in \mathcal{R}^N$.

We can also see the term 'tangent cone' is similar to what we term the derivative of a function at a given point as a 'tangent line' of the function at the given point. Now, the functional derivative of the mapping Π_Λ is a projection onto the cone $\Pi_{T_{\Lambda,\beta_0}}$ (rather than a line). Therefore, it is called the 'tangent cone' of Λ at β_0 and is denoted as T_{Λ,β_0}. For readers' convenience, we give the formal definition of tangent cone of Λ at $\theta \in \mathcal{R}^N$ below:

$$T_{\Lambda,\theta} = \overline{\cup_{\alpha \geq 0} \alpha\{\Lambda - \Pi_\Lambda \theta\}}, \tag{A.4}$$

where for any set $A \in \mathcal{R}^N$, $\overline{A}$ is the closure of A ($\overline{A}$ is the smallest closed set that contains A).

It can be easily checked that for our synthetic control estimation problem, the tangent cone of Λ at β_0 is the same as the asymptotic range of $\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$.

In Lemma C.1 of the Appendix C we show that

$$\hat{\beta}_{T_1} = \tilde{\beta}_{T_1} + o_p(T_1^{-1/2}) = \Pi_\Lambda \hat{\beta}_{OLS} + o_p(T_1^{-1/2}). \tag{A.5}$$

Theorem 4.3.2 follows from (A.3) and (A.5).

A.2. Proof of Theorem 4.3.3

First, we write $\hat{A} = \sqrt{T_2}(\hat{\Delta}_1 - \Delta_1)$ defined in (4.4.2) as $\hat{A} = \hat{A}_1 + \hat{A}_2$, where

$$\hat{A}_1 = -\left[\frac{1}{T_2}\sum_{t=T_1+1}^{T} x_t'\right]\sqrt{\frac{T_2}{T_1}}\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0), \qquad \hat{A}_2 = \frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T} v_{1t}. \tag{A.6}$$

We know that $\hat{A}_2 \xrightarrow{d} Z_2$ by assumption 2, where Z_2 is distributed as $N(0, \Sigma_2)$. By Theorem 4.3.2 and assumption 1, we have $\hat{A}_1 \xrightarrow{d} A_1 = -\phi E(x_t')\Pi_{T_{\Lambda,\beta_0}} Z_1$, where $\phi = \lim_{T_1,T_2\to\infty}\sqrt{T_2/T_1}$ and Z_1 is the weak limit of $\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0)$, i.e., $\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0) \xrightarrow{d} Z_1$. Also, by Lemma A.1 and Theorem 3.2 of essay 1, we know that Z_1 and Z_2 are asymptotically

independent with each other, this implies that $A_1 = -\phi E(x_t)\Pi_{T_{\Lambda,\beta_0}} Z_1$ is asymptotically independent of Z_2. Hence, we have $\hat{A} \xrightarrow{d} -\phi E(x_t')\Pi_{T_{\Lambda,\beta_0}} Z_1 + Z_2$.

A.3. Proof of Theorem 4.4.1

The proof that $\hat{A}^*$ can be used to approximate the distribution of $\hat{A}$ consists of the following arguments. First, we show that one can consistently estimate Σ_2 by $\hat{\Sigma}_2 = \frac{1}{T_2}\sum_{t=T_1+1}^{T} \hat{v}_{1t}^2$ (when v_{1t} is serially uncorrelated), where $\hat{v}_{1t} = \hat{\Delta}_{1t} - \hat{\Delta}_1$. From $\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^0 = x_t'(\beta_0 - \hat{\beta}_{T_1}) + \Delta_{1t} + u_{1t} = \Delta_{1t} + u_{1t} + O_p(T_1^{-1/2})$ and $\hat{\Delta}_1 = \bar{x}'(\beta_0 - \hat{\beta}_{T_1}) + \bar{\Delta}_1 + \bar{u}_1 = \Delta_1 + O_p(T_1^{-1/2} + T_2^{-1/2})$, we have

$$\hat{\Sigma}_2 = \frac{1}{T_2}\sum_{t=T_1+1}^{T} (\Delta_{1t} + u_{1t} - \Delta_1)^2 + O_p(T_1^{-1/2} + T_2^{-1/2}) = \Sigma_2 + O_p(T_1^{-1/2} + T_2^{-1/2}).$$

Next, it is obvious that $T_2^{-1/2}\sum_{t=T_1+1}^{T} v_{1t}^* \overset{d}{\sim} T_2^{-1/2}\sum_{t=T_1+1}^{T} v_{1t} \xrightarrow{d} Z_2$, where $A \overset{d}{\sim} B$ means that A and B have the same asymptotic distribution. By the conditions that $m \to \infty$, $m/T_1 \to 0$ as $T_1 \to \infty$ and the weak convergence result of Theorem 4.3.2, we know that $\sqrt{m}(\hat{\beta}_m^* - \hat{\beta}_{T_1}) \overset{d}{\sim} \sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)$ by Theorem 2.2.1 of Politis et al. (1999). It follows that $\hat{A}^*$ defined in (4.4.3) and $\hat{A}$ defined in (4.4.2) have the same asymptotic distribution.

Appendix B: Uniqueness of the synthetic control estimator

B.1. Assumptions

We first list assumptions that are used in deriving the main results of the essay.

Assumption 1. The data $\{x_t\}_{t=1}^{T}$ is a weakly dependent stationary process so that laws of large number holds: $T_1^{-1}\sum_{t=1}^{T_1} x_t \xrightarrow{p} E(x_t)$ 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 t^{th} row given by $x_t' = (1, y_{2t}, ..., y_{Nt})$. Let $\phi = \lim_{T_1,T_2\to\infty}\sqrt{T_2/T_1}$, then ϕ is a finite non-negative constant.

Assumption 2. $\{u_{1t}\}_{t=1}^{T}$ is zero mean, serially uncorrelated 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, serially uncorrelated and satisfies a central limit theorem: $T_2^{-1/2} \sum_{t=T_1+1}^{T} v_{1t} \xrightarrow{d} N(0, \Sigma_2)$, where $\Sigma_2 = \lim_{T_2 \to \infty} T_2^{-1} \sum_{t=T_1+1}^{T} \sum_{s=T_1+1}^{T} E(v_{1t} v_{1s})$.

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

Assumptions 1 and 2 imply that $\sqrt{T_1}(\hat{\beta}_{OLS} - \beta_0) \xrightarrow{d} N(0, A^{-1} \Sigma_1 A^{-1})$, 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 essay 1, this assumption ensures that the estimator $\hat{\beta}_{T_1}$ using the pre-treatment data is asymptotically independent with an quantity that involves the post-treatment sample average of the de-mean treatment effects and the idiosyncratic error.

B.2. A projection of the unconstrained estimator

We write the regression model in a matrix form:

$$Y = X\beta_0 + u,$$

where Y and u are both $T_1 \times 1$ vectors, X is of dimension $T_1 \times N$ and has a full column rank, β_0 is of dimension $N \times 1$. We assume that the true parameter $\beta_0 \in \Lambda$, where Λ is a closed and convex set ($\Lambda = \Lambda_{Syn}$ or Λ_{Msyn} in our applications).

We denote the constrained least squares estimator as $\hat{\beta}_{T_1}$, i.e.,

$$\hat{\beta}_{T_1} = \arg\min_{\beta \in \Lambda} (Y - X\beta)'(Y - X\beta) \equiv \arg\min_{\beta \in \Lambda} \|Y - X\beta\|^2,$$

where $||A||^2 = A^{'}A$ for a vector A.

We denote the unconstrained least squares estimator as $\hat{\beta}_{OLS} = arg\min_{\beta \in \mathcal{R}^N}(Y - X\beta)^{'}(Y - X\beta)$, i.e., $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$. By the definition of $\hat{\beta}_{OLS}$, we may write

$$Y = X\hat{\beta}_{OLS} + \hat{u},$$

where $\hat{u} = Y - X\hat{\beta}_{OLS}$. It follows that

$$\begin{aligned} f(\beta) &\stackrel{def}{=} \|Y - X\beta\|^2 \\ &= \|X(\hat{\beta}_{OLS} - \beta) + \hat{u}\|^2 \\ &= \|X(\hat{\beta}_{OLS} - \beta)\|^2 + 2\hat{u}^{'}X(\hat{\beta}_{OLS} - \beta) + \|\hat{u}\|^2 \\ &= \|X(\hat{\beta}_{OLS} - \beta)\|^2 + \|\hat{u}\|^2 \\ &\equiv (\hat{\beta}_{OLS} - \beta)^{'}X^{'}X(\hat{\beta}_{OLS} - \beta) + \|\hat{u}\|^2, \end{aligned} \tag{B.1}$$

where the fourth equality follows from $\hat{u}^{'}X = 0$ (least squares residual $\hat{u}$ is orthogonal to X).

Since $\|\hat{u}\|^2$ is unrelated to β, the minimizer of $f(\beta)$ is identical to the minimizer of $(\hat{\beta}_{OLS} - \beta)^{'}X^{'}X(\hat{\beta}_{OLS} - \beta)$. Thus, we have

$$\begin{aligned} \hat{\beta}_{T_1} &= \arg\min_{\beta \in \Lambda}(\hat{\beta}_{OLS} - \beta)^{'}X^{'}X(\hat{\beta}_{OLS} - \beta) \\ &= \arg\min_{\beta \in \Lambda}(\hat{\beta}_{OLS} - \beta)^{'}(X^{'}X/T_1)(\hat{\beta}_{OLS} - \beta) \\ &= \arg\min_{\beta \in \Lambda}||\hat{\beta}_{OLS} - \beta||_X^2, \end{aligned}$$

where the second equality follows since $T_1 > 0$.

B.3. The uniqueness of the (modified) synthetic control estimator

We first give the definition of a strictly convex function. A function f is said to be *strictly convex* if $f(\alpha x + (1 - \alpha)y)) < \alpha f(x) + (1 - \alpha)f(y)$ for all $0 < \alpha < 1$ and for all $x \neq y$,

$x, y \in D$, where D is the domain of f.

Under the assumption that the data matrix $X_{T_1 \times N}$ has a full column rank, we show below that $f(\beta) \stackrel{def}{=} \sum_{t=1}^{T_1}(y_{1t} - x_t'\beta)^2$ is a strictly convex function. Since the objective function is a convex function and the constrained domains for β, Λ_{Syn} and Λ_{Msyn}, are convex sets, then the constrained minimization problem has an unique (global) minimizer. To see this, we argue by contradiction. Suppose that we have two local minimizers $z_1 \neq z_2$. Then for any convex combination $z_3 = \alpha z_1 + (1-\alpha)z_2$, we have $f(z_3) < \alpha f(z_1) + (1-\alpha)f(z_2)$ for all $\alpha \in (0,1)$. This contradicts the fact that z_1 and z_2 are two minimizers. Hence, we must have $z_1 = z_2$ and the minimizer is unique.

It remains to show that $f(\beta) = (\hat{\beta}_{OLS} - \beta)'X'X(\hat{\beta}_{OLS} - \beta)$ is a strictly convex function (we ignore the irrelevant constant term $\|\hat{u}\|^2$ in $f(\beta)$ defined in (B.1)). We first establish an intermediate result. For $\beta, \gamma \in \mathcal{R}^N$ with $\beta \neq \gamma$, because $A \equiv X'X$ is positive definite, we have

$$\begin{aligned} 0 &< (\beta - \gamma)' A(\beta - \gamma) \\ &= ((\beta - \hat{\beta}_{OLS}) - (\gamma - \hat{\beta}_{OLS}))' A((\beta - \hat{\beta}_{OLS}) - (\gamma - \hat{\beta}_{OLS})) \\ &= (\beta - \hat{\beta}_{OLS})'A(\beta - \hat{\beta}_{OLS}) + (\gamma - \hat{\beta}_{OLS})'A(\gamma - \hat{\beta}_{OLS}) \\ &\quad -2(\beta - \hat{\beta}_{OLS})'A(\gamma - \hat{\beta}_{OLS}) \\ &= f(\beta) + f(\gamma) - 2(\hat{\beta}_{OLS} - \beta)' A(\hat{\beta}_{OLS} - \gamma). \end{aligned} \tag{B.2}$$

Then for all $\alpha \in (0,1)$, we have

$$\begin{aligned} f(\alpha\beta+(1-\alpha)\gamma) &= (\hat{\beta}_{OLS}-(\alpha\beta+(1-\alpha)\gamma))' A(\hat{\beta}_{OLS}-(\alpha\beta+(1-\alpha)\gamma)) \\ &= (\alpha(\hat{\beta}_{OLS}-\beta)+(1-\alpha)(\hat{\beta}_{OLS}-\gamma))' A(\alpha(\hat{\beta}_{OLS}-\beta)+(1-\alpha)(\hat{\beta}_{OLS}-\gamma)) \\ &= \alpha^2(\hat{\beta}_{OLS}-\beta)' A(\hat{\beta}_{OLS}-\beta)+(1-\alpha)^2(\hat{\beta}_{OLS}-\gamma)' A(\hat{\beta}_{OLS}-\gamma) \\ &\quad +2\alpha(1-\alpha)(\hat{\beta}_{OLS}-\beta)' A(\hat{\beta}_{OLS}-\gamma) \\ &= \alpha^2 f(\beta)+(1-\alpha)^2 f(\gamma)+2\alpha(1-\alpha)(\hat{\beta}_{OLS}-\beta)' A(\hat{\beta}_{OLS}-\gamma) \\ &< \alpha^2 f(\beta)+(1-\alpha)^2 f(\gamma)+\alpha(1-\alpha)[f(\beta)+f(\gamma)] \\ &= \alpha f(\beta)+(1-\alpha)f(\gamma), \end{aligned} \tag{B.3}$$

where the above inequality follows from (B.2). Equation (B.3) shows that $f(\cdot)$ is a strictly convex function.

Appendix C: Two useful lemmas

C.1. Two useful lemmas

In this appendix we prove two lemmas that are used to prove Theorem 4.3.2.

Lemma C.1 *Under the same conditions as in Theorem 4.3.2, we have*

$$\hat{\beta}_{T_1} = \tilde{\beta}_{T_1} + o_p(T_1^{-1/2}) = \Pi_\Lambda \hat{\beta}_{OLS} + o_p(T_1^{-1/2}).$$

Proof: For any fixed $\epsilon > 0$, suppose that $\sqrt{T_1}\|\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}\| > \epsilon$, then we have

$$\sqrt{T_1}(\hat{\beta}_{T_1}-\hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1}-\hat{\beta}_{OLS}) < \sqrt{T_1}(\tilde{\beta}_{T_1}-\hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\tilde{\beta}_{T_1}-\hat{\beta}_{OLS}), \tag{C.1}$$

where the strict inequality is due to uniqueness of the projection and the assumption that $\epsilon > 0$ which implies that $\hat{\beta}_{T_1} \neq \tilde{\beta}_{T_1}$. By simple algebra (adding/subtracting terms), we

have:

$$\begin{aligned}
& \sqrt{T_1}(\hat{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1} - \hat{\beta}_{OLS}) \\
= \; & \sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1} + \tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1} + \tilde{\beta}_{T_1} - \hat{\beta}_{OLS}) \\
= \; & \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS}) \\
& + \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})'(X'X/T_1)\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}) \\
& + 2\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}). \qquad \text{(C.2)}
\end{aligned}$$

By (C.1) and (C.2) we know that the sum of the last two terms in (C.2) is negative, i.e.,

$$\begin{aligned}
D_{T_1} \; \stackrel{def}{=} \; & \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})' \left(\frac{1}{T_1}X'X\right) \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}) \\
& + 2\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})' \left(\frac{1}{T_1}X'X\right) \sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) \\
\equiv \; & D_{1,T_1} + D_{2,T_1} < 0. \qquad \text{(C.3)}
\end{aligned}$$

Let $\mathcal{S}^N = \{a \in \mathcal{R}^N : \|a\| = 1\}$ denote the unit sphere in $\mathcal{R}^N$, we have:

$$\begin{aligned}
D_{1,T_1} &= \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})' \left(\frac{1}{T_1}X'X\right) \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}) \\
&= \|\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})\|^2 \left[\frac{\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})'}{\|\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})\|} \left(\frac{1}{T_1}X'X\right) \frac{\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})}{\|\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{T_1})\|}\right] \\
&\geq T_1\|\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}\|^2 \inf_{a \in \mathcal{S}^N} a' \left(\frac{1}{T_1}X'X\right) a \\
&= T_1\|\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}\|^2 \lambda_{\min} \left(\frac{1}{T_1}X'X\right) \\
&\geq \epsilon^2 \lambda_{\min} \left(\frac{1}{T_1}X'X\right) \\
&\stackrel{p}{\to} \epsilon^2 \lambda_{\min}[E(X_t X_t')] > 0 \; , \qquad \text{(C.4)}
\end{aligned}$$

because $\sqrt{T_1}\|\tilde{\beta}_{T_1} - \hat{\beta}_{T_1}\| \geq \epsilon$ and $E(X_t X_t')$ is nonsingular, where $\lambda_{\min}(A)$ denotes the minimum eigenvalue of a square matrix A, the third equality used Lemma C.2 which is

proved at the end of this appendix.

By writing $(X'X/T_1) = E(X_tX_t') + (X'X/T_1) - E(X_tX_t')$, the second term in (C.3) can be rewritten as:

$$\begin{aligned} D_{2,T_1} &= 2\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1)\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) \\ &= 2\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'[E(X_tX_t')]\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) \\ &\quad +2\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'(X'X/T_1 - E[X_tX_t'])\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) \\ &= D_{2,1,T_1} + D_{2,2,T_1}. \end{aligned} \tag{C.5}$$

By the definition of $\tilde{\beta}_{T_1}$ and Lemma 1.1 in Zarantonello (1971) (page 239),

$$D_{2,1,T_1} = \sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS})'[E(X_tX_t')]\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) \geq 0\ . \tag{C.6}$$

By a law of large numbers, $X'X/T_1 - E(X_tX_t') = o_p(1)$. Also, $\sqrt{T_1}(\tilde{\beta}_{T_1} - \hat{\beta}_{OLS}) = O_p(1)$ and $\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}) = O_p(1)$ because:

$$\begin{aligned} \|\sqrt{T_1}(\hat{\beta}_{T_1} - \tilde{\beta}_{T_1})\| &\leq \|\sqrt{T_1}(\hat{\beta}_{T_1} - \beta_0)\| + \|\sqrt{T_1}(\tilde{\beta}_{T_1} - \beta_0)\| \\ &= \|\sqrt{T_1}(\Pi_{\Lambda,T_1}\hat{\beta}_{OLS} - \beta_0)\| + \|\sqrt{T_1}(\Pi_{\Lambda}\hat{\beta}_{OLS} - \beta_0)\| \\ &\leq \sqrt{T_1}\|\hat{\beta}_{OLS} - \beta_0\|_{T_1} + \sqrt{T_1}\|\hat{\beta}_{OLS} - \beta_0\| = O_p(1)\ , \end{aligned}$$

where we used the Lipschitz continuity of projection operators Zarantonello (1971) (page 241), first display in equation (1.8)), and Π_{Λ,T_1} is the projection onto Λ with respect to the aforementioned random norm $\|a\|_{T_1} = \sqrt{a'(X'X/T_1)a}$. Hence, we have $D_{2,2,T_1} = o_p(1)$. Combining $D_{2,2,T_1} = o_p(1)$ and (C.6), we obtain

$$D_{2,T_1} \geq o_p(1). \tag{C.7}$$

Thus, we have shown that: if $\sqrt{T_1}\|\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}\| > \epsilon$, then $D_{T_1} < 0$, which implies that (if A implies B, then $P(A) \leq P(B)$, this argument is used twice in (C.8) below)

$$\begin{aligned} P(\sqrt{T_1}\|\hat{\beta}_{T_1} - \tilde{\beta}_{T_1}\| > \epsilon) &\leq P(D_{T_1} < 0) \\ &\leq P(o_p(1) + \epsilon^2 \lambda_{\min}\left(\frac{1}{T_1}X'X\right) < 0) \\ &\to P(\epsilon^2 \lambda_{\min}\left(E(X_t X_t')\right) \leq 0) \\ &= 0, \end{aligned} \tag{C.8}$$

where the second inequality above follows from $D_{T_1} = D_{1,T_1} + D_{2,T_1} \geq \epsilon^2 \lambda_{min}(X'X/T_1) + o_p(1)$ by (C.4) and (C.7), hence, $D_{T_1} < 0$ implies $\epsilon^2 \lambda_{min}(X'X/T_1) + o_p(1) < 0$.

Equation (C.8) is equivalent to $\hat{\beta}_{T_1} - \tilde{\beta}_{T_1} = o_p(T_1^{-1/2})$, or

$$\hat{\beta}_{T_1} = \Pi_\Lambda \hat{\beta}_{OLS} + o_p(T_1^{-1/2}) \ . \tag{C.9}$$

This finishes the proof of Lemma C.1.

Lemma C.2 *Let A be an $N \times N$ positive definite matrix, $\mathcal{S}^N = \{a \in \mathcal{R}^N : \|a\| = 1\}$ denote the unit sphere in $\mathcal{R}^N$, then we have* $\inf_{a \in \mathcal{S}^N} a'Aa = \lambda_{min}(A)$.

Proof: Let $v_1, ..., v_N$ be N eigen-vectors of A with corresponding eigen-values $\lambda_1, ..., \lambda_N$ so that $Av_j = \lambda_j v_j$ for $j = 1, ..., N$. Then since $v_1, ..., v_N$ form an orthonormal base for $\mathcal{S}^N$, we have for any $a \in \mathcal{S}^N$, $a = \sum_{i=1}^N c_i v_i$ with $\sum_{i=1}^N c_i^2 = 1$ since $a'a = 1$ and $v_i'v_j = \delta_{ij}$ (the Kronecker delta). Then we have

$$\begin{aligned} a'Aa &= \sum_{i=1}^N \sum_{j=1}^N c_i v_i' A c_j v_j = \sum_{i=1}^N \sum_{j=1}^N c_i v_i' c_j A v_j \\ &= \sum_{i=1}^N \sum_{j=1}^N c_i c_j \lambda_j v_i' v_j = \sum_{i=1}^N \lambda_j c_j^2 \\ &\geq \lambda_{min} \sum_{j=1}^N c_j^2 = \lambda_{min}, \end{aligned} \tag{C.10}$$

which implies (i) $\inf_{a\in\mathcal{S}^N} a'Aa \geq \lambda_{min}$.

On the other hand, pre-multiplying $Av_j = \lambda_j v_j$ by v_j', we get $\lambda_j = v_j'Av_j \geq \inf_{a\in\mathcal{S}^N} a'Aa$ for all $j = 1, ..., N$, which implies (ii) $\lambda_{min} \geq \inf_{a\in\mathcal{S}^N} a'Aa$. Combining (i) and (ii) we finish the proof of Lemma C.2.

Appendix D: Asymptotic theory with trend stationary data

D.1. Asymptotic theory with trend stationary data

The trend-stationary data generating process can also be motivated using a factor model framework. Let $\{y_{it}^0\}$, for $i = 1, ..., N$ and $t = 1, ..., T$, be generated by some common factors with one of the factor being a time trend and the remaining factors being weakly dependent stationary variables. Following Hsiao et al. (2012) we assume that $y_t^0 = (y_{1t}^0, y_{2t}^0, ..., y_{Nt}^0)'$ is generated via a factor model

$$y_t^0 = \delta_0 + Bf_t + \epsilon_t, \tag{D.1}$$

where $\delta_0 = (\delta_{01}, ..., \delta_{0N})'$ is an $N \times 1$ vector of intercepts, B is an $N \times K$ factor loading matrix, $f_t = (f_{1t}, ..., f_{Kt})'$ is a $K \times 1$ vector of common factors, $\epsilon_t = (\epsilon_{1t}, ..., \epsilon_{Nt})'$ is an $N \times 1$ vector of idiosyncratic errors. We assume that $f_{1t} = t$ and all other factors are stationary variables. Also, ϵ_t is a zero mean, weakly dependent stationary process with finite fourth moment. Hence, y_t^0 follows a trend-stationary process.

Hsiao et al. (2012), and essay 1 show that, under the condition that $rank(B) = K$, one can replace the unobservable factor f_t by $x_t = (1, y_{2t}, ..., y_{Nt})'$ to estimate the counterfactual outcome y_{1t}^0. Specifically, one can estimate the following regression model

$$y_{1t} = x_t'\delta + u_{1t}, \qquad (t = 1, ..., T_1), \tag{D.2}$$

where $x_t = (1, y_{2t}, ..., y_{Nt})'$ and $\delta = (\delta_1, ..., \delta_N)'$.

To facility the asymptotic analysis, below we consider the time trend component explicitly.

We write $y_{jt} = c_{0,j} + c_{1,j}t + \eta_{jt}$, where η_{jt} is a weakly dependent stationary process (detrended from y_{jt}) for $j = 2, ..., N$. Let $\tilde{y}_t = (y_{2t}, ..., y_{Nt})'$ and $\tilde{\delta} = (\delta_2, ..., \delta_N)'$. Then in vector notation, we have $\tilde{y}_t = \tilde{c}_0 + \tilde{c}_1 t + \tilde{\eta}_t$, $\tilde{c}_0 = (c_{0,2}, ..., c_{0,N})'$, $\tilde{c}_1 = (c_{1,2}, ..., c_{1,N})'$ and $\tilde{\eta} = (\eta_{2t}, ..., \eta_{Nt})'$. Then we can write $\tilde{y}_t'\tilde{\delta} = (\tilde{c}_0 + \tilde{c}_1 t + \tilde{\eta}_t)'\tilde{\delta}$. Hence, we can re-write (D.2) as

$$\begin{aligned} y_{1t} &= \delta_1 + \tilde{y}_t'\tilde{\delta} + u_{1t} \\ &= \alpha_0 t + \beta_1 + \tilde{\delta}'\tilde{\eta}_t + u_{1t} \\ &= \alpha_0 t + z_t'\beta_0 + u_{1t} \qquad\qquad t = 1, ..., T_1, \end{aligned} \tag{D.3}$$

where $\alpha_0 = \tilde{c}_1'\tilde{\delta}$, $\beta_1 = \delta_1 + \tilde{c}_0'\tilde{\delta}$, $\beta_0 = (\beta_1, \tilde{\delta}')'$ and $z_t = (1, \tilde{\eta}')' \equiv (1, \eta_{2t}, ..., \eta_{Nt})'$.

Below we derive the asymptotic distribution of the ATE estimator $\hat{\Delta}_1$ defined in (4.3.7). For the post-treatment period, we have $y_{1t}^1 = y_{1t}^0 + \Delta_{1t}$. Hence, we have for $t = 1, ..., T$,

$$y_{1t} = \alpha t + z_t'\beta + d_t\Delta_{1t} + v_{1t}, \tag{D.4}$$

where $d_t = 0$ for $t \leq T_1$ and $d_t = 1$ for $t \geq T_1 + 1$.

Let $\hat{\alpha}$ and $\hat{\beta}$ be the least squares estimators of α and β based on (D.3). Then it is well established that (e.g., Hamilton (1994), Chapter 16) $\hat{\alpha} - \alpha = O_p(T_1^{-3/2})$ and $\hat{\beta} - \beta = O_p(T_1^{-1/2})$. Thus, using (4.3.7) and (D.4) we have

$$\begin{aligned} \hat{\Delta}_1 - \Delta_1 &= \frac{1}{T_2}\sum_{t=T_1+1}^{T}\left[y_{1t} - \hat{y}_{1t}^0\right] - \Delta_1 \\ &= -\frac{1}{T_2}\sum_{t=T_1+1}^{T}\left[(\hat{\alpha}_{T_1} - \alpha_0)t - z_t'(\hat{\beta}_{T_1} - \beta_0) + \Delta_{1t} - \Delta_1 + v_{1t}\right] \\ &= -\left[\frac{2T_1 + T_2 + 1}{2}\right](\hat{\alpha} - \alpha) - [E(z_t') + o_p(1)](\hat{\beta} - \beta) + \frac{1}{T_2}\sum_{t=T_1+1}^{T} v_{1t}, \end{aligned} \tag{D.5}$$

where we used $\sum_{t=T_1+1}^{T} t = (T_1 + 1 + T)T_2/2 = (2T_1 + T_2 + 1)T_2/2$, $v_{1t} = \Delta_{1t} - \Delta_1 + u_{1t}$.

Hence,

$$\begin{aligned}\sqrt{T_2}(\hat{\Delta}_1-\Delta_1) &= -\sqrt{T_2/T_1}\left[\frac{2+T_2/T_1}{2}\right]\sqrt{T_1^3}(\hat{\alpha}_{T_1}-\alpha_0)-\sqrt{T_2/T_1}E(z_t')\sqrt{T_1}(\hat{\beta}_{T_1}-\beta_0)\\ &\quad+\frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T}v_{1t}+o_p(1)\\ &\quad+\frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T}v_{1t}+o_p(1)\\ &= -c'M_{T_1}(\hat{\gamma}_{T_1}-\gamma_0)+\frac{1}{\sqrt{T_2}}\sum_{t=T_1+1}^{T}v_{1t}+o_p(1), \qquad \text{(D.6)}\end{aligned}$$

where $c=(\sqrt{\phi}(2+\phi)/2,\sqrt{\phi}E(z_t'))'$, $\phi=\lim_{T_1,T_2\to\infty}T_2/T_1$, $\hat{\gamma}_{T_1}=(\hat{\alpha}_{T_1},\hat{\beta}_{T_1}')'$ and $\gamma_0=(\alpha_0,\beta_0')'$, $M_{T_1}=\sqrt{T_1}diag(T_1,1,...,1)$ which is an $(N+1)\times(N+1)$ diagonal matrix with the first diagonal element equals to $T_1^{3/2}$ and all other diagonal elements equal to $\sqrt{T_1}$.

To establish the asymptotic distribution of $\sqrt{T_2}(\hat{\Delta}_1-\Delta_1)$, we make the following assumptions.

Assumption D1. Let $z_t=(1,\eta_{2t},...,\eta_{Nt})'$. We assume that (i) $\{z_t\}_{t=1}^{T}$ is a weakly dependent and weakly stationary process, $T_1^{-1}\sum_{t=1}^{T_1}z_tz_t'\xrightarrow{p}E(z_tz_t')$ as $T_1\to\infty$, and $[E(z_tz_t')]$ is invertible; (ii) $M_{T_1}(\hat{\gamma}_{OLS}-\gamma)\xrightarrow{d}N(0,\Omega)$, where Ω is a positive definite matrix.

Assumption D2. Let $v_{1t}=\Delta_{1t}-\Delta_1+v_{1t}$. Then $T_2^{-1/2}\sum_{t=T_1+1}^{T}v_{1t}\xrightarrow{d}N(0,\Sigma_2)$ as $T_2\to\infty$, where $\Sigma_2=\lim_{T_2\to\infty}T_2^{-1}\sum_{t=T_1+1}^{T}\sum_{s=T_1+1}^{T}E(v_{1t}v_{1s})$.

Assumption D3. Let $w_t=(v_{1t},\eta_{2t},...,\eta_{Nt})'$. We assume that w_t is a ρ-mixing process with the mixing coefficient $\rho(\tau)$ satisfies the condition: $\rho(\tau)\le C\,\lambda^{\tau}$ for some finite positive constants $C>0$ and $0<\lambda<1$, where $\rho(\tau)=\max_{1\le i,j\le N}\frac{|Cov(w_{it},w_{j,t+\tau})|}{\sqrt{Var(w_{it})Var(w_{j,t+\tau})}}$, w_{it} is the i^{th} component of w_t for $i=1,...,N$.

Assumptions D1 and D2 are not restrictive. They require that (z_t,v_{1t}) to be a weakly dependent stationary process so that law of large numbers and central limit theorem hold for their (partial) sums. If $E(z_tz_t')$ is not invertible, we can remove the linearly dependent

regressors and redefine z_t as a subset of $(1, \eta_{2t}, ..., \eta_{Nt})'$ such that assumption 1 holds. Assumption D3 further imposes an exponential decay rate for the ρ-mixing processes. Many ARMA processes are known to be ρ-mixing with exponential decay rate.

By Assumption D3 and the proof of Theorem 3.2 and Lemma 1 in essay 1, we know that $\hat{\gamma} - \gamma$ is asymptotic independent with $T_2^{-1/2} \sum_{t=T_1+1}^{T} v_{1t}$. Therefore, applying the projection theory to (D.6) we immediately have the following result.

Under assumptions D1 to D3 and note that $\gamma_0 \in \Lambda$, we have

$$\begin{aligned}
\sqrt{T_2}(\hat{\Delta}_1 - \Delta_1) &= -c' M_{T_1}(\hat{\gamma}_{T_1} - \gamma_0) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t}) \\
&= -c' M_{T_1}(\Pi_\Lambda \hat{\gamma}_{OLS} - \Pi_\Lambda \gamma_0) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} \\
&= -c' \Pi_{T_{\Lambda,\gamma_0}} M_{T_1}(\hat{\gamma}_{OLS} - \gamma_0) + \frac{1}{\sqrt{T_2}} \sum_{t=T_1+1}^{T} v_{1t} + o_p(1) \\
&\xrightarrow{d} -c' \Pi_{T_{\Lambda,\gamma_0}} Z_3 + Z_2, \qquad \text{(D.7)}
\end{aligned}$$

where Z_3 is the weak limit of $M_{T_1}(\hat{\gamma}_{OLS} - \gamma_0)$ as described in Assumption C1, Z_2 is independent with Z_3, and is normally distributed with a zero mean and variance Σ_2.

Appendix E: Explanation of subsampling method works for a wide range of subsample sizes

In this appendix, we explain why the subsampling method works well for our estimated ATE estimator for a wide range of subsample size m values.

E.1. A simple example from Andrews (2000)

We consider a simple example as considered in Andrews (2000), where Y_i, for $i = 1, ..., n$, is iid $N(\mu_0, 1)$ with $\mu_0 \geq 0$, i.e., $Y_i = \mu_0 + u_i$ with u_i iid $N(0, 1)$ and that $\mu_0 \in \Lambda = \mathcal{R}^+ \stackrel{def}{=} \{y : y \geq 0\}$. The constrained least squares estimator of μ_0 is $\hat{\mu}_n = \max\{\bar{Y}_n, 0\}$, where

$\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$. It is easy to show that

$$\hat{S}_n \stackrel{def}{=} \sqrt{n}(\hat{\mu}_n - \mu_0) \stackrel{d}{\to} \begin{cases} Z & \text{if } \mu_0 > 0 \\ \max\{Z, 0\} & \text{if } \mu_0 = 0, \end{cases} \tag{E.1}$$

where Z denotes a standard normal random variable. Let Y_i^* be random draws from $\{Y_j\}_{j=1}^n$, then a bootstrap analogue of $(E.1)$ is $\sqrt{n}(\tilde{\mu}_n^* - \hat{\mu}_n)$, where $\hat{\mu}_n^* = \max\{\bar{Y}_n^*, 0\}$, where $\bar{Y}_n^* = n^{-1}\sum_{i=1}^n Y_i^*$. Andrews (2000) shows that this standard resampling bootstrap method as well as several parametric bootstrap methods do not work in the sense that, when $\mu_0 = 0$,, $\tilde{S}_n^* = \sqrt{n}(\tilde{\mu}_n^* - \hat{\mu}_n)$ will not converge to $\max\{Z, 0\}$, the limiting distribution of $\hat{S}_n$. In fact, Andrews (2000) shows that $\hat{S}_n^*$ converges to a distribution that is to the left of $\max\{Z, 0\}$.

Andrews (2000) also suggests a few re-sampling methods that overcome the problem. One particular easy-to-implement method is a parametric subsampling method. Specifically, for m satisfies that $m \to \infty$ and $m/n \to 0$ as $n \to \infty$, one can use $\tilde{S}_m^* = \sqrt{m}(\hat{\mu}_m^* - \hat{\mu}_n)$ to approximate the distribution of $\sqrt{n}(\hat{\mu}_n - \mu_0)$, where $\hat{\mu}_m = \max\{\bar{Y}_m^*, 0\}$, $\bar{Y}_m^* = m^{-1}\sum_{i=1}^m Y_i^*$ with Y_i^* is iid draws from $N(\bar{Y}_n, 1)$, i.e., $Y_i^* = \bar{Y}_n + u_i^*$ with u_i^* iid $N(0,1)$. To see that the subsampling method indeed works, we have that, conditional on $\{Y_i\}_{i=1}^n$,

$$\begin{aligned}
\hat{S}_m^* &\stackrel{def}{=} \sqrt{m}(\hat{\mu}_m^* - \hat{\mu}_n) \\
&= \max\left\{\sqrt{m}\,\bar{Y}_m^*, 0\right\} - \sqrt{m}\,\hat{\mu}_n \\
&= \max\left\{\sqrt{m}\,\bar{Y}_m^*, 0\right\} - \sqrt{m}\,\mu_0 - \sqrt{m}(\hat{\mu}_n - \mu_0) \\
&= \max\left\{\sqrt{m}(\bar{Y}_m^* - \bar{Y}_n + \bar{Y}_n - \mu_0), -\sqrt{m}\,\mu_0\right\} - \sqrt{m}(\hat{\mu}_n - \mu_0) \\
&= \max\left\{\sqrt{m}(\bar{Y}_m^* - \bar{Y}_n) + \sqrt{m/n}\sqrt{n}(\bar{Y}_n - \mu_0), -\sqrt{m}\,\mu_)\right\} - \sqrt{m/n}\sqrt{n}(\hat{\mu}_n - \mu_0) \\
&= \max\left\{\sqrt{m}(\bar{Y}_m^* - \bar{Y}_n) + o_p(1), -\sqrt{m}\,\mu_0\right\} + o_p(1) \\
&\stackrel{d}{\to} \begin{cases} Z & \text{if } \mu_0 > 0 \\ \max\{Z, 0\} & \text{if } \mu_0 = 0, \end{cases}
\end{aligned} \tag{E.2}$$

where the second equality follows from the definition of $\hat{\mu}_m^*$, we add/subtract $\sqrt{m}\,\mu_0$ at the third equality, the fourth equality follows from $\max\{a,b\}-c=\max\{a-c,b-c\}$, the sixth equality follows from $m/n=o(1)$, $\sqrt{n}(\bar{Y}_n-\mu_0)=O_p(1)$ and $o(1)O_p(1)=o_p(1)$. The last equality follows from the fact that $Y_i^*-\bar{Y}_n=u_i^*$ is iid $N(0,1)$. Hence, $\sqrt{m}(\hat{Y}_m^*-\bar{Y}_n)\overset{d}{\sim} N(0,1)\equiv Z$ for any value of m. If $\{Y_i^*\}_{i=1}^m$ is iid with mean $\bar{Y}_n$ and unit variance but is not normally distributed, then we need m to be large so that $\sqrt{m}(\hat{Y}_m^*-\bar{Y}_n)\overset{d}{\to} N(0,1)\equiv Z$ by virtue of a central limit theorem argument (as $m\to\infty$).

Comparing (E.1) and (E.2), we see that subsampling method works under very mild conditions that $m\to\infty$ and $m/n\to 0$ as $n\to\infty$.

E.2. Testing zero ATE by subsampling method

We conduct simulations to examine the finite sample performances of the subsampling method. We generate Y_i iid $N(0,1)$ (i.e., $\mu_0=0$) for $i=1,...,n$ and we choose $n=100$ and conduct 5000 simulations. Within each simulation, we generate 2000 subsampling samples with subsample sizes $m\in\{5,10,20,30,50,100\}$. Note that we select the largest $m=n=100$ because we want to show numerically that the standard bootstrap method does not work. For each fixed value m, we sort the 2000 subsampling statistics in an ascending order such that $\hat{S}_{m,(1)}^*\le\hat{S}_{m,(2)}^*\le...\le\hat{S}_{m,(2000)}^*$, then we get right-tail α-percentile value by $\hat{S}_{((1-\alpha)(2000))}^*$. We record rejection rate as the percentage that $\hat{S}$ is greater or equal to $\hat{S}_{((1-\alpha)(2000))}^*$ for $\alpha\in\{0.01,0.05,0.1,0.2\}$. We consider two cases: (i) We generate Y_i iid $N(0,1)$ and $Y_i^*=\bar{Y}_n+v_i$ with v_i iid $N(0,1)$; and (ii) We generate Y_i uniformly distributed over $[-\sqrt{3},\sqrt{3}]$ (so that it has zero mean and unit variance) and $Y_i^*=\bar{Y}_n+v_i$ with v_i iid uniformly distributed over $[-\sqrt{3},\sqrt{3}]$. The results for the two cases are almost identical. To save space we only report the normally distributed v_i case in Table 15.

First, we see that the subsampling method with $5\le m\le 20$ seem to work well. Second, we see clearly that using $m=n$ or m close to n ($m\ge 50$) do not work. For example, when $m=n$, it gives estimated rejection rates double that of the nominal levels. Andrews (2000)

Table 15: Estimated sizes ($Y_i^* \sim N(\bar{Y}_n, 1)$)

	m=5	m=10	m=20	m=30	m=50	m=100
1%	.0132	. 0126	.0124	.0130	.0136	.0248
5%	.0516	.0518	.0518	.0532	.0658	.1032
10%	.0960	.0968	.1006	.1104	.1346	.2014
20%	.1936	.2004	.2278	.2588	.3164	.4020

shows that the distribution of $\sqrt{n}(\hat{\mu}_n^* - \hat{\mu}_n)$ is to the left of that of $\sqrt{n}(\hat{\mu}_n - \mu_0)$. Hence, the bootstrap method will lead to over rejection of the null hypothesis. Our simulation results verifies Andrews' (2000) theoretical analysis.

The simulation results seem to be in contradiction to the simulation results reported in Section 4.5 where even for $m = n$, the subsampling method seems to be fine. We explain the seemingly contradictory result in the next subsection.

E.3. Not all parameters are at the boundary

Our simulations reported in Section 4.5 corresponds to the case of $\beta_{0,j} > 0$ for $j = 2, ..., 7$ and $\beta_{0,j} = 0$ for $j = 8, ..., 11$. The constrained estimators $\hat{\beta}_{T_1,j}$ ($\hat{\beta}^*_{m,j}$) for $j = 8, 9, 10, 11$ can cause problems for the standard bootstrap method not to work. However, notice that our ATE estimator also depends on $\hat{\beta}_{T_1,j}$ ($\hat{\beta}^*_{m,j}$) for $j = 1, ..., 7$, which does not take boundary value 0. This helps to improve subsampling method for large value of m. More importantly, our ATE estimator also contains a term not related to $\hat{\beta}_{T_1}$ (see the second term at the right hand side of (4.4.5)) and the existence of this term further improves the performance of the subsampling method when m is close to or equal to n. This is the reason why in our simulations even when $m = n$ the subsampling method seems to work fine. To numerically verify this conjecture, we generate a sequence of iid $Z_1, Z_2 \sim N(0, \sigma_v^2)$ random variables and add them to $\hat{S}_n$ and $\hat{S}^*_m$, i.e., $\tilde{S}_n = \hat{S}_n + Z_1$ and $\tilde{S}^*_m = \hat{S}^*_m + Z_2$, we then repeat the simulations to compute the estimated sizes. The results for $\sigma_v = 1$ and 5 are reported in Table 16. We observe the performance of the subsampling statistic $\tilde{S}^*_m$ improves significantly over $\hat{S}^*_m$ for $m = 50$ and 100. Consider the case of $\sigma_v = 1$ and $m = n$, the rejection rates based on $\tilde{S}^*_m$ is about 20% higher than that of the nominal levels whereas it was 100% higher than that

of nominal levels based on $\hat{S}_m^*$.

From Table 16 we see that when σ_v^2 is large, Z_1 and Z_2 becomes the dominating components of $\tilde{S}_n$ and $\tilde{S}_m^*$, therefore, the subsampling method works well for all values of m including $m = n$. Also note that the estimated sizes for $\sigma_v^2 = 1$ only slightly over sized compared to $\sigma_v^2 = 25$ shows that the significant improvements in the estimated sizes (over the case of $\sigma_v^2 = 0$) does not require adding a regular component with large dominating variance.

Table 16: Estimated sizes: Adding a $N(0, \sigma_v^2)$ to $\hat{S}_n$ and $\hat{S}_m^*$

	m=5	m=10	m=20	m=30	m=50	m=100
	$\sigma_v = 1$					
1%	.0104	.0110	.0112	.0128	.0122	.0114
5%	.0550	.0562	.0562	.0590	.0600	.0648
10%	.1066	.1098	.1140	.1168	.1198	.1236
20%	.2170	.2244	.2320	.2372	.2440	.2520
	$\sigma_v = 5$					
1%	.0112	.0116	.0116	.0110	.0124	.0128
5%	.0518	.0521	.0528	.0530	.0542	.0556
10%	.1030	.1044	.1046	.1048	.1060	.1074
20%	.2070	.2082	.2030	.2102	.2126	.2160

Appendix F: Inferences when T_2 is small

F.1. Inferences when T_2 is small

In this section, we consider case of large $T_1 = 100$, 200 and small $T_2 = 3$, 5. We use Andrews (2003) end-of-sample instability test discussed in Section 4.4.2 to test the null hypothesis H_0: $\Delta_{1t} = 0$ ($\Delta_{1,0} = 0$) against the one-sided alterative H_1: $\Delta_{1t} > 0$ for all $t = T_1 + 1, ..., T$. The data is generated by the three-factor model (DGP1) as discussed in Section 4.5.1, and the treatment effects is generated via (4.5.2) with $\alpha_0 = 0$ under H_0, and $\alpha_0 = 0.5$, 1 under H_1. The number of simulation is 10,000. The simulations results are reported in Table 17.

Andrews' (2003) test is expected to give good estimated sizes when T_1 is large. As expected, we see from Table 17 that the test is over sized for $T_1 = 100$, its estimated sizes improve as T_1 increase to 200. Another result worth noticing from Table 17 is that, if we fix T_1,

Table 17: Coverage probabilities for DGP1 (Andrews' (2003) instability test)

	H_0: $\alpha_0 = 0$					
	$T_2 = 3$			$T_2 = 5$		
T_1	5%	10%	20%	5%	10%	20%
100	0.0849	0.1362	0.2366	0.0935	0.1497	0.2440
200	0.0652	0.1161	0.2191	0.0711	0.1250	0.2273
	H_1: $\alpha_0 = 0.5$					
	$T_2 = 3$			$T_2 = 5$		
T_1	5%	10%	20%	5%	10%	20%
100	0.2892	0.4076	0.6656	0.3492	0.4753	0.6985
	H_1: $\alpha_0 = 1$					
	$T_2 = 3$			$T_2 = 5$		
T_1	5%	10%	20%	5%	10%	20%
100	0.5416	0.6573	0.7937	0.6994	0.7939	0.8853

the estimated sizes deteriorate as T_2 increases. That is understandable because this test is designed for large T_1 and small T_2.

As Andrews (2003) points out, this statistic is not a consistent test for small values of T_2.[7] Note that a large T_1 helps to give better estimated sizes, it does not increase the power of the test. Therefore, we only consider $T_1 = 100$ for power calculation because for $T_1 = 200$ or even larger T_1, the powers of the test are similar. When T_1 is large, the power of the test increases with T_2 and also depends on the magnitude of $\sum_{t=T_1+1}^{T}(\Delta_{1t} - \Delta_{1,0})$ under H_1. From Table 17 we see that the estimated power increases with T_2 as well as with α_0 (the magnitude of Δ_{1t}). However, a large T_2 adversely affects the estimated sizes of Andrews' (2003) test.

We also conducted simulations of Andrews' (2003) test under DGP1 using $T_1 = 90$ and $T_2 = 20$ (same T_1 and T_2 as in our empirical data). Based on 10,000 simulations with $\alpha_0 = 0$, the estimated sizes are 0.1660 and 0.1964 for nominal levels 5% and 10%, respectively. We see that for $T_2 = 20$, $T_1 = 90$ is not large enough for the test to have good estimated sizes, because an error term of order $\sqrt{T_2/T_1}$ is not negligible which causes Andrews' (2003) test invalid in our context. Therefore, the end-of-sample stability testing and the subsampling

[7] A test is said to be a consistent test if, when the null hypothesis is false, the probability of rejecting the (false) null hypothesis converges to one as sample size goes to infinity ($T_2 \to \infty$).

testing procedures are complement to each other. The former can be used when T_2 is small, while the later is preferred when T_2 is not small.

Remark F.1 *Here for our synthetic control ATE estimator with panel data, large T_2 invalidates Andrews' (2003) test due an error term of order $\sqrt{T_2/T_1}$ becoming non-negligible. This differs from the time series model considered by Andrews (2003), where when T_2 is also large, testing a possible structure break at T_1 becomes a simple and standard problem.*

Appendix G: Additional robust check results

G.1. Adding Covariates

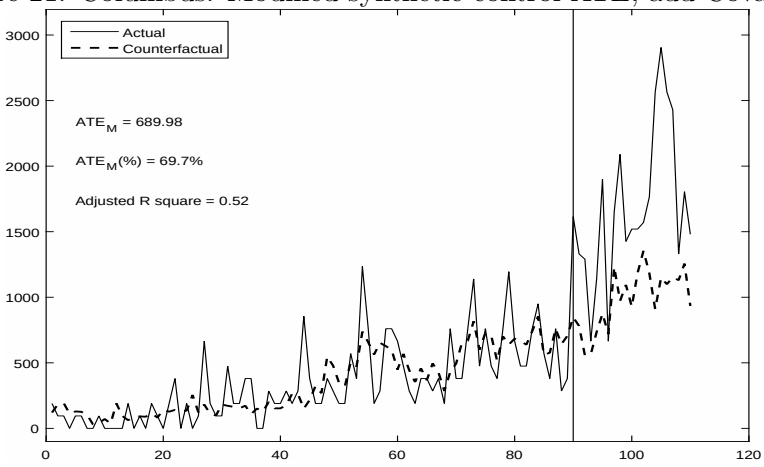
We collect monthly data on unemployment rate (Unemp), labor force (LF) and average weekly earnings (Inc) for Columbus, and linear extrapolate them to weekly data. The data is downloaded from the Bureau of Labor Statistics website (bls.gov). The estimation model is

$$y_{1t} = x_t'\beta_0 + z_{1t}'\gamma_0 + u_{1t}, \qquad t = 1, ..., T_1 \tag{G.1}$$

where $x_t = (1, y_{2t}, ..., y_{Nt})'$, $z_{1t} = (Unemp_t, LF_t, Inc_t)'$, β_0 and γ_0 are $N \times 1$ and 3×1 vector of parameters, respectively. Since obviously that opening a showroom has no (or negligible) effects on z_{1t}, we can use the above model to predict post-treatment counterfactual sale for the treated city. Specifically, we estimate model (G.1) under the restriction $\beta_j \geq 0$ for $j \geq 2$ using the pre-treatment data $t = 1, ..., T_1$ (there is no restriction for other parameters). Let $\hat{\beta}_{T_1}$ and $\hat{\gamma}_{T_1}$ denote the corresponding estimators. We estimate the counterfactual outcome y_{1t}^0 by $\hat{y}_{1t}^0 = x_t'\hat{\beta}_{T_1} + z_{1t}'\hat{\gamma}_{T_1}$ for $t = T_1 + 1, ..., T$ and estimate ATE by $T_2^{-1}\sum_{t=T_1+1}^{T}(y_{1t} - \hat{y}_{1t}^0)$.

Figure 21 plots the estimation result for Columbus. The ATE becomes 69.7% which is quite close to the original result of 67%. However, the adjusted R^2 decreased slightly from 0.528 to 0.520, indicating that the three covariates do not have additional explanatory power to explain sale. The virtually same ATE estimation result with added covariates again supports our original ATE estimation result.

Figure 21: Columbus: Modified synthetic control ATE, add Covariates

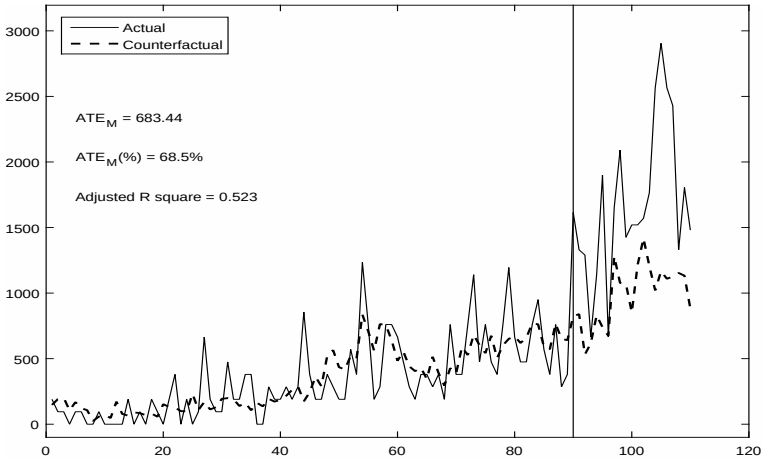


G.2. Select control units based on covariate matching

In this subsection we first select cities whose covariates are close to the covariates of the treated city. Then we select the number of control cities by comparing adjusted R^2. Finally we estimate ATE using the selected control units. We explain this procedure in more details below.

For each $j = 1, 2, 3$ (corresponding to Unemp, LF, Inc), we regress $z_{1,jt}$ on $z_{i,jt}$ using the pre-treatment data and obtain the goodness-of-fit $R^2_{i,j}$ for $i = 2, ..., 11$. We obtain a total R-square for city i by $R^2_i = R^2_{i,1} + R^2_{i,2} + R^2_{i,3}$. We sort them in a non-increasing order: $R^2_{(2)} \geq R^2_{(3)} \geq ... \geq R^2_{(11)}$. Their corresponding sales are denoted by $y_{(2),t}$,...,$y_{(11),t}$ for $t = 1, ..., T_1$. Next, we regress y_{1t} on $y_{(2),t}$ and obtain an adjusted $\bar{R}^2_{(2)}$; and we regress y_{1t} on $(y_{(2),t}, y_{(3),t})$ and obtain an adjusted $\bar{R}^2_{(2),(3)}$; continuing this way until we regress y_{1t} on all $(y_{(2),t}, ..., y_{(11),t})$. We choose a model with the largest adjusted $\bar{R}^2$. For Columbus, this method selects seven cities (Portland, Houston and Atlanta are not selected) gives the largest adjusted $\bar{R}^2$. Using the seven selected cities as control group, the modified synthetic control method's estimation result is plotted in figure 22. The ATE estimation result is 68.5% which is quite close to the original result of 67%. The robust check shows that our ATE estimation result is not sensitive to the selection of different control units.

Figure 22: Columbus: ATE Estimation Based on Covariates Matching



BIBLIOGRAPHY

A. Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19, 2005.

A. Abadie and J. Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(1):113–132, 2003.

A. Abadie, A. Diamond, and J. Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of californias tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

E. T. Anderson, N. M. Fong, D. I. Simester, and C. E. Tucker. How sales taxes affect customer and firm behavior: The role of search on the internet. *Journal of Marketing Research*, 47(2):229–239, 2010.

D. W. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, pages 817–858, 1991.

D. W. K. Andrews. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405, 2000.

D. W. K. Andrews. End-of-sample instability tests. *Econometrica*, 71(6):1661–1694, 2003.

O. Ashenfelter. Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, pages 47–57, 1978.

O. Ashenfelter and D. Card. Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, 67(4):648–60, 1985.

S. Athey and G. W. Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006.

S. Athey and G. W. Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017.

J. Avery, T. J. Steenburgh, J. Deighton, and M. Caravella. Adding bricks to clicks: Predicting the patterns of cross-channel elasticities over time. *Journal of Marketing*, 76(3): 96–111, 2012.

C. Bai, Q. Li, and M. Ouyang. Property taxes and home prices: A tale of two cities. *Journal of Econometrics*, 180(1):1–15, 2014a.

J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

J. Bai, K. Li, et al. Theory and methods of panel data models with interactive effects. *The Annals of Statistics*, 42(1):142–170, 2014b.

D. R. Bell, S. Gallino, and A. Moreno. Offline showrooms in omnichannel retail: Demand and operational benefits. *Management Science*, 2017.

M. Bertrand, E. Duflo, and S. Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275, 2004.

P. J. Bickel and A. Sakov. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, pages 967–985, 2008.

P. J. Bickel, Y. Ritov, A. B. Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

B. J. Bronnenberg, J.-P. H. Dubé, and M. Gentzkow. The evolution of brand preferences: Evidence from consumer migration. *American Economic Review*, 102(6):2472–2508, 2012.

M. Busse, J. Silva-Risso, and F. Zettelmeyer. 1,000 cash back: The pass-through of auto manufacturer promotions. *American Economic Review*, 96(4):1253–1270, 2006.

D. Card and A. B. Krueger. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. *American Economic Review*, 84:772–784, 1994.

M. Carrasco and X. Chen. Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory*, 18(1):17–39, 2002.

V. Chernozhukov, K. Wuthrich, and Y. Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *arXiv preprint arXiv:1712.09089*, 2017.

J. A. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354, 2006.

T. G. Conley and C. R. Taber. Inference with difference in differences with a small number of policy changes. *The Review of Economics and Statistics*, 93(1):113–125, 2011.

S. G. Donald and K. Lang. Inference with difference-in-differences and other panel data. *The Review of Economics and Statistics*, 89(2):221–233, 2007.

N. Doudchenko and G. W. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research, 2016.

Z. Du and L. Zhang. Home-purchase restriction, property tax and housing price in china: A counterfactual analysis. *Journal of Econometrics*, 188(2):558–568, 2015.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Z. Fang and A. Santos. Inference on directionally differentiable functions. *arXiv preprint arXiv:1404.3763*, 2016.

B. Ferman and C. Pinto. Inference in differences-in-differences with few treated groups and heteroskedasticity. 2015.

B. Ferman and C. Pinto. Revisiting the synthetic control estimator. 2016.

C. Forman, A. Ghose, and A. Goldfarb. Competition between local and electronic markets: How the benefit of buying online depends on where you live. *Management science*, 55(1): 47–57, 2009.

M. Forni and L. Reichlin. Let's get real: a factor analytical approach to disaggregated business cycle dynamics. *The Review of Economic Studies*, 65(3):453–473, 1998.

A. Goldfarb and C. E. Tucker. Privacy regulation and online advertising. *Management science*, 57(1):57–71, 2011.

A. W. Gregory and A. C. Head. Common and country-specific fluctuations in productivity, investment, and the current account. *Journal of Monetary Economics*, 44(3):423–451, 1999.

J. Hahn and R. Shi. Synthetic control and inference. *Econometrics*, 5(4):52, 2017.

J. D. Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

H. Hong and J. Li. The numerical delta method and bootstrap. Technical report, Working paper, 2015.

C. Hsiao, S. Ching, and K. S. Wan. 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(5):705–740, 2012.

J. Huang, S. Ma, and C.-H. Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618, 2008.

G. W. Imbens and J. M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

B. Mantin and E. Rubin. Fare prediction websites and transaction prices: Empirical evidence from the airline industry. *Marketing Science*, 35(4):640–655, 2016.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009.

W. K. Newey and K. D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(5):703–708, 1987.

M. Ouyang and Y. Peng. The treatment-effect estimation: A case study of the 2008 economic stimulus package of china. *Journal of Econometrics*, 188(2):545–557, 2015.

O. C. Ozturk, S. Venkataraman, and P. K. Chintagunta. Price reactions to rivals local channel exits. *Marketing Science*, 35(4):588–604, 2016.

J.-S. Pischke. The impact of length of the school year on student performance and earnings: Evidence from the german short school years. *The Economic Journal*, 117(523):1216–1242, 2007.

D. N. Politis, J. P. Romano, and M. Wolf. Subsampling springer series in statistics, 1999.

J. Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494, 1993.

K. Wang and A. Goldfarb. Can offline stores drive online sales? *Journal of Marketing Research*, 54(5):706–719, 2017.

H. White. *Asymptotic theory for econometricians*. Academic press, 1984.

E. H. Zarantonello. 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*, pages 237–424. Elsevier, 1971.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.