Synthetic Control Method and LASSO under Staggered Adoption with Dynamic Treatment Effect

Advisor: Professor Andrii Babii* Student: Yi Niu†

Abstract

It is often hard to analyze the effect of a policy when the policy is staggered-adopted by states, or the treatment effect is dynamic. Researchers usually make strong assumptions on the staggered adoption and the dynamic treatment effect to make evaluation possible. One popular estimation method that is robust to dynamic treatment effect is Callaway and Sant'Anna's Difference-in-Difference (Callaway and Sant'Anna (2021)). This paper identifies the limitations of CS-DID and proposes two new methods/algorithms based on SCM, LASSO, and CS-DID (namely: SCM-CS-DID and LASCM-CS-DID), that are robust to abnormal units in the treatment group and relax the assumptions from CS-DID.

1 Introduction

Synthetic Control Method (SCM) by Abadie et al. (2010) a statistical technique used in the field of causal inference and policy evaluation. It is designed to estimate the causal effect of a treatment or intervention on an individual agent by comparing it to a weighted combination

*Department of Economics, University of North Carolina at Chapel Hill - Gardner Hall, CB 3305 Chapel Hill, NC 27599-3305. Email: babii.andrii@gmail.com

†University of North Carolina at Chapel Hill. Email: n1y1@emai.unc.edu

of control units that did not receive the treatment. Least Absolute Shrinkage and Selection Operator (LASSO) by Tibshirani (1996) is a statistical technique used for linear regression and feature selection. It adds a penalty term to the standard linear regression cost function, which combines the sum of squared residuals with a penalty based on the absolute values of the regression coefficients. SCM and LASSO are similar in construction, shown in Section 4, but they are seldom compared in context of policy evaluation.

Analyzing the effects of policies is often complex when they're adopted at different times or have evolving impacts. One popular estimation method that is robust to dynamic treatment effect is Callaway and Sant'Anna's Difference-in-Difference (Callaway and Sant'Anna (2021)). We list the assumptions and main theorem of Callaway and Sant'Anna (2021) in the Appendix. However, we argue that the assumptions in Callaway and Sant'Anna (2021) could be uneasy to achieve in practice in Section 2. Motivated by this, we discuss and compare SCM and LASSO in context of policy evaluation under staggered adoption with dynamic treatment effect. We come up with two new algorithms that combine SCM with CS-DID and LASSO with SCM with CS-DID (namely SCM-CS-DID and LASCM-CS-DID) which are robust under staggered adoption with dynamic treatment effect. Properties of our methods are supported by two Monte-Carlo simulations.

Additionally, this paper employs The 1984 National Minimum Drinking Age Act on Beer Consumption as a case study and finds the effect of the policy is $-2,270,766$ gallons of beer on average across all states. We recommend researchers to perform our proposed algorithms at first to take advantage of the abnormality detection property.

This paper starts from discussing the Literature Review in Section 2. Section 3 contains two simulation setups. Under the first simulation setup, we introduces SCM-CS-DID and LASSO-CS-DID. We show the usual methods, including CS-DID, would produce a biased estimate under the DGP in Section 3, while SCM-CS-DID and LASSO-CS-DID would do well. Under the second simulation setup, we demonstrate LASSO-CS-DID would be problematic, and that motivate us to introduce LASCM-CS-DID. Section 4 summarizes important sta-

tistical models. Section 5 compares SCM, LASSO, and a proxy estimation method LCR in terms of predictability and interpretability. Section 6 applies the usual estimators, SCM-CS-DID, and LASCM-CS-DID to estimate the effect of the 1984 National Minimum Drinking Age Act on beer consumption. Section 7 synthesizes the limitations that arise in this paper. Finally, the Appendix contains definitions, assumptions, and theorems from Callaway and Sant'Anna (2021) and Goodman-Bacon (2021), Tables, and Figures.

2 Literature Review

2.1 Methodology

In policy evaluation, the staggered adoption and the dynamic treatment effect are problematic if researchers naively apply econometric tools without dealing with them properly. Goodman-Bacon (2021) gives intuition that, under staggered adoption, the Two-Way-Fixed-Effect Difference-in-Difference (TWFE-DID) estimator is biased, and the true estimator is a variance-weighted average of all combinations of treatment windows (early vs late in the early window, late vs untreated in the late window, untreated vs early in the early window, untreated vs early in the late window, where the last comparison is the source of the bias). We attached the major theorem in Goodman-Bacon (2021) in our Appendix. This intuition can be extended to the Synthetic Control Method (SCM) and LASSO with staggered adoption: We should not directly apply SCM or LASSO under staggered adoption framework. It is quite surprising that there has been very few papers discussing SCM under staggered adoption framework. Ben-Michael et al. (2021) attempts to use SCM on staggered adoption. They use a pooled SCM method under a linear factor model and an $AR(L)$ process. This is the only literature I found that attempts to discover the potential of staggered SCM.

To gain intuition behind the staggered adoption and dynamic treatment effect framework, we want to employ the Monte-Carlo Simulation and see how various estimators behave under different simulation setups. Baker et al. (2022) uses two groups of three Monte Carlo

simulations to show that the TWFE-DID is unbiased under Not Staggered + Constant τ, Not Staggered + Dynamic τ, where the τ increases with time at a constant rate, and Staggered + Constant/Equal τ. They then show that the TWFE-DID is biased under Staggered + Constant/Unequal τ, Staggered + Dynamic/Equal τ, and Staggered + Constant/Unequal τ. This literature summarizes the common approaches to solve the bias in TWFE-DID: Callaway and Sant'Anna (2021) Estimator (CS-DID), Sun and Abraham (2021) Estimator, and Stacked Regression Estimators. All three estimators give an unbiased estimation of τ under the later three simulations. Baker et al. (2022) inspire this paper to run simulations similar to their setups in Section 3.

One crucial estimator that we employ (or combine with other estimators to form a new algorithm in Section 3) is CS-DID. Callaway and Sant'Anna (2021) show that their CS-DID, under Assumptions 1 to 6, can recover the τ under heterogeneous treatment effect over time. However, some of their assumptions are relatively hard to observe in reality. Using their notations, Assumption 2 requires

$$\{Y_{i,1}, Y_{i,2}, \ldots, Y_{i,\mathcal{T}}, X_i, D_{i,1}, D_{i,2}, \ldots, D_{i,\mathcal{T}}\}_{i=1}^{n}$$

to be i.i.d., which generally require strong experimental designs and might not be easily achieved in policy evaluation programs. Also, assumptions 4 and 5 generalize the parallel trend assumption to the conditional parallel trend assumptions with respect to different groups (Never-treated group or not-yet-treated group). As mentioned in Remark 2 in Callaway and Sant'Anna (2021): "In some applications, practitioners may not be comfortable with using 'never-treated' units as part of the comparison group because they behave very differently from the other 'eventually treated' units." I agree that "treated or not" contains important information itself, and I would suggest that it is reasonable to consider the order/sequence of treatments in some applications (e.g. the agent with noticeable degree of problem would be treated first). Therefore, generalization of assumptions 4 and 5 is needed

if differences exist between the early-treated, late-treated, and never-treated groups. We will generalize the conditional parallel trends assumptions from a different angle in Section 3, and we propose our Assumptions 7 and 8 in the Appendix.

2.2 Useful Properties

We want to mention a specific property of SCM that plays crucial role in our paper. Abadie (2021) provides practical guidance to researchers employing synthetic control methods. Abadie summarizes the advantages of SCM, including *No Extrapolation* and *Transparency of the Fit*. In our proposed algorithms, the first property serves great importance.

3 Simulations

3.1 Setup 1

Analogous to the simulation setup in Baker et al. (2022), we generate 4 simulation frameworks to investigate the performance of different estimation methods under different setups. The 4 simulation frameworks are generated by $\{\text{NS}, \text{S}\} \times \{\text{C}, \text{D}\}$, where NS and S denote the "Not Staggered" and "Staggered" status, i.e. there does not/does exist the staggered adoption, and C and D denote the "Constant" and "Dynamic" treatment effect statuses, i.e. there does not/does exist the dynamic treatment effect. In this section, dynamic treatment effect contains variation in treatment effect only with respect to time, i.e. treatment effect only varies with time but does not go across different units.

This paper will first gain insights by running $5,000$ Monte Carlo simulations by the following setup: Assume we have N units and T periods under a balanced panel data setup. The outcome variable for the i-th unit at time t is denoted by Y_{it}, and it is generated based on a non-stationary model as

$$Y_{it}(T_i) = v_t(T_i) + \tau_k D_{it} + u_{it}$$

where $T_i \in \{t_1, t_2, \infty\}$ is the treatment-group indicator and $0 \leq t_1 < t_2 \leq T$, $D_{it} \stackrel{\text{def}}{=} \mathbf{1}_{t \geq T_i}$ is the treatment dummy variable, and $\tau_k = 300 - 10k$ is the time-varying treatment effect variable where the index $k \stackrel{\text{def}}{=} \max\{0, t - T_i\}$ is the event-time variable, and $u_{it} \sim_{i.i.d.} N(0,1)$ is the noise term. The $v_t(T_i)$ is defined as

$$v_t(T_i) = \begin{cases} t & \text{if } T_i = t_2 \\ 1.5t & \text{if } T_i = t_1 \\ 2t & \text{if } T_i = \infty \end{cases}$$

In the simulation setup, we construct the outcome variable based on a linear factor model and separate the outcome variables of early-treated, late-treated, and never-treated groups by a different $v_t(T_i)$, so that the conditional parallel trends assumptions (i.e. Assumptions 4 and 5) in Callaway and Sant'Anna (2021) are all violated.

In the first simulation (Simulation 1 $\stackrel{\text{def}}{=}$ Not Staggered and Constant Treatment Effect τ_k), we take the first 20 units as treated states $\stackrel{\text{def}}{=} T_{i \in \{1,\ldots,20\}} = 10$ and the rest 30 as never-treated units $\stackrel{\text{def}}{=} T_{i \in \{21,\ldots,50\}} = \infty$. The outcome variable Y_{it} is generated as mentioned above, and the treatment effect $\tau_k = 300 \; \forall k \geq 0$ and $\tau_k = 0$ otherwise. Therefore, the outcome variable of the never-treated group $Y_{it}(\infty)$ is composed of time and unit fixed effects and a stochastic error term. The outcome variable of the treated group $Y_{it}(10)$ has a 300-unit increase in addition to the composition of $Y_{it}(\infty)$ when $t \geq 10$.

In our second simulation (Simulation 2 $\stackrel{\text{def}}{=}$ Not Staggered and Dynamic τ_k), we take the first 20 units as treated states $\stackrel{\text{def}}{=} T_{i \in \{1,\ldots,20\}} = 10$ and the rest 30 as never-treated units $\stackrel{\text{def}}{=} T_{i \in \{21,\ldots,50\}} = \infty$. The outcome variable Y_{it} is generated as usual, and the treatment effect $\tau_k = \max\{0, 300 - 10k\} \; \forall k \geq 0$ and $\tau_k = 0$ otherwise, where 10 is the time-varying coefficient and k is the event-time variable. Therefore, the outcome variable of the treated group $Y_{it}(10)$ has a 300-unit increase in addition to the composition of $Y_{it}(\infty)$ when $t \geq 10$. Then the size of the increment decreases linearly by 10 with respect to every 1 year increase until the effect vanishes after the treatment launches.

In the third and forth simulations (Simulation 3 $\stackrel{\text{def}}{=}$ Staggered and Constant τ_k; Simulation 4 $\stackrel{\text{def}}{=}$ Staggered and Dynamic τ_k), we take the first 10 units as early-treated states $\stackrel{\text{def}}{=} T_{i\in\{1,\dots,10\}} = 10$, the next 10 units as late-treated states $\stackrel{\text{def}}{=} T_{i\in\{11,\dots,20\}} = 20$, and the rest 30 as never-treated units $\stackrel{\text{def}}{=} T_{i\in\{21,\dots,50\}} = \infty$. The outcome variable Y_{it} and the treatment effect τ_k are defined similarly to those in Simulation 1 and 2.

For all of the simulations, we run TWFE-DID, SCM, CS-DID, SCM-DID, SCM-CS-DID, and LASSO-CS-DID, where the last three estimators are our proposed methods at this stage.

First, we define the algorithms of SCM-DID, SCM-CS-DID, and LASSO-CS-DID:

Algorithm 1 SCM-DID

Input: Data = $\{\boldsymbol{Y}\}$, Trajectory Fit Criteria
Output: $\hat{\tau}^{SCM-DID}$
for each unit i in each treated group **do**
    Form proper $\boldsymbol{Y}_{c,pre}$ by choosing $\boldsymbol{Y}_{j,pre}$ s.t. $j \notin$ same group as i and $j \in$ Not-Yet-Treated group
    SCM on $\{\boldsymbol{Y}_{i,pre}, \boldsymbol{Y}_{c,pre}\}$ and compute the synthetic counterpart $\boldsymbol{Y}_{synthetic}$
    Compute the Trajectory Fit score $TF_i \stackrel{\text{def}}{=} \frac{\sqrt{||(\boldsymbol{Y}_{i,pre} - \boldsymbol{Y}_{c,pre})||^2/\#(pre)}}{||\boldsymbol{Y}_{i,pre}||/\#(pre)}$
    if $TF_i \leq$ Criteria **then**
        TWFE-DID on $\{\boldsymbol{Y}_i, \boldsymbol{Y}_{synthetic}\}$ to obtain $\hat{\tau}_i^{SCM-DID}$
    else
        Next i
    end if
end for
Calculate $\hat{\tau}^{SCM-DID} = \text{mean}(\hat{\tau}_i^{SCM-DID})$

The intuition behind running SCM then running CS-DID is that SCM, if perform well, will construct a synthetic counterpart follows the trajectory of the outcome variable of the treated unit. This naturally implies the conditional parallel trends assumptions in Callaway and Sant'Anna (2021) will hold. We include SCM-DID to provide an analogous argument parallel with DID, SCM, and LASSO.

Here are some remarks on Algorithm 1: First, *Trajectory Fit Criteria* is a value set by the practitioners. Through this paper, we would like to propose a range between 0.1 to 0.2. Secondly, TF_i is defined intuitively: it measures the pre-treatment discrepancies between the outcome variable of the synthetic counterpart and that of the treated unit (i.e. how good

SCM performs). Notice this paper intentionally uses a Root-Mean-Squared-Error type of construction on the numerator of TF_i, and an absolute mean of the pre-treatment outcome variable of the treated on the denominator. We want to distinguish TF_i from the traditional R^2 and emphasize the extent of the pre-treatment trajectory fit. However, one can easily modify TF_i to any traditional forms of measures for discrepancy WLOG. Thirdly, the reason for the inclusion of TF_i in SCM-related estimators is that: in our DGP, recall the generation of $v_t(T_i)$. It is obvious that $v_t(10) = 0.5v_t(20) + 0.5v_t(\infty)$, so SCM will theoretically perform perfectly on the units in the early-treated group. However, there is no way that SCM can offer a good match on the units in the late-treated group. Abadie (2021) advises not to use SCM if the pre-treatment fit (the TF_i in our example) is poor. Thus, the use of TF_i is to exclude states with poor pre-treatment fit. Finally, SCM-DID is a biased estimation method under staggered adoption (since running TWFE-DID on staggered units will be biased by Theorem A.1). Therefore, we only use SCM-DID to smoothly introduce the main proposed estimators of our paper: SCM-CS-DID and LASCM-CS-DID.

Algorithm 2 SCM-CS-DID

Input: Data = $\{\boldsymbol{Y}\}$, Trajectory Fit Criteria
Output: $\hat{\tau}^{SCM-CS-DID}$
for each unit i in each treated group **do**
    Form proper $\boldsymbol{Y}_{c,pre}$ by choosing $\boldsymbol{Y}_{j,pre}$ s.t. $j \notin$ same group as i and $j \in$ Not-Yet-Treated group
    SCM on $\{\boldsymbol{Y}_{i,pre}, \boldsymbol{Y}_{c,pre}\}$ and compute the synthetic counterpart $\boldsymbol{Y}_{synthetic}$
    Compute the Trajectory Fit score TF_i
    if $TF_i \leq$ Criteria **then**
        CS-DID on $\{\boldsymbol{Y}_i, \boldsymbol{Y}_{synthetic}\}$ to obtain $\hat{\tau}_i^{SCM-CS-DID}$
    else
        Next i
    end if
end for
Calculate $\hat{\tau}^{SCM-CS-DID} = \text{mean}(\hat{\tau}_i^{SCM-CS-DID})$

Algorithm 2 is analogous to Algorithm 1, and the only difference is that we conduct CS-DID instead of TWFE-DID on $\{\boldsymbol{Y}_i, \boldsymbol{Y}_{synthetic}\}$. We would expect this method to perform well under staggered adoption with heterogeneous treatment effect, since it combines the

benefits of SCM and CS-DID. Notice SCM-CS-DID will work under Assumption 7 (See Appendix).

Algorithm 3 LASSO-CS-DID

1: **Input:** Data = $\{\boldsymbol{Y}\}$
2: **Output:** $\hat{\tau}^{LASSO-CS-DID}$
3: **for** each unit i in each treated group **do**
4: Form proper $\boldsymbol{Y}_{c,pre}$ by choosing $\boldsymbol{Y}_{j,pre}$ s.t. $j \notin$ same group as i and $j \in$ Not-Yet-Treated group
5: LASSO on $\{\boldsymbol{Y}_{i,pre}, \boldsymbol{Y}_{c,pre}\}$ and compute the synthetic counterpart $\boldsymbol{Y}_{synthetic}$
6: CS-DID on $\{\boldsymbol{Y}_i, \boldsymbol{Y}_{synthetic}\}$ to obtain $\hat{\tau}_i^{LASSO-CS-DID}$
7: **end for**
8: Calculate $\hat{\tau}^{LASSO-CS-DID} = \text{mean}(\hat{\tau}_i^{LASSO-CS-DID})$

Algorithm 3 is analogous to Algorithm 2, and the only difference is that we conduct LASSO instead of SCM on $\{\boldsymbol{Y}_{i,pre}, \boldsymbol{Y}_{c,pre}\}$. Notice we do not worry about the trajectory fit of the synthetic counterpart by LASSO, because the issue of the late-treated group in the SCM does not matter in the LASSO: $v_t(20) = 0.5v_t(\infty)$, so LASSO will theoretically perform perfectly on the units in the late-treated group. Therefore, there is no intention to include TF_i at this moment. Notice LASSO-CS-DID will work under Assumption 8 (See Appendix).

3.2 Simulation Results for Setup 1

Figure 1 shows the outcome paths of the four simulations. The colored lines are paths of outcome variable averages clustered by early-treated, late-treated, and never-treated groups. For each simulation, we generate $5,000$ datasets (each dataset contains $N = 50$ and $T = 30$ observations) and estimate τ by performing TWFE-DID, SCM, CS-DID, SCM-DID, SCM-CS-DID, and LASSO-CS-DID. The set of estimates by different estimation methods of τ, $\hat{\tau}^i$ where $i \in$ {TWFE-DID, SCM, CS-DID, SCM-DID, SCM-CS-DID, LASSO-CS-DID}, are then compared based on their distance between the τ. In other words, we care about the extent of biasedness of the $\hat{\tau}^i$s,

Figures 2 to 7 are the kernel density plots of each $\hat{\tau}^i$ from $5,000$ Monte Carlo simulations for each data generating processes. Figures 2 and 4 suggest that TWFE-DID and CS-DID provide biased estimates of the τ under all simulation settings since we form the DGP by violating the conditional parallel trend assumptions.

Figures 3 and 5 show that SCM and SCM-DID give unbiased estimates of τ when there is no dynamics in treatment effect. Moreover, we can observe that SCM-DID offers a less varied distribution of the estimated τ than SCM because SCM-DID, in principle, averages all values in the post-treatment window, and thus becomes more efficient.

Figures 6 and 7 show that both SCM-CS-DID and LASSO-CS-DID can offer unbiased estimates of the τ under all simulation setups. Notice that the density plot of LASSO-CS-DID is less spread out than that of SCM-CS-DID. As proved in Proposition 5.1, LASSO-CS-DID gives more efficient result is expected.

If LASSO-CS-DID is so good, why we care about SCM-CS-DID? To illustrate the practical importance of SCM-CS-DID, we include a second simulation setup that has important insights discussed in Section 5.2.

3.3 Setup 2

We generate a simulation framework $\{\text{S} \times \text{D}\}$, different from the four-simulation setup previously, to investigate the performance of SCM-CS-DID and LASSO-CS-DID. We will demonstrate why LASSO-CS-DID (using Cross Validation or BIC) will be biased under this setup and naturally introduce a new method: LASCM-CS-DID.

Assume we have $N = 50$ units and $T = 30$ periods under a balanced panel data setup. The outcome variable for i-th unit at time t is denoted by Y_{it} and generated based on a non-stationary model as:

$$Y_{it}(T_i) = v_t(T_i, i) + \tau_k D_{it} + u_{it} + h_t(i)$$

where $T_i \in \{10, 20, \infty\}$ is the treatment-group indicator, $D_{it} \overset{\text{def}}{=} \mathbf{1}_{t \geq T_i, i \notin \{1,\ldots,5,11,\ldots,15\}}$ is the treatment dummy variable, and $\tau_k = 300 - 10k$ is the time-varying treatment effect variable where the index $k \overset{\text{def}}{=} \max\{0, t - T_i\}$ is the event-time variable, and $u_{it} \sim_{i.i.d.} N(0,1)$ is the noise term. The $v_t(T_i, i)$ is defined as

$$v_t(T_i, i) = \begin{cases} t & \text{if } T_i = t_2 \text{ and } i \notin \{1, \ldots, 5, 11, \ldots, 15\} \\ 1.5t & \text{if } T_i = t_1 \text{ and } i \notin \{1, \ldots, 5, 11, \ldots, 15\} \\ 2t & \text{if } T_i = \infty \text{ and } i \notin \{1, \ldots, 5, 11, \ldots, 15\}. \end{cases}$$

The $h_t(i)$ is defined as:

$$h_t(i) \sim_{i.i.d.} U(-10, 10) \text{ if } i \in \{1, \ldots, 5, 11, \ldots, 15\}$$

In this simulation setup, we construct the outcome variable based on a linear factor model, and separate the outcome variables of early-treated, late-treated, and never-treated groups by a different $v_t(T_i, i)$, so that the conditional parallel trends assumptions (i.e. Assumptions 4 and 5) in Callaway and Sant'Anna (2021) are all violated. Additionally, we force the first half units in the early-treated and late-treated group to be absurd: The outcome variables of these units follow a uniform distribution $U(-10, 10)$ for all periods. Even though these units are treated, their outcome variables respond with no change. Obviously, we want to exclude these observations if we try to run SCM-CS-DID and LASSO-CS-DID. From the SCM-CS-DID in the previous setup, we know that Trajectory Fit is a good method to decide which synthetic counterparts by SCM are good and which states can stay in the new data set for the CS-DID. It seems reasonable to extend the Trajectory Fit method to LASSO-CS-DID as well.

Therefore, the improved-version of LASSO-CS-DID is defined as Algorithm 4:

We will show that even the improved-version of LASSO-CS-DID fail to exclude these problematic states and leads to biasedness.

Algorithm 4 LASSO-CS-DID (New)

Input: Data = $\{\boldsymbol{Y}\}$, Trajectory Fit Criteria
Output: $\hat{\tau}^{LASSO-CS-DID}$
for each unit i in each treated group **do**
    Form proper $\boldsymbol{Y}_{c,pre}$ by choosing $\boldsymbol{Y}_{j,pre}$ s.t. $j \notin$ same group as i and $j \in$ Not-Yet-Treated group
    LASSO on $\{\boldsymbol{Y}_{i,pre}, \boldsymbol{Y}_{c,pre}\}$ and compute the synthetic counterpart $\boldsymbol{Y}_{synthetic}$
    Compute the Trajectory Fit score TF_i
    if $TF_i \leq$ Criteria **then**
        CS-DID on $\{\boldsymbol{Y}_i, \boldsymbol{Y}_{synthetic}\}$ to obtain $\hat{\tau}_i^{LASSO-CS-DID}$
    else
        Next i
    end if
end for
Calculate $\hat{\tau}^{LASSO-CS-DID} = \text{mean}(\hat{\tau}_i^{LASSO-CS-DID})$

3.4 Simulation Results for Setup 2

For this simulation, we generate 5,000 Monte-Carlo Simulations (each dataset contains $N = 50$ and $T = 30$ observations) and estimate τ by performing SCM-CS-DID and LASSO-CS-DID. Since the LASSO can sometimes produce a wonderful Trajectory Fit even if the unit is having abnormal distributions/patterns, Figure 8 shows that LASSO-CS-DID is biased for all $TFC \geq 0.02$, and is severely biased when $TFC = 0.01$. Therefore, LASSO-CS-DID (by using Cross Validation) cannot detect those problematic units and is not performing well.

One might argue that BIC prevents the LASSO from overfitting, but Figure 8 presents that the LASSO-CS-DID (by using BIC) only performs well when $TFC = 0.02$, and becomes severely biased even when a subtle increase of TFC occurred. This is intuitively correct: BIC penalizes heavily on new member of control units into the synthetic component (if the new member is not "good"). Thus, when $TFC = 0.02$, the synthetic counterparts must be extremely close to the treated unit. However, since BIC cannot exclude those abnormal units by nature, with TFC increases, the estimated τ is inevitably biased. Also, we cannot determine which value of TFC to use in reality. In comparison with SCM-CS-DID, which gives unbiased estimates of τ for $TFC \leq 0.25$, LASSO-CS-DID by BIC is not favourable for

practical reasons.

From the first simulation (and latter in Section 5), we see how the LASSO-CS-DID outperforms the SCM-CS-DID, so we do want to combine the advantages of SCM-CS-DID (abnormalities detection) and LASSO-CS-DID (Efficiency and unbiasedness) together. This motivates us to propose the final important method: LASCM-CS-DID (LASSO-SCM-CS-DID), which is defined by Algorithm 5:

Algorithm 5 LASCM-CS-DID

Input: Data = $\{\boldsymbol{Y}\}$, Trajectory Fit Criteria
Output: $\hat{\tau}^{LASCM-CS-DID}$
for each unit i in each treated group **do**
    Form proper $\boldsymbol{Y}_{c,pre}$ by choosing $\boldsymbol{Y}_{j,pre}$ s.t. $j \notin$ same group as i and $j \in$ Not-Yet-Treated group
    SCM on $\{\boldsymbol{Y}_{i,pre}, \boldsymbol{Y}_{c,pre}\}$ and compute the synthetic counterpart $\boldsymbol{Y}^{scm}_{synthetic}$
    LASSO on $\{\boldsymbol{Y}_{i,pre}, \boldsymbol{Y}_{c,pre}\}$ and compute the synthetic counterpart $\boldsymbol{Y}^{lasso}_{synthetic}$
    Compute the Trajectory Fit score TF_i using $\{\boldsymbol{Y}_{i,pre}, \boldsymbol{Y}^{scm}_{synthetic}\}$
    if $TF_i \leq$ Criteria **then**
        CS-DID on $\{\boldsymbol{Y}_i, \boldsymbol{Y}^{lasso}_{synthetic}\}$ to obtain $\hat{\tau}_i^{LASCM-CS-DID}$
    else
        Next i
    end if
end for
Calculate $\hat{\tau}^{LASCM-CS-DID} = \text{mean}(\hat{\tau}_i^{LASCM-CS-DID})$

Notice we run both SCM and LASSO in LASCM-CS-DID. For any treated unit i in a treated group, we use the synthetic counterpart $\boldsymbol{Y}^{scm}_{synthetic}$ and actual pre-treatment outcome variable to compute TF_i. Thus, we can detect abnormalities. For those units without abnormalities, we run CS-DID on $\boldsymbol{Y}^{lasso}_{synthetic}$ and $\boldsymbol{Y}_i$ to take the advantage of the efficiency and unbiasedness of LASSO-CS-DID. Figure 8 demonstrates an unbiased estimate of τ when $TFC \leq 0.25$ for LASCM-CS-DID, which supports our results.

4 Usual Statistical Models

The following subsections will consider a panel data setting with $N+1$ units within t time periods, where $t = 1, \ldots, T$. Using the potential outcome setup (Rubin (1974); Holland (1986); Imbens and Rubin (2015); Doudchenko and Imbens (2016)), each of the $N+1$ units is assigned with a pair of potential outcomes $Y_{i,t}(0)$ and $Y_{i,t}(1)$, where $Y_{i,t}(0)$ denotes the outcome for an individual unit i in time t in the untreated status, and $Y_{i,t}(1)$ denotes the outcome for an individual unit i in time t in the treated status. Let $\tau_{i,t} = Y_{i,t}(1) - Y_{i,t}(0)$, for $i = 0, 1, \ldots, N$ and $t = 1, \ldots, T$, which stands for the causal effects at time t and unit i.

4.1 Notation

This paper adopts the same structure of notation as Doudchenko and Imbens (2016). Let $i = 1, \ldots, N$ be the N units which do not receive the treatment through all time windows. Let the unit $i = 0$ be the unit which is untreated when $t = 1, \ldots, T_0$, and receives treatment when $t = T_0 + 1, \ldots, T_0 + T_1$, and the total time periods $T = T_0 + T_1$. Doudchenko and Imbens (2016) denote the treatment received by $W_{i,t}$, satisfying:

$$W_{i,t} = \begin{cases} 1 & \text{if } i = 0, \text{ and } t \in \{T_0 + 1, \ldots, T\}, \\ 0 & \text{otherwise.} \end{cases}$$

The observed outcome for unit i in period t is $Y_{i,t}^{\text{obs}}$, which is described by $W_{i,t}$ as:

$$Y_{i,t}^{\text{obs}} = Y_{i,t}\left(W_{i,t}\right) = \begin{cases} Y_{i,t}(0) & \text{if } W_{i,t} = 0 \\ Y_{i,t}(1) & \text{if } W_{i,t} = 1 \end{cases}$$

Doudchenko and Imbens (2016) denote the time-invariant individual-level characteristics by X_i, an $M \times 1$ column vector $(X_{i,1}, \ldots, X_{i,M})^\top$, for $i = 0, \ldots, N$. They denote the covariate matrix for the control group by $\boldsymbol{X}_{\text{c}}$, an $N \times M$ matrix with the $(i,m)^{\text{th}}$ entry equal to $X_{i,m}$, for $i = 1, \ldots, N$. They denote the covariate matrix for the treatment group by $\boldsymbol{X}_{\text{t}}$, an M-row

vector with the m^{th} entry equal to $X_{0,m}$, so that $\boldsymbol{X} = (\boldsymbol{X}_{\text{t}}, \boldsymbol{X}_{\text{c}})$. Similarly, $\boldsymbol{Y}^{\text{obs}}_{\text{c, pre}}$ denotes the $N \times T_0$ matrix with the $(i,t)^{\text{th}}$ entry equal to $Y^{\text{obs}}_{i,T_0-t+1}$. $\boldsymbol{Y}^{\text{obs}}_{\text{t,pre}}$ denotes a T_0-vector with the t-th entry equal to $Y^{\text{obs}}_{0,t}$. The same applies to $\boldsymbol{Y}^{\text{obs}}_{\text{c, post}}$ and $\boldsymbol{Y}^{\text{obs}}_{\text{t,post}}$ for the post-treatment period. Combining these matrices:

$$\boldsymbol{Y}^{\text{obs}} = \begin{pmatrix} \boldsymbol{Y}^{\text{obs}}_{\text{t,post}} & \boldsymbol{Y}^{\text{obs}}_{\text{c,post}} \\ \boldsymbol{Y}^{\text{obs}}_{\text{t,pre}} & \boldsymbol{Y}^{\text{obs}}_{\text{c,pre}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Y}_{\text{t,post}}(1) & \boldsymbol{Y}_{\text{c, post}}(0) \\ \boldsymbol{Y}_{\text{t,pre}}(0) & \boldsymbol{Y}_{\text{c,pre}}(0) \end{pmatrix}, \quad \text{and } \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{\text{t}} & \boldsymbol{X}_{\text{c}} \end{pmatrix}$$

Immediately, we can see that $\hat{\tau}_{0,t}$ depends on $\boldsymbol{Y}_{\text{t,post}}(1)$ and $\boldsymbol{Y}_{\text{t, post}}(0)$. But we can only observe $\boldsymbol{Y}_{\text{t,post}}(1)$ and $\boldsymbol{Y}_{\text{t, post}}(0)$ is unobservable. Therefore, various econometric approaches are used for an accurate prediction of the relationship between the observed $\boldsymbol{Y}_{\text{t,post}}(1)$ and unobserved $\boldsymbol{Y}_{\text{t, post}}(0)$ by using the information in the observed $\boldsymbol{Y}_{\text{t,pre}}(0)$ and $\boldsymbol{Y}_{\text{c,pre}}(0)$.

4.2 Constraints

Most of the literature imputes the unobserved $Y_{0,T}(0)$ by a linear combination of other observed outcomes, that is:

$$\hat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{N} \omega_i \cdot Y^{\text{obs}}_{i,T} \, .$$

where $Y_{i,T}^{\text{obs}}$ consists of $\boldsymbol{Y}_{\text{c, post}}^{\text{obs}}$, $\boldsymbol{Y}_{\text{t, pre}}^{\text{obs}}$, and $\boldsymbol{Y}_{\text{c, pre}}^{\text{obs}}$. Doudchenko and Imbens (2016) focus on this popular setting and discuss five constraints that are usually imposed on estimators:

$$\mu = 0, \qquad \text{(NO-INTERCEPT)}$$

$$\sum_{i=1}^{N} \omega_i = 1, \qquad \text{(ADDING-UP)}$$

$$\omega_i \geq 0, i = 1, \ldots, N, \qquad \text{(NON-NEGATIVITY)}$$

$$\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} = \mu + \omega^{\top} \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}}, \qquad \text{(EXACT-BALANCE)}$$

$$\omega_i = \bar{\omega}, i = 1, \ldots, N \qquad \text{(CONSTANT-WEIGHTS)}$$

4.3 Difference-in-Difference

DID method, with restrictions of (ADDING-UP), (NON-NEGATIVITY), and (CONSTANT-WEIGHTS), is solving the following:

$$\left(\hat{\mu}^{\text{did}}, \hat{\omega}^{\text{did}}\right) = \arg\min_{\mu,\omega} \left\{ \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} - \mu - \omega^{\top} \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}}\right) \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} - \mu - \omega^{\top} \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}}\right)^{\top} \right\}.$$

In Doudchenko and Imbens (2016), they point out that the $\hat{\omega}^{did}$ does not depend on the data due to the restrictions (ADDING-UP), (NON-NEGATIVITY), and (CONSTANT-WEIGHTS) imposed. As a result, we can readily obtain:

$$\hat{\omega}_i^{\text{did}} = \frac{1}{N}, i = 1, \ldots, N,$$
$$\hat{\mu}^{\text{did}} = \frac{1}{T_0} \sum_{s=1}^{T_0} Y_{0,s}^{\text{obs}} - \frac{1}{N \cdot T_0} \sum_{s=1}^{T_0} \sum_{i=1}^{N} Y_{i,s}^{\text{obs}}.$$

As clearly stated in Doudchenko and Imbens (2016), the estimates for $Y_{0,t}(0)$, for the periods

$t \geq T_0 + 1$, are equal to:

$$\begin{aligned}\hat{Y}_{0,t}^{\text{did}}(0) &= \hat{\mu}^{\text{did}} + \sum_{i=1}^{N} \hat{\omega}_i^{\text{did}} \cdot Y_{i,t}^{\text{obs}} \\ &= \left(\frac{1}{T_0} \sum_{s=1}^{T_0} Y_{0,s}^{\text{obs}} - \frac{1}{N \cdot T_0} \sum_{s=1}^{T_0} \sum_{i=1}^{N} Y_{i,s}^{\text{obs}} \right) + \frac{1}{N} \sum_{i=1}^{N} Y_{i,t}^{\text{obs}}.\end{aligned}$$

Remark 1: If $T = 2$, then the TWFE regression is equivalent to DID.

4.4 Constrained Regression

Constrained Regression, with restrictions of (NO-INTERCEPT), (ADDING-UP), and (NON-NEGATIVITY) (as a special case of ADH SCM, all of the restrictions can be easily relaxed), is solving the following:

$$\begin{aligned}\hat{\omega}^{\text{constr}} =& \arg\min_{\mu,\omega} \left\{ \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} - \mu - \omega^{\top} \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \right) \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} - \mu - \omega^{\top} \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \right)^{\top} \right\} \\ & \text{s.t. } \mu = 0, \quad \sum_{i=1}^{N} \omega_i = 1 \quad \text{and} \quad \omega_i \geq 0, i = 1, \ldots, N.\end{aligned}$$

The Lagrangian form of Constrained Regression is:

$$L(\omega, \lambda, \gamma) = \left(Y_{t,\text{pre}}^{\text{obs}} - \omega^T Y_{c,\text{pre}}^{\text{obs}} \right) \left(Y_{t,\text{pre}}^{\text{obs}} - \omega^T Y_{c,\text{pre}}^{\text{obs}} \right)^T + \lambda \left(1 - \sum_{i=1}^{N} \omega_i \right) - \sum_{i=1}^{N} \gamma_i \omega_i$$

where λ and γ are the Lagrange multipliers. The dual problem is then solving:

$$\begin{aligned}(\hat{\omega}, \hat{\lambda}, \hat{\gamma}) =& \arg\max_{\lambda,\gamma} \{ \arg\min_{\omega} \ L(\omega, \lambda, \gamma) \} \\ & \text{s.t. } \lambda \geq 0, \gamma_i \geq 0, i = 1, \ldots, N.\end{aligned}$$

The first observation is that, without any constraints, the objective function is equivalent to that of Ordinary Least Squares.

By loosening the NON-NEGATIVE constraint, the optimization problem becomes a

quadratic programming problem:

$$\min \left\{ \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} - \mu - w^{\top} \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \right) \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} - \mu - w^{\top} \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \right)^{\top} \right\}$$
$$\text{s.t. } \mu = 0, \ \sum_{i=1}^{N} w_i = 1$$

Let $Q = \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \left(\boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \right)^{T}$, $q = -\boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} \right)^{T}$, and ι be the vector of ones $\in \mathbb{R}^N$, the quadratic programming would have the following expressions:

$$\begin{cases} Q\hat{\omega} + q + \iota\hat{\lambda} &= 0 \\ \iota^{\top}\hat{\omega} - 1 &= 0 \end{cases} \implies \begin{bmatrix} Q & \iota \\ \iota^{\top} & 0 \end{bmatrix} \begin{bmatrix} \hat{\omega} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} -q \\ 1 \end{bmatrix}$$
$$\begin{bmatrix} \hat{\omega} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} Q & \iota \\ \iota^{\top} & 0 \end{bmatrix}^{-1} \begin{bmatrix} -q \\ 1 \end{bmatrix}$$

where $\hat{\lambda}$ is the Lagrangian multiplier. We call this estimator Less Constrained Regression (LCR) throughout the paper. Note that in the following discussions, when we mention SCM, we intrinsically ignore covariates to keep things simple. So we will only discuss the Constrained Regression estimator.

4.5 The Abadie-Diamond-Hainmueller Synthetic Control Method

SCM method, with restrictions of (NO-INTERCEPT), (ADDING-UP), and (NON-NEGATIVITY), is solving the following:

$$\begin{aligned} (\hat{\omega}(V), \hat{\mu}(V)) = & \arg\min_{\omega,\mu} \left\{ (\boldsymbol{X_t} - \mu - \omega^T \boldsymbol{X_c})^T V (\boldsymbol{X_t} - \mu - \omega^T \boldsymbol{X_c}) \right\} \\ & s.t. \sum_{i=1}^{N} \omega_i = 1 \text{ and } \omega_i \geq 0, i = 1, \ldots, N, \mu = 0 \end{aligned} \tag{1}$$

where

$$\begin{aligned}\hat{V} = \underset{V=\text{diag}(v_1,\ldots,v_M)}{\arg\min} & \left\{ \left(\boldsymbol{Y}^{\text{obs}}_{\text{t,pre}} - \hat{\omega}(V)^\top \boldsymbol{Y}^{\text{obs}}_{\text{c,pre}} \right)^\top \left(\boldsymbol{Y}^{\text{obs}}_{\text{t,pre}} - \hat{\omega}(V)^\top \boldsymbol{Y}^{\text{obs}}_{\text{c,pre}} \right) \right\} \\ & \text{s.t. } \sum_{m=1}^{M} v_m = 1 \text{ and } v_m \geq 0, m = 1, \ldots, M.\end{aligned} \tag{2}$$

4.6 Least Absolute Shrinkage and Selection Operator

The LASSO regression is solving the following:

$$\begin{aligned}(\hat{\omega}, \hat{\mu}) = & \underset{\mu,\omega}{\arg\min} \left\{ \left(\boldsymbol{Y}^{\text{obs}}_{\text{t,pre}} - \mu - \omega^\top \boldsymbol{Y}^{\text{obs}}_{\text{c,pre}} \right) \left(\boldsymbol{Y}^{\text{obs}}_{\text{t,pre}} - \mu - \omega^\top \boldsymbol{Y}^{\text{obs}}_{\text{c,pre}} \right)^\top \right\} \\ & \text{s.t. } \|\omega\|_1 \leq t, \mu = 0\end{aligned}$$

where $\|u\|_p = \left(\sum_{i=1}^{N} |u_i|^p \right)^{1/p}$ is the standard ℓ^p norm. The Lagrangian form of LASSO is:

$$(\hat{\omega}, \hat{\mu}) = \underset{\mu,\omega}{\arg\min} \left\{ \left(\boldsymbol{Y}^{\text{obs}}_{\text{t,pre}} - \mu - \omega^\top \boldsymbol{Y}^{\text{obs}}_{\text{c,pre}} \right) \left(\boldsymbol{Y}^{\text{obs}}_{\text{t,pre}} - \mu - \omega^\top \boldsymbol{Y}^{\text{obs}}_{\text{c,pre}} \right)^\top + \lambda \|\omega\|_1 \right\}$$

5 Comparison Between SCM, LASSO, and LCR

We include LCR (not important in other sections) for the following discussion to better illustrate our results. Again, our incentive to employ SCM and LASSO is to form a group of synthetic counterparts that follows the trajectories of the treated units well. We will then be able to perform CS-DID on the treated units and their corresponding synthetic counterparts. Therefore, predictability, the ability of an estimator to give a close fit of the pre-treatment trajectory would be one of our interests. Since for most economists, interpretability is the core of policy evaluations, we want to study how SCM outperform LASSO in terms of interpretability.

5.1 Predictability

In terms of predictability, LASSO will outperform SCM and LCR when the tuning parameter $t \geq 1$, always giving a no-worse-than value (a smaller or equal value) of the objective function than those of SCM and LCR. Recall for these estimators: SCM is solving:

$$\begin{aligned} &\min \left\{ \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} - \mu - \omega^\top \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \right) \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} - \mu - \omega^\top \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \right)^\top \right\} \\ &\text{s.t. } \mu = 0, \quad \sum_{i=1}^{N} \omega_i = 1 \quad \text{and} \quad \omega_i \geq 0, i = 1, \ldots, N. \end{aligned}$$

LASSO is solving:

$$\begin{aligned} &\min \left\{ \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} - \mu - \omega^\top \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \right) \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} - \mu - \omega^\top \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \right)^\top \right\} \\ &\text{s.t. } \|\omega\|_1 \leq t, \quad \mu = 0 \end{aligned}$$

LCR is solving:

$$\begin{aligned} &\min \left\{ \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} - \mu - w^\top \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \right) \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} - \mu - w^\top \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \right)^\top \right\} \\ &\text{s.t. } \mu = 0, \ \sum_{i=1}^{N} w_i = 1 \end{aligned}$$

We are interested in which objective function fed with respective optimal weights gives the smallest value. Denote $f(\omega) = \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} - w^\top \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \right) \left(\boldsymbol{Y}_{\text{t,pre}}^{\text{obs}} - w^\top \boldsymbol{Y}_{\text{c,pre}}^{\text{obs}} \right)^\top$, notice that the objective function f is a standard quadratic function which is convex and continuous. The constrains in SCM, LASSO, and LCR are obvious compact sets, so we can directly state the following result:

Lemma 5.1. *Given* $1 \leq t < \infty$, $f(\omega^*_{LASSO}) \leq f(\omega^*_{LCR}) \leq f(\omega^*_{SCM})$.

The proof is straightforward. Notice that the feasible set of LASSO Ω_{LASSO} is a superset of that of the LCR Ω_{LCR} when $1 \leq t < \infty$, and Ω_{LCR} is a superset of that of the SCM Ω_{SCM} , so the LASSO always yield a no-worse-than value of the objective function than LCR

than SCM. The existences of the global minimums are direct results from the Weierstrass Theorem:

Theorem 5.2. *If $g : \mathbb{R}^n \to \mathbb{R}$ is continuous on compact set X, then the maximum and the minimum of $g(X)$ can be obtained (i.e. $\sup g(X) \in g(X)$ and $\inf g(X) \in g(X)$).* ■

We then want to state the relationship of $f(\omega^*_{LASSO})$, $f(\omega^*_{LCR})$, and $f(\omega^*_{SCM})$ with the tuning parameter $0 < t < 1$ (the case where $t = 0$ is trivial in practice). Notice the argument for $f(\omega^*_{LCR}) \leq f(\omega^*_{SCM})$ is the same as before, so we only need to discuss the relationship between $f(\omega^*_{LASSO})$ and $f(\omega^*_{LCR})$. The usual way to select t is by Cross Validation. Suppose CV chooses $t = \hat{t}$, let's define the feasible sets of the LASSO (given $t = \hat{t}$), LCR, and SCM:

$$\Omega_{LASSO} \stackrel{\text{def}}{=} \{\omega : ||\omega||_1 \leq \hat{t}\}$$
$$\Omega_{LCR} \stackrel{\text{def}}{=} \{\omega : \sum_{i=1}^{N} \omega_i = 1\}$$
$$\Omega_{SCM} \stackrel{\text{def}}{=} \{\omega : \sum_{i=1}^{N} \omega_i = 1, \omega_i \in \mathbb{R}_+\}$$

Denote the true optimal weights $\omega^* \stackrel{\text{def}}{=} \arg\min_{\omega \in \mathbb{R}^N} \left\{ \left(\boldsymbol{Y}^{\text{obs}}_{\text{t,pre}} - \omega^\top \boldsymbol{Y}^{\text{obs}}_{\text{c,pre}} \right) \left(\boldsymbol{Y}^{\text{obs}}_{\text{t,pre}} - \omega^\top \boldsymbol{Y}^{\text{obs}}_{\text{c,pre}} \right)^\top \right\}$. We either have $\omega^* \in \Omega_{LASSO}$ or $\omega^* \notin \Omega_{LASSO}$.

In the first case, $\omega^* = \arg\min_{\omega \in \Omega_{LASSO}} \left\{ \left(\boldsymbol{Y}^{\text{obs}}_{\text{t,pre}} - \omega^\top \boldsymbol{Y}^{\text{obs}}_{\text{c,pre}} \right) \left(\boldsymbol{Y}^{\text{obs}}_{\text{t,pre}} - \omega^\top \boldsymbol{Y}^{\text{obs}}_{\text{c,pre}} \right)^\top \right\} = \omega^*_{LASSO}$ by definition. Therefore, $f(\omega^*_{LASSO}) \leq f(\omega^*_{LCR})$.

In the second case, we have two sub-cases: Either $\omega^* \in \Omega_{LCR}$ or $\omega^* \in (\mathbb{R}^N \setminus (\Omega_{LASSO} \cup \Omega_{LCR}))$. In the first case, by the same superset argument, we can conclude that $f(\omega^*_{LCR}) \leq f(\omega^*_{LASSO})$ (If $\omega^* \in \Omega_{SCM}$, then conclusion becomes $f(\omega^*_{SCM}) = f(\omega^*_{LCR}) \leq f(\omega^*_{LASSO})$); However, we do not have a decisive conclusion on the relationship between $f(\omega^*_{LASSO})$ and $f(\omega^*_{LCR})$ in the second case. We invite researchers to generalize our Lemma 5.3 and Proposition 5.1. Specifically, by showing the inequality relationship holds under $\omega^* \in (\mathbb{R}^N \setminus (\Omega_{LASSO} \cup \Omega_{LCR}))$.

Lemma 5.3. *Suppose t chosen by CV s.t. $0 < t < 1$, then $g^{obj}(\omega^*_{LASSO}) \leq h^{obj}(\omega^*_{LCR}) \leq f^{obj}(\omega^*_{SCM})$ only if $\omega^* \in \Omega_{LASSO}$. $f(\omega^*_{SCM}) = f(\omega^*_{LCR}) \leq f(\omega^*_{LASSO})$* only if $\omega^* \in \Omega_{SCM}$.

Combining Lemma 5.1 and 5.3, we have the following Proposition:

Proposition 5.1. *If t chosen by CV satisfies $1 \leq t < \infty$,* then $g^{obj}(\omega^*_{LASSO}) \leq h^{obj}(\omega^*_{LCR}) \leq f^{obj}(\omega^*_{SCM})$. *If t chosen by CV satisfies $0 < t < 1$,* then $g^{obj}(\omega^*_{LASSO}) \leq h^{obj}(\omega^*_{LCR}) \leq f^{obj}(\omega^*_{SCM})$ only if $\omega^* \in \Omega_{LASSO}$; $f(\omega^*_{SCM}) = f(\omega^*_{LCR}) \leq f(\omega^*_{LASSO})$ only if $\omega^* \in \Omega_{SCM}$ ■.

Since CV chooses $\hat{t}$ based on MSE (here the data input are validation group and test group), we suspect there will at least be some reduction in the objective function of LASSO when $0 < \hat{t} < 1$. Again, we invite researchers for further rigorous generalization.

Figure 9 is an example when the SCM, LASSO, and LCR yield the same optimal sets of weights ω^*. All three sub-figures are under the programming that

$$\min \left\{ \left(\boldsymbol{Y}^{\text{obs}}_{\text{t,pre}} - \omega^\top \boldsymbol{Y}^{\text{obs}}_{\text{c,pre}} \right) \left(\boldsymbol{Y}^{\text{obs}}_{\text{t,pre}} - \omega^\top \boldsymbol{Y}^{\text{obs}}_{\text{c,pre}} \right)^\top \right\}$$
$$\text{s.t.} \quad \sum_{i=1}^{N} \omega_i = 1$$

where $\boldsymbol{Y}^{\text{obs}}_{\text{t,pre}} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$, $\boldsymbol{Y}^{\text{obs}}_{\text{c,pre}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and therefore the true optimal weights $\omega^* = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$. Sub-figures 9a, 9b, and 9c show that ω^* are equal, and any other weights ω^+ would not be the optimal weights.

5.2 Interpretability

In Section 4 of Abadie (2021), Abadie compares the advantages of using SCM over the traditional Linear Regression model. We include the most relevant advantages he summarizes within the scope of our discussion on SCM, LASSO, and LCR: *No Extrapolation* and *Transparency of the Fit.*

No Extrapolation. Because of the non-negativity and adding-up-to-one properties of ω^{SCM}, it is easy to see that the synthetic control counterparts would be inside of the convex

hull of the control units in the donor pool, so no extrapolation is guaranteed. Meanwhile, LASSO and LCR, since there is no restriction on the sign of the weights, cannot guarantee an interpolation within the support of the data. This property becomes extremely useful when there is abnormalities in the dataset as in Section 3.3 and 6. SCM uses this property to detect and exclude abnormal units based on Trajectory Fit, where then CS-DID can be applied.

Transparency of the Fit. Since SCM protects the synthetic counterparts from the extrapolation, we can calculate the discrepancies between the counterparts and the corresponding treated units during the pre-treatment period. Based on the size of the discrepancies, we are able to conclude the ability of the units in the control group to approximate the treated units by interpolation only. If the disparities between the counterparts and the treated units are non-negligible, then Abadie et al. (2010, 2015) suggest not to employ SCM.

6 Case Study: The Effect of The 1984 National Minimum Drinking Age Act on Beer Consumption

In this section, we present an empirical analysis of our framework, focusing on the impact of The 1984 National Minimum Drinking Age Act on the average beer consumption across 50 states. The key variable of interest in our study is the beer consumption, which is measured in gallons. We apply TWFE-DID, CS-DID, SCM-CS-DID, LASSO-CS-DID, and LASCM-CS-DID and compare the results to provide a straightforward analysis of the policy's impact on beer consumption.

6.1 Background

The 1984 National Minimum Drinking Age Act was a significant piece of U.S. legislation that aimed to reduce alcohol-related accidents by making it illegal for any individuals under the age of 21 to purchase or publicly possess alcoholic beverages.

In the early 1970s, many states in the U.S. lowered their drinking ages, primarily in response to the argument that if 18-year-olds could be drafted to fight in wars (e.g. Vietnam War), they should also be allowed to drink. However, this led to some unintended consequences. States that had lowered the drinking age saw an increase in alcohol-related accidents among young drivers. This soon became a major public health concern (Carpenter and Dobkin (2011)).

As the issue continued to gain prominence and relevance, an undeniable crescendo of societal and stakeholder pressure began to mount, necessitating a thorough examination and resolution. Various advocacy groups campaigned vigorously for a uniform national drinking age of 21 (Grant (1984)).

In response to this mounting concern, in 1984, Congress passed the National Minimum Drinking Age Act. Rather than directly imposing a restriction on the national drinking age, the act used a clever mechanism: it threatened to withhold 10% of federal highway construction funds from states that didn't enforce the minimum legal drinking age of 21. Given the significant amount of money at stake, all states eventually complied (Carpenter and Dobkin (2011)).

6.2 Data Description

Our data source is the National Institute on Alcohol Abuse and Alcoholism, which contains a balanced-panel data on apparent alcoholic beverage consumption by state-level and types of alcoholic beverage from 1970 through 2021. Our variables include State ID, Year, and Gallons of Beer. We pick the consumption of beer as our variable of interest because the main type of alcoholic consumption for youths is beer. Reduced rates of alcohol use among youth after the 1984 Act was primarily evident in decreased rates of beer consumption (Toomey et al. (1996); Berger and Snortum (1985)). However, few research have discussed the average amount of beer consumption decrease after the Act. This motivates us to focus our discussion on estimating the average beer consumption. For simplicity and consistency

with the simulation, no covariate is included.

We also classify states in the dataset. Some of states have multiple modifications of the minimum drinking age (i.e. continuous treatment), so we exclude them from the dataset. 25 states remain after the removal. For the remaining states, there are differences in their adoption time, as indicated in Figure **10**, specifically:

The state which raised the drinking age to 21 in 1980: Illinois.

The state which raised the drinking age to 21 in 1982: Maryland.

States which raised the drinking age to 21 in 1984: Alaska, Delaware.

States which raised the drinking age to 21 in 1985: Arizona, Kansas.

States which raised the drinking age to 21 in 1986: Alabama, Hawaii, Mississippi, Vermont.

The state which raised the drinking age to 21 in 1987: Idaho.

States which raised the drinking age to 21 in 1988: Colorado, Wyoming.

The never treated states which do not raise the drinking age (they have set 21 as the minimum age for beer consumption long before the Act): Arkansas, California, Indiana, Kentucky, Missouri, Nevada, New Mexico, North Dakota, Oregon, Pennsylvania, Utah, and Washington.

After finalizing the states and their groups, we will demonstrate how we choose which time periods to include. First, let's begin with a simple case: Suppose there are only two states {NC, NY} in the framework, where NC is the treated state that decreased the minimum drinking age to 19 at 1975 and then adopted The 1984 National Minimum Drinking Age Act at 1984, and NY is the never-treated state, then we can easily choose the time windows that we want to keep:

1. The pre-treatment periods between 1975 and 1984, which are the periods under no policy change.

2. The post-treatment periods.

Similarly, in the staggered adoption framework, we can divide the time windows into

three parts:

1. The pre-treatment periods between the time that the first state adopted the Act under no policy change (1975) and the time that the first state adopted the Act.

2. The post-treatment periods for the last state that adopted the Act.

3. Any periods between 1. and 2.

Notice the post-treatment (1989-2023) is too long for a discussion of the beer consumption effect, and since we do not add time or state dummies except for TWFE-DID, we would like to confine the total number of post-treatment periods within 10. Therefore, we establish a balanced-panel dataset with 25 states from 1975 to 1995.

One remark that we want to mention is: Some of the states have exceptions regarding drinking age. For example, in 2017, North Carolina signed the "Brunch Bill", making it legal for bars, restaurants, and retail stores to sell alcohol earlier on Sundays. For simplicity, we ignore these subtle differences between states local policies. The ignorability is justified since most of the discrepancies between local policies took place after 20^{th} century, beyond the time periods we included in our dataset.

6.3 Results

Table 1 describes the summary statistics of the outcome variable *beer*. We can see a giant spread of *beer* and a huge discrepancy between the maximum and minimum values of *beer*. This may indicate outlier states in beer consumption, which makes sense due to the difference in population and alcoholic culture. At this point, we do not want to remove outliers and we will explain this later in this section.

Let's first consider the validity of our baseline models TWFE-DID and CS-DID. Notice it is obvious that the TWFE-DID would be biased, since it is under the staggered adoption framework and by Theorem A.1. Table **11** is the outcome path plot of *beer* for Alabama (AL), Arizona (AZ), and Indiana (IN) which are in group 1986, 1985, and NV respectively. This path plot shows no obvious parallel trend ($\equiv$ conditional parallel trend condition since

we do not have covariates) during the pretreatment windows of any two of these three states. Therefore, neither Conditional Parallel Trends Based on "Not-Yet-Treated"/"Never-Treated" Groups (Assumptions 5 or 4) is triggered. Thus, we would expect CS-DID to be biased.

Table 2 shows that the $\hat{\tau}_{did} = -4,610,101$, which is statistically significant only under 10% level of significance. Table 3 shows that the $\hat{\tau}_{csdid} = -725,521.3$, which is statistically insignificant under all standard levels of significance, and we know both of these two baseline estimation methods are biased.

Moving on to SCM-CS-DID, LASSO-CS-DID, and LASCM-CS-DID. Table 4 shows the estimated τ under Trajectory Fit Criterion $= 0.1$ using the three tools and the corresponding selected states chosen by the mechanism described in the second simulation. First, notice an abnormal $\hat{\tau} = 121,524.5$ by LASSO-CS-DID. As demonstrated in the second simulation, this is an expected result since LASSO can sometimes produce a wonderful Trajectory Fit even if the state is having abnormal distributions/patterns. For example, AL and AZ are in the selected states in LASSO-CS-DID, but from Figure **11**, we can directly observe abnormalities in the outcome paths of AL and AZ: Given that they adopted the Act in 1986 and 1985 respectively, the beer consumption increased (though with a decreasing rate) in the following year in AL, and increased largely with no decreasing rate in AZ in the following year after the adoption. The beer consumption in AZ decreased dramatically in 1991 which is 5 years after the adoption. Without controlling the covariates, there seems to be no reason that AL and AZ should be in the selected states. However, as Table 5 demonstrates, the Trajectory Fits by LASSO, especially that of AZ, are fantastic, so they stay in the pool and cause problems, which makes LASSO-CS-DID invalid.

In comparison to LASSO-CS-DID which uses states selected by LASSO, SCM-CS-DID and LASCM-CS-DID use states selected by SCM, which ensures interpolation and excludes abnormal states. Returning to the example of AL and AZ, we can see in Table 5, that they are both eliminated by SCM, since SCM cannot give a good pre-treatment Trajectory Fit

less than TF Criterion based on all other states in the original dataset. Also, recall that we choose to keep outliers in the dataset. Though outliers can have an inherent abnormal distribution (which causes problems), they could have the same distribution as the other states with an abnormal parameter. For example, AL is the largest state, which implies a relatively large beer consumption. But it could still have the same underlying distribution of beer consumption as some others states with different parameters. We do not want to arbitrarily drop any information that is available to us, since it is possible that an "outlier" could be a good component of a synthetic counterpart, decided by SCM, for some treated states. One immediate result of this is that SCM-CS-DID and LASSO-CS-DID are robust to outliers and can utilize more information than those estimation methods which are not. This property is desirable when there are a few individuals in the dataset, since any removal will lead to a relatively large proportional drop in the number of the observations. Table 5 also discloses the similarity between SCM-CS-DID and LASCM-CS-DID. This is consistent with the results shown in the second simulation, since the set of states used in LASCM-CS-DID is the same as SCM-CS-DID. By Proposition 5.1, we are good to conclude an average amount of $-2,270,766$ on gallons of beer by The 1984 National Minimum Drinking Age Act.

7 Discussion and Limitations

There are several limitations that we recognize and want to demonstrate in this section.

Firstly, recall in the simulation DGP, we construct the outcome variables of the early-treated, late-treated, and never-treat groups to satisfy SCM conditions. We recognize that DGP are manipulated and unrealistic. As suggested by Abadie (2021), SCM should not be considered if the pre-treatment trajectory is not fitted well. So one should disregard SCM-CS-DID and LASCM-CS-DID if they found no units are selected even under large TF.

Secondly, Proposition 5.1 is not fully generalized in my opinion. Again, we invite researchers for generalization.

Thirdly, in the case study, there are a bunch of issues that need discussion and elaboration. We include, for simplicity, only the outcome variable *beer* in our dataset. However, we admit covariates (e.g. population) can play an inseparable role in giving an unbiased estimate of the effect the 1984 Act on beer consumption. Moreover, there could have been anticipation effect in the data, since the citizens in the late-adopted states should have changed their beer consumption behavior based on information from the citizens from in the early-adopted states. Also, Table 4 indicates only five states passed the TF score, which means a huge reduction in observations for CS-DID.

Finally, we would like to invite researchers to further SCM-CS-DID and LASCM-CS-DID with valid inference techniques.

References

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

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *American Statistical Association 105*(490), 493–505.

Abadie, A., A. Diamond, and J. Hainmueller (2015). Comparative politics and the synthetic control method. *American Journal of Political Science 59*(2), 495–510.

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

Ben-Michael, E., A. Feller, and J. Rothstein (2021, June). Synthetic controls with staggered adoption. Working Paper 28886, National Bureau of Economic Research.

Berger, D. E. and J. R. Snortum (1985). Alcoholic beverage preferences of drinking-driving violators. *Journal of Studies on Alcohol 46*(3), 232–239. PMID: 4010301.

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230. Themed Issue: Treatment Effect 1.

Carpenter, C. and C. Dobkin (2011). The minimum legal drinking age and public health. *The journal of economic perspectives*.

Doudchenko, N. and G. W. Imbens (2016, October). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Working Paper 22791, National Bureau of Economic Research.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277. Themed Issue: Treatment Effect 1.

Grant, D. P. (1984). Evidence and evaluation: The national minimum drinking age act of 1984. *Sam Houston State University*.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199. Themed Issue: Treatment Effect 1.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Toomey, T. L., C. Rosenfeld, and A. C. Wagenaar (1996). The minimum legal drinking age: History, effectiveness, and ongoing debate. *Alcohol Health and Research World 20*(4), 213–218.

A Appendix

A.1 Assumptions and Theorems

(The following notations in Assumption 1 to 6 and Theorem A.2 come from Callaway and Sant'Anna (2021))

Assumption 1 (Irreversibility of Treatment). $D_1 = 0$ almost surely (a.s.). For $t = 2, \ldots, \mathcal{T}$, $D_{t-1} = 1$ implies that $D_t = 1$ a.s.

Assumption 2 (Random Sampling). $\{Y_{i,1}, Y_{i,2}, \ldots, Y_{i,\mathcal{T}}, X_i, D_{i,1}, D_{i,2}, \ldots, D_{i,\mathcal{T}}\}_{i=1}^{n}$ is independent and identically distributed (i.i.d.).

Assumption 3 (Limited Treatment Anticipation). There is a known $\delta \geq 0$ such that

$$\mathbb{E}\left[Y_t(g) \mid X, G_g = 1\right] = \mathbb{E}\left[Y_t(0) \mid X, G_g = 1\right] \text{ a.s. for all } g \in \mathcal{G}, t \in \{1, \ldots, \mathcal{T}\} \text{ such that } t < g - \delta$$

Assumption 4 (Conditional Parallel Trends Based on a "Never-Treated" Group). Let δ be as defined in Assumption 3. For each $g \in \mathcal{G}$ and $t \in \{2, \ldots, \mathcal{T}\}$ such that $t \geq g - \delta$

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \mid X, G_g = 1\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \mid X, C = 1\right] \text{ a.s.}$$

Assumption 5 (Conditional Parallel Trends Based on "Not-Yet-Treated" Groups). Let δ be as defined in Assumption 3. For each $g \in \mathcal{G}$ and each $(s, t) \in \{2, \ldots, \mathcal{T}\} \times \{2, \ldots, \mathcal{T}\}$ such that $t \geq g - \delta$ and $t + \delta \leq s < \bar{g}$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \mid X, G_g = 1\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \mid X, D_s = 0, G_g = 0\right] \text{ a.s.}$$

Assumption 6 (Overlap). For each $t \in \{2, \ldots, \mathcal{T}\}, g \in \mathcal{G}$, there exist some $\varepsilon > 0$ such that $P\left(G_g = 1\right) > \varepsilon$ and $p_{g,t}(X) < 1 - \varepsilon$ a.s.

We propose our Assumptions 7 and 8 for SCM-CS-DID, LASSO-CS-DID, and LASCM-CS-DID to hold with the notations in Callaway and Sant'Anna (2021).

Assumption 7 (Treated Units in Convex Hull of Not-Yet-Treated/Never-Treated Units). For each $i \in \{i : G_i \notin \{\infty \cup \bar{g}\}\}$, let $\mathfrak{J} = \{j : G_j < G_i\}$ and $t \in \{2, \ldots, \min\{G_n : n \in \{G_n > G_i\}\}\} = \mathfrak{T}$; for i s.t. $G_i = \bar{g}$, let $\mathfrak{J} = \{j : G_j = \infty\}$ and $t \in \{1, \ldots, n\} = \mathfrak{T}$

$$Y_{i,t}(0) = \omega^T Y_{\mathfrak{J},t}(0)$$

where $\omega \in \mathbb{R}^{\#\mathfrak{J}}, \omega \geq 0, \sum \omega_i = 1$.

Assumption 8 (Representation of Treated Units as a Linear Combination of Not-Yet-Treated/Never-Treated Units). For each $i \in \{i : G_i \notin \{\infty \cup \bar{g}\}\}$, let $\mathfrak{J} = \{j : G_j < G_i\}$ and $t \in \{2, \ldots, \min\{G_n : n \in \{G_n > G_i\}\}\} = \mathfrak{T}$; for i s.t. $G_i = \bar{g}$, let $\mathfrak{J} = \{j : G_j = \infty\}$ and $t \in \{1, \ldots, n\} = \mathfrak{T}$

$$Y_{i,t}(0) = \omega^T Y_{\mathfrak{J},t}(0)$$

where $\omega \in \mathbb{R}^{\#\mathfrak{J}}, \sum \omega_i \leq \lambda$, where $\lambda \in \mathbb{R}_+$

Notice Assumptions 7 and 8 are essential conditions under which Theorem A.2 will hold for unit i. In other words, for each unit i defined in Assumptions 7, we can represent the outcome variable of i from period 1 to the start of the next treated group as a linear combination of Not-Yet-Treated/Never-Treated units, and the weights are subject to non-negativity and summing-up-to-one. Therefore, Assumption 4 will be invoked since for $g \in \mathcal{G} \cap \mathfrak{T}$:

$$\begin{aligned}
\mathbb{E}\left[Y_{i,t}(0) - Y_{i,t-1}(0) \mid X_i, G_{i,g} = 1\right] &= \mathbb{E}\left[\omega^T Y_{\mathfrak{J},t}(0) - \omega^T Y_{\mathfrak{J},t-1}(0) \mid X_i, G_{i,g} = 1\right] \\
&= \mathbb{E}\left[\omega^T Y_{\mathfrak{J},t}(0) - \omega^T Y_{\mathfrak{J},t-1}(0) \mid X_i, G_{\mathfrak{J},g} = 0\right] \text{ a.s.} \\
&= \mathbb{E}\left[Y_{i,t}(0) - Y_{i,t-1}(0) \mid X_i, C = 1\right] \text{ a.s.}
\end{aligned}$$

where the first equality is by Assumption 7. The second equality is by Assumption 7 and the fact that $G_{i,g} = 1 \implies G_{\mathfrak{J},g} = 0$, The third equality is by Assumption 7 again and the definition of C when the time window is restricted to $\mathfrak{T}$. Notice $Y_{\mathfrak{J},t}(0)$, by construction, are

observable. Therefore, Theorem A.2 holds for unit i.

The argument for Assumption 8 is similar.

Remark. If we filter the pair of $\{Y_{i,t}(0), \omega^T Y_{\mathfrak{J},t}(0)\}$ based on TF_i, then Assumption 2 can be relaxed to $\{Y_{i,1}, Y_{i,2}, \ldots, Y_{i,\mathcal{T}}, X_i, D_{i,1}, D_{i,2}, \ldots, D_{i,\mathcal{T}}\}_{i=1}^{m}$ is independent and identically distributed (i.i.d.). for some $1 < m \leq n$. (i.e. we allow abnormalities as discussed in Section 3 and 6)

Theorem A.1 (Difference-in-Differences Decomposition Theorem)**.** Assume that the data contain $k = 1, \ldots, K$ timing groups of units ordered by the time when they receive a binary treatment, $k \in (1, T]$. There may be one timing group, U, that includes units that never receive treatment. The OLS estimate, $\hat{\beta}^{DD}$, in a two-way fixed-effects regression is a weighted average of all possible TWFE-DID estimators.

$$\hat{\beta}^{DD} = \sum_{k \neq U} s_{kU} \hat{\beta}^{2\times2}_{kU} + \sum_{k \neq U} \sum_{\ell > k} \left[s^k_{k\ell} \hat{\beta}^{2\times2,k}_{k\ell} + s^\ell_{k\ell} \hat{\beta}^{2\times2,\ell}_{k\ell} \right].$$

where the TWFE-DID estimators are:

$$\begin{aligned}
\hat{\beta}^{2\times2}_{kU} &\equiv \left(\bar{y}_k^{\text{POST }(k)} - \bar{y}_k^{PRE(k)} \right) - \left(\bar{y}_U^{\text{POST }(k)} - \bar{y}_U^{\text{PRE }(k)} \right), \\
\hat{\beta}^{2\times2,k}_{k\ell} &\equiv \left(\bar{y}_k^{\text{MID}(k,\ell)} - \bar{y}_k^{PRE(k)} \right) - \left(\bar{y}_\ell^{\text{MID }(k,\ell)} - \bar{y}_\ell^{\text{PRE }(k)} \right), \\
\hat{\beta}^{2\times2,\ell}_{k\ell} &\equiv \left(\bar{y}_\ell^{\text{POST }(\ell)} - \bar{y}_\ell^{\text{MID }(k,\ell)} \right) - \left(\bar{y}_k^{\text{POST }(\ell)} - \bar{y}_k^{\text{MID}(k,\ell)} \right).
\end{aligned}$$

The weights are:

$$\begin{aligned}
s_{kU} &= \frac{(n_k + n_U)^2 \overbrace{n_{kU}\left(1 - n_{kU}\right) \bar{D}_k \left(1 - \bar{D}_k\right)}^{\hat{V}^D}}{\hat{V}^D_{kU}}, \\
s^k_{k\ell} &= \frac{\left(\left(n_k + n_\ell\right)\left(1 - \bar{D}_\ell\right)\right)^2 \overbrace{n_{k\ell}\left(1 - n_{k\ell}\right)}^{\hat{V}^{D,k}_{k\ell}} \frac{\bar{D}_\ell}{1-\bar{D}_\ell} \frac{1-\bar{D}_k}{1-\bar{D}_\ell}}{\hat{V}^D},
\end{aligned}$$

and $\sum_{k \neq U} s_{kU} + \sum_{k \neq U} \sum_{\ell > k} \left[s^k_{k\ell} + s^\ell_{k\ell} \right] = 1$ (See Goodman-Bacon (2021) for more information)

Theorem A.2. Let Assumptions 1-3 and 6 hold. (i) If Assumption 4 holds, then, for all g and t such that $g \in \mathcal{G}_\delta, t \in \{2, \dots \mathcal{T} - \delta\}$ and $t \geq g - \delta$,

$$ATT(g,t) = ATT^{nev}_{ipw}(g,t;\delta) = ATT^{nev}_{or}(g,t;\delta) = ATT^{nev}_{dr}(g,t;\delta).$$

(ii) If Assumption 5 holds, then, for all g and t such that $g \in \mathcal{G}_\delta, t \in \{2, \dots \mathcal{T} - \delta\}$ and $g - \delta \leq t < \bar{g} - \delta$,

$$ATT(g,t) = ATT^{ny}_{ipw}(g,t;\delta) = ATT^{ny}_{or}(g,t;\delta) = ATT^{ny}_{dr}(g,t;\delta).$$

(See Callaway and Sant'Anna (2021) for more information)

A.2 Tables

Table 1: Variable Description and Summary Statistics

Variable	**Symbol**	**Description**	**Type**
Beer Consumption	*beer*	Total annual gallons of beer consumption in a particular state	Continuous
State ID	*state*	State ID variable	Discrete
Year	*year*	Time variable	Continuous

Variable	Summary Statistics				
	N	Mean	St. Dev.	Min	Max
beer	525	94,273,414	127,350,317	7,829,000	691,050,432

Table 2: **Two-Way-Fixed-Effect Difference-in-Difference Estimation of the Effect of the 1984 Act on Beer Consumption**

	Dependent variable:	
	beer	
	Mean	Standard Deviation
$\hat{\tau}_{did}$	$-4,610,101^{*}$	$2,059,478$
Observations	525	
R^2	0.991	
Adjusted R^2	0.9901	
Residual Std. Error	12,650,000 (df = 479)	
F Statistic	1,170 *** (df = 45 ; 479)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 3: **Callaway-Sant'Anna Difference-in-Difference Estimation of the Effect of the 1984 Act on Beer Consumption**

	Dependent variable:	
	beer	
	Mean	Standard Deviation
$\hat{\tau}_{csdid}$	$-725,521.3$	$1,236,963$
Observations	525	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 4: **SCM/LASSO/LASCM-CS-DID Estimation of the Effect of the 1984 Act on Beer Consumption**

Model	$\hat{\tau}$	Selected States
SCM-CS-DID	-2,119,416	CO, DE, IL, MD, VT
LASSO-CS-DID	121,524.5	AL, AK, AZ, CO, DE
		HI, ID, IL, KS, MD
		MS, VT, WY
LASCM-CS-DID	-2,270,766	CO, DE, IL, MD, VT

Note: The Trajectory Fit Criterion is set to be 0.1. The number of observation for SCM-CS-DID is 59; The number of observation for LASSO-CS-DID is 161; The number of observation for LASCM-CS-DID is 59

Table 5: **States Selected by LASSO/SCM Based on Trajectory Fit**

state	TF by LASSO	TF by SCM
AK	0.051	-
AL	0.015	-
AZ	0.037	-
CO	0.081	0.068
DE	0.011	0.039
HI	0.051	-
ID	0.013	-
IL	0.017	0.014
KS	0.002	-
MD	0.024	0.094
MS	0.016	-
VT	0.020	0.096
WY	0.028	-

Note: The Trajectory Fit Criterion is set to be 0.1. The symbol "-" stands for states that are selected by LASSO, but not selected by the SCM. The number of observation for SCM and LASSO is 525

A.3 Figures

Figure 1: Simulation: Estimation Methods Under Uniform/Staggered Treatment Timing and Treatment Effect Homogeneity/Heterogeneity - Trends in Outcome Path

This figure plots the average outcome paths by early-treated (line 1), late-treated (line 2 if applicable), and never-treated groups (line 0) for unit $N = 50$ over time $T = 30$. The outcome variable is generated under a linear factor model, increasing linearly in time t, with a noise term $\sim_{i.i.d} N(0, 1)$. Simulation 1 and 2 are under no staggered adoption setting, and the treatment happens at the vertical dashed line $t = 10$, whereas they differs in terms of homogeneity of the treatment effect τ (Simulation 1 has a constant τ while Simulation 2 has a dynamic τ). Simulation 3 and 4 are under staggered adoption setting, and the early treatment takes place at the red vertical dashed line $t = 10$ and the late treatment happens at the green vertical dashed line $t = 20$. The difference across Simulations 3 and 4 is similar to that between Simulations 1 and 2.

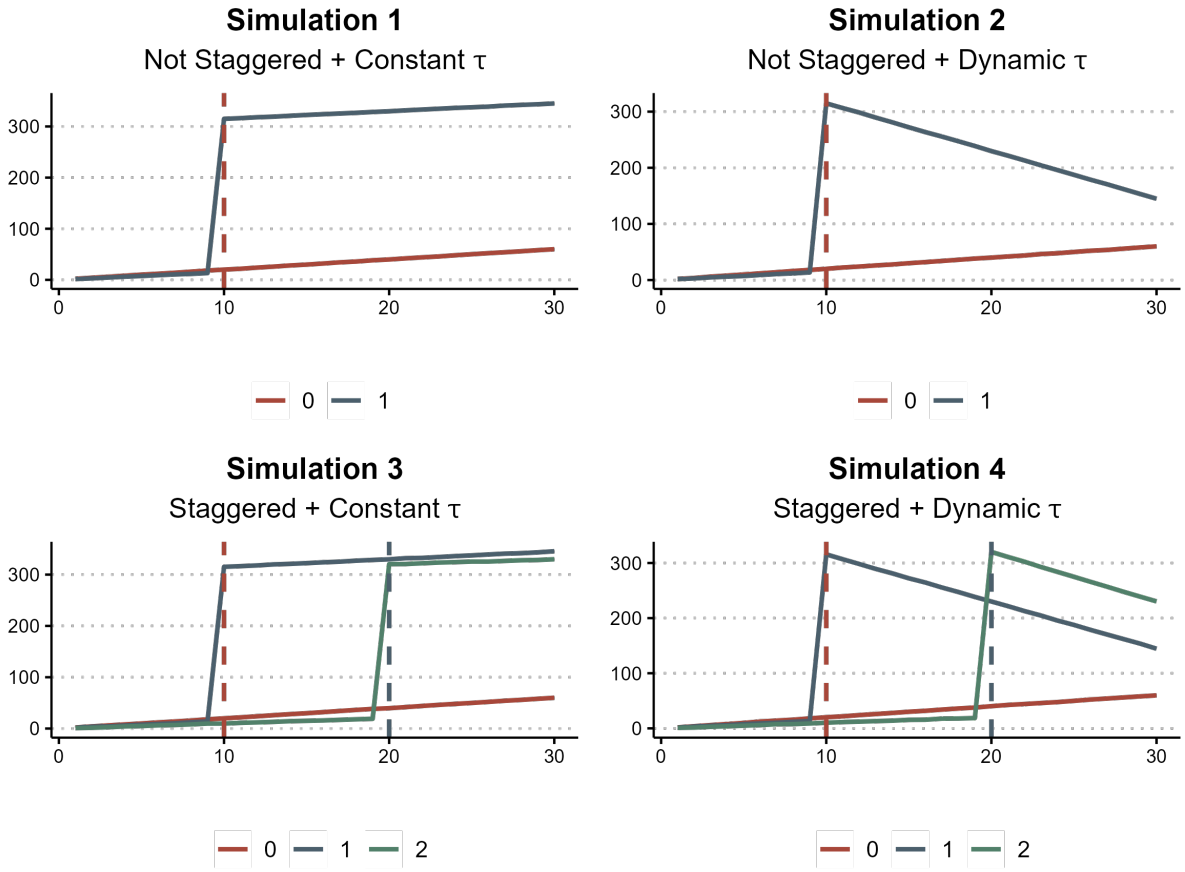


Figure 2: Simulation: Estimation Methods Under Uniform/Staggered Treatment Timing and Treatment Effect Homogeneity/Heterogeneity - TWFE-DID Density Plots

This figure draws kernel density estimate of the $\hat{\tau}^{DID}$ by the Two-way Fixed Effect Difference-in-Difference from 5,000 Monte Carlo simulations for each data generating processes. The distribution of the $\hat{\tau}^{DID}$ is represented by the curve, while the true $\tau = 300$ is indicated by the red vertical dashed line. In Simulation 1 and 3, TWFE-DID is biased towards 0. Additionally, in Simulation 2 and 4, the $\hat{\tau}^{DID}$ is almost 0 or even has a wrong sign (the density plots are invisible on the left of Simulation 2 and 4). Simulation 2 severely underestimates the τ since TWFE-DID cannot capture the dynamics in τ. Simulation 4 underestimates the τ, because the early-treated group is used as the control group in TWFE-DID estimation on the late-treated group in the late-treatment window (after $t = 20$). All simulations underestimates the τ since the parallel trend assumption is not hold by construction.

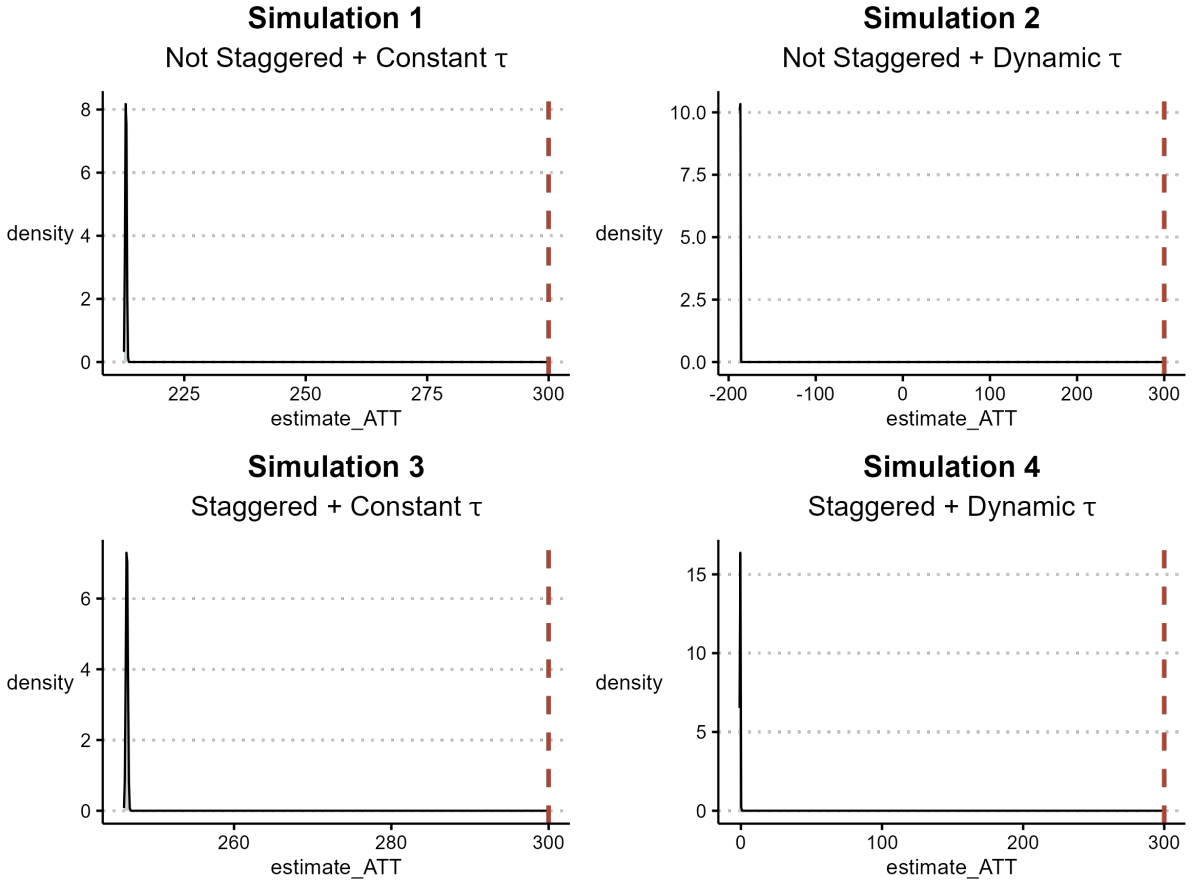


Figure 3: Simulation: Estimation Methods Under Uniform/Staggered Treatment Timing and Treatment Effect Homogeneity/Heterogeneity - SCM Density Plots

This figure draws kernel density estimate of the τ by Synthetic Control Method from 5,000 Monte Carlo simulations for each data generating processes. The distribution of the $\hat{\tau}^{SCM}$ is represented by the curve, while the true $\tau = 300$ is indicated by the red vertical dashed line. In Simulation 1 and 3, SCM gives unbiased but skewed estimates of τ. However, in Simulation 2 and 4, the $\hat{\tau}^{SCM}$ is biased towards 0 since SCM cannot capture dynamics in treatment effect.

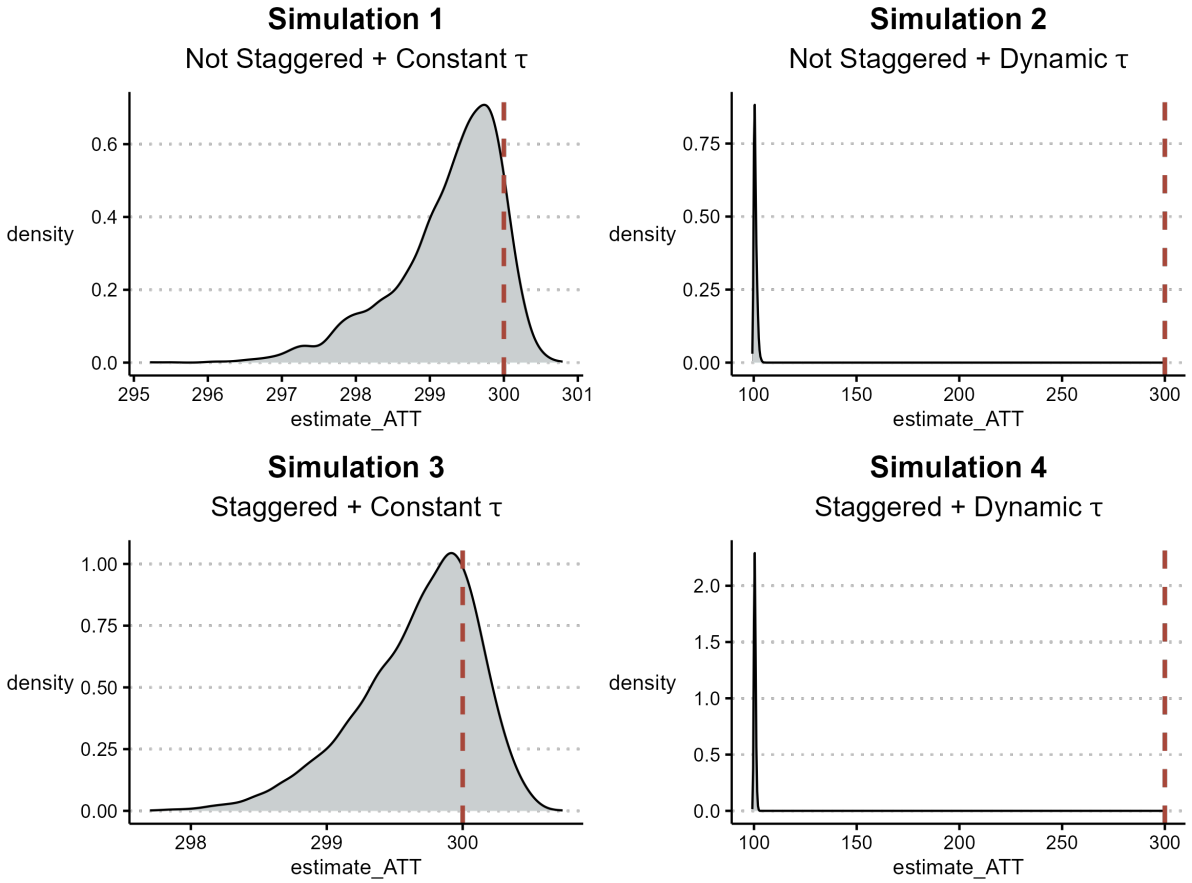


Figure 4: Simulation: Estimation Methods Under Uniform/Staggered Treatment Timing and Treatment Effect Homogeneity/Heterogeneity - CS-DID Density Plots

This figure draws kernel density estimate of the τ by the Callaway and Sant'Anna's Difference-in-Difference from 5,000 Monte Carlo simulations for each data generating processes. The distribution of the $\hat{\tau}^{CS}$ is represented by the curve, while the true $\tau = 300$ is indicated by the red vertical dashed line. In all simulations, CS-DID gives biased estimates of τ since the conditional parallel trend assumptions do not hold by construction.

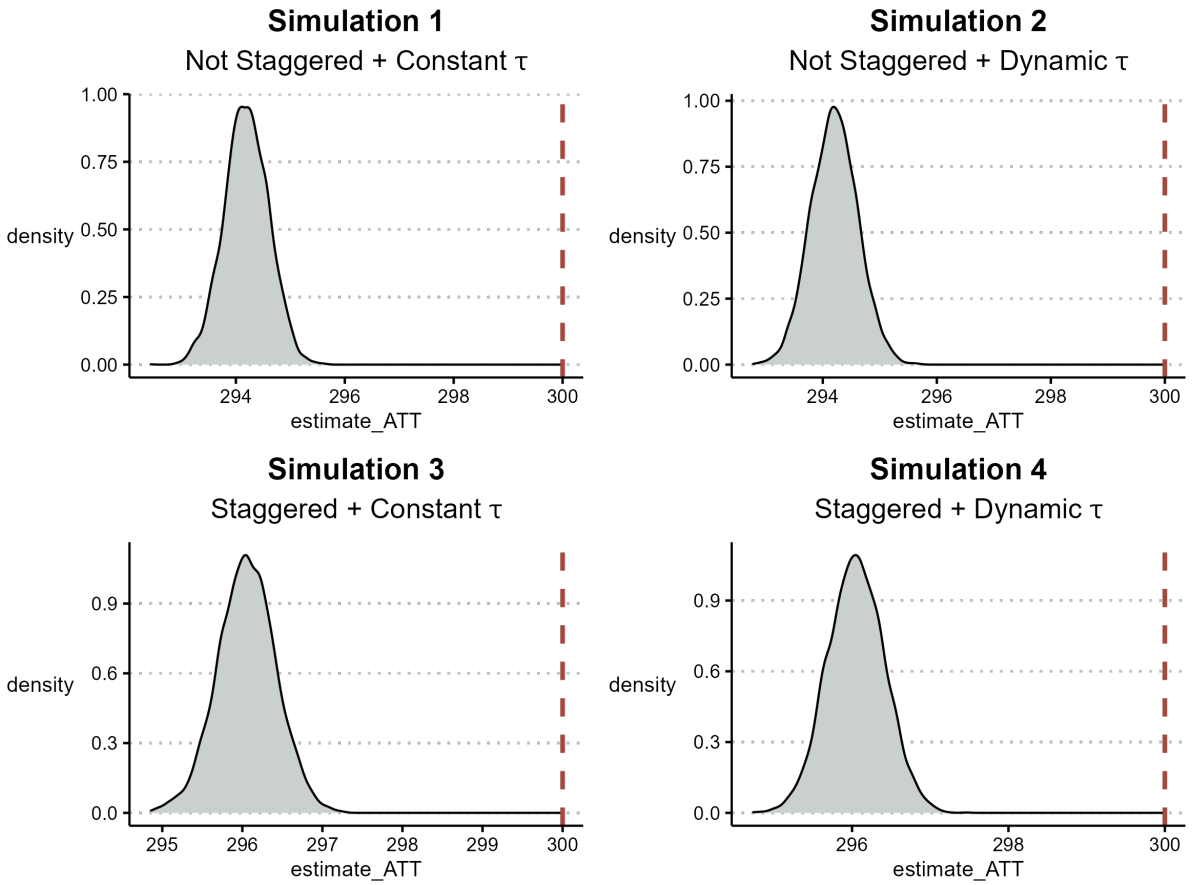


Figure 5: Simulation: Estimation Methods Under Uniform/Staggered Treatment Timing and Treatment Effect Homogeneity/Heterogeneity - SCM-DID Density Plots

This figure draws kernel density estimate of the τ by SCM-DID from 5,000 Monte Carlo simulations for each data generating processes. The distribution of the $\hat{\tau}^{SCM-DID}$ is represented by the curve, while the true $\tau = 300$ is indicated by the red vertical dashed line. In Simulation 1 and 3, SCM-DID gives unbiased estimates of τ. However, in Simulation 2 and 4, the $\hat{\tau}^{SCM-DID}$ is biased towards 0 since neither SCM nor TWFE-DID can capture dynamics in treatment effect.

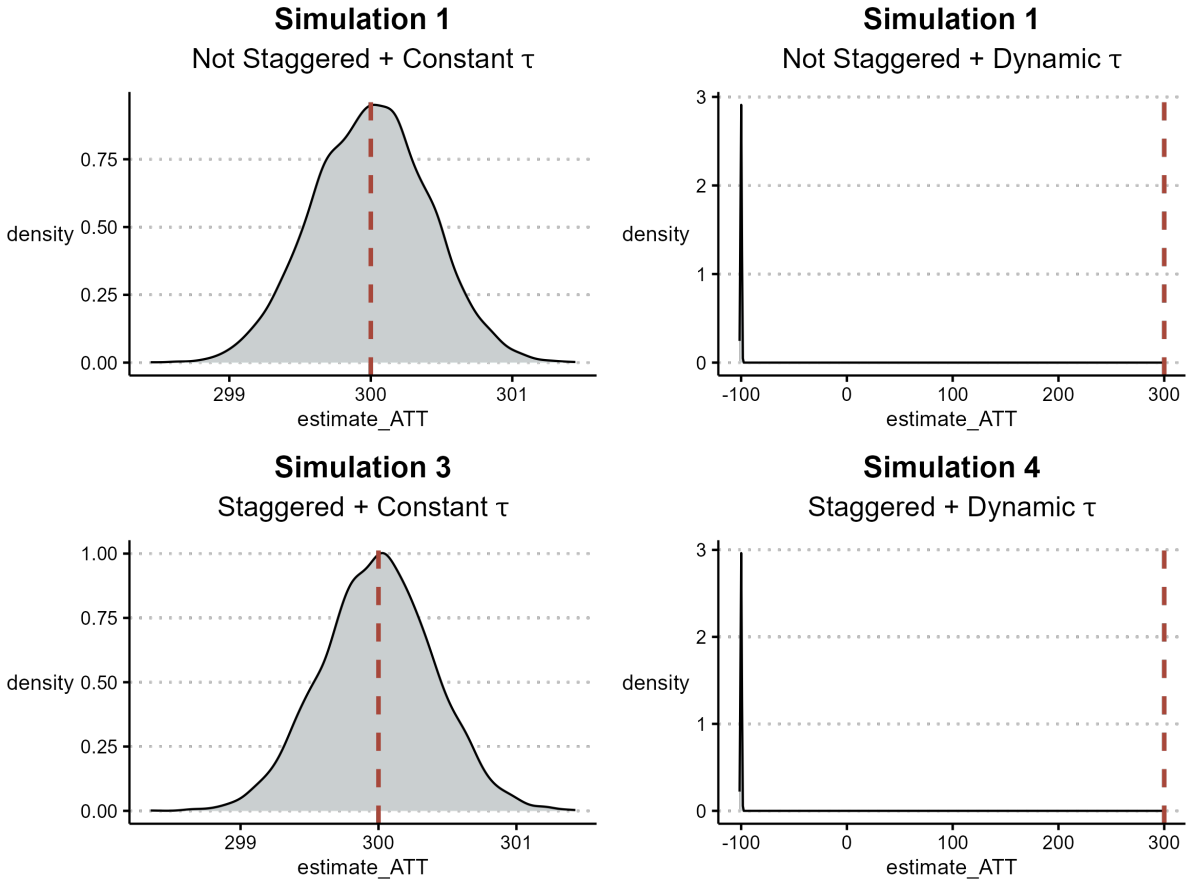


Figure 6: Simulation: Estimation Methods Under Uniform/Staggered Treatment Timing and Treatment Effect Homogeneity/Heterogeneity - SCM-CS-DID Density Plots

This figure draws kernel density estimate of the τ by SCM-CS-DID from 5,000 Monte Carlo simulations for each data generating processes. The distribution of the $\hat{\tau}^{CSCM}$ is represented by the curve, while the true $\tau = 300$ is indicated by the red vertical dashed line. In all simulations, SCM-CS-DID gives unbiased estimates of τ.

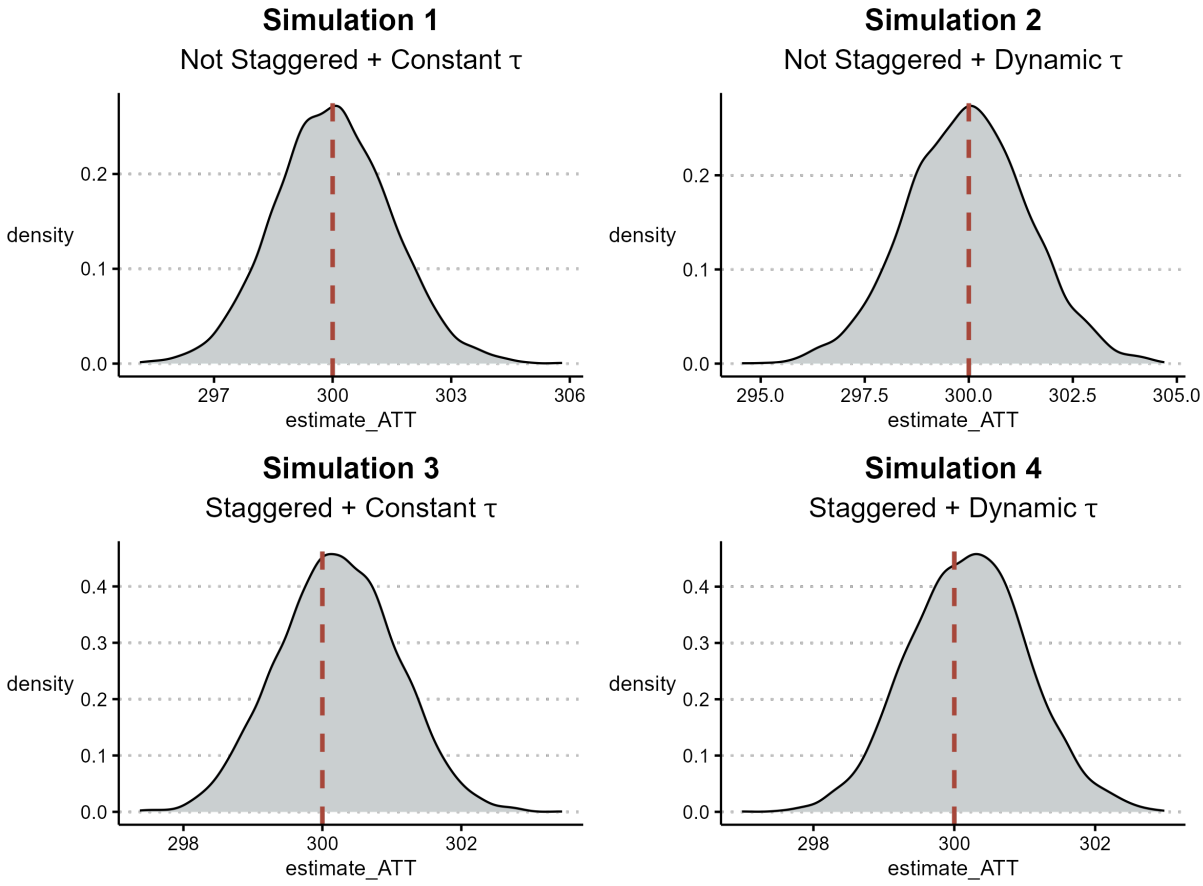


Figure 7: Simulation: Estimation Methods Under Uniform/Staggered Treatment Timing and Treatment Effect Homogeneity/Heterogeneity - LASSO-CS-DID Density Plots

This figure draws kernel density estimate of the τ by LASSO-CS-DiD from 5,000 Monte Carlo simulations for each data generating processes. The distribution of the $\hat{\tau}^{LASSO-CS-DID}$ is represented by the curve, while the true $\tau = 300$ is indicated by the red vertical dashed line. In all simulations, LASSO-CS-DID gives unbiased estimates of τ.

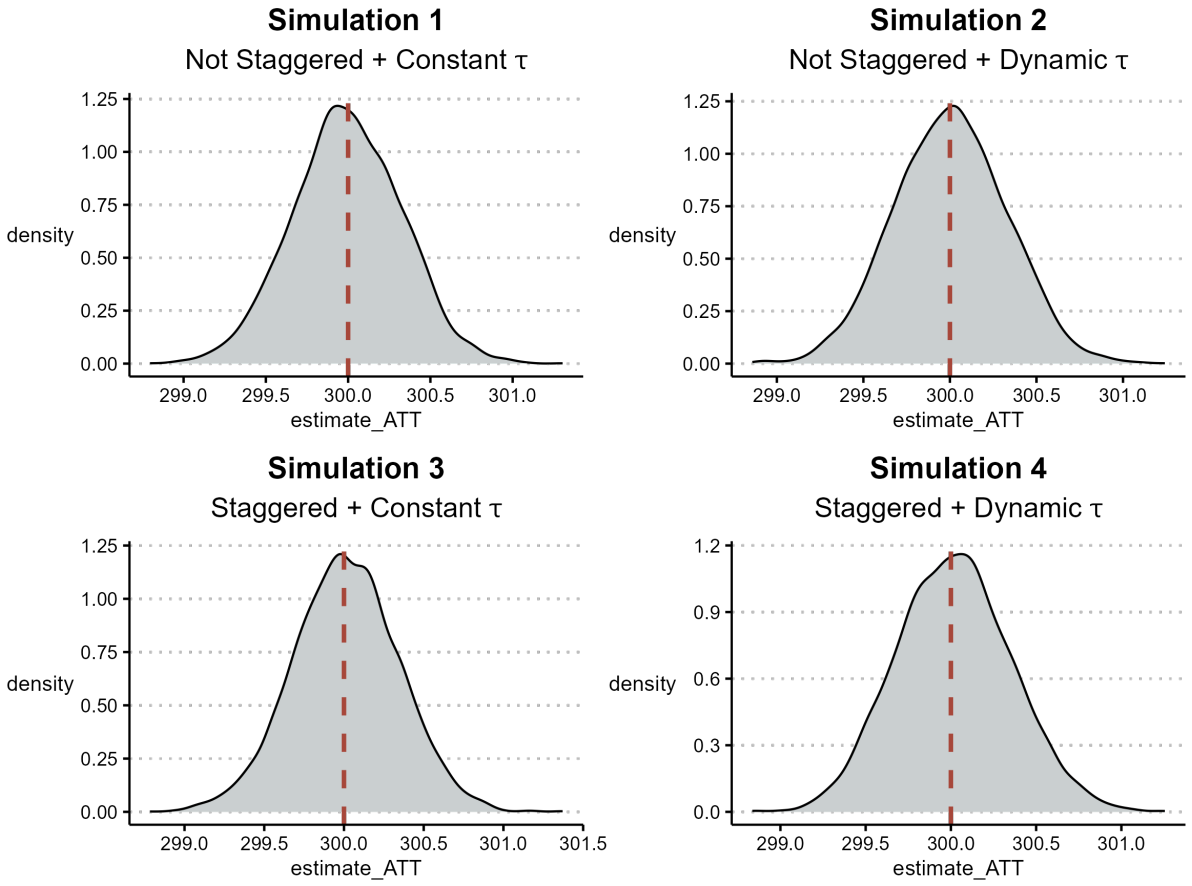


Figure 8: Simulation: Estimated ATT from Different Estimators against Trajectory Fit Criteria

This figure draws estimate of the τ by SCM-CS-DID, LASSO-CS-DID (using Cross Validation), LASSO-CS-DID (using BIC), and LASCM-CS-DID from 5,000 Monte Carlo simulations. The true $\tau = 300$. Notice that LASCM-CS-DID and SCM-CS-DID are overlapping and giving unbiased estimates of τ when TFC is less than 0.25, and gradually decreases when TFC increases. LASSO-CS-DID by CV gives a 292.61 estimate of τ when TFC is 0.02, and a disastrous estimate $= 5.02$ when TFC is 0.01. LASSO-CS-DID by BIC gives an unbiased estimate of τ when TFC is 0.02, but the estimate becomes biased dramatically with a slight increase in TFC from 0.02.

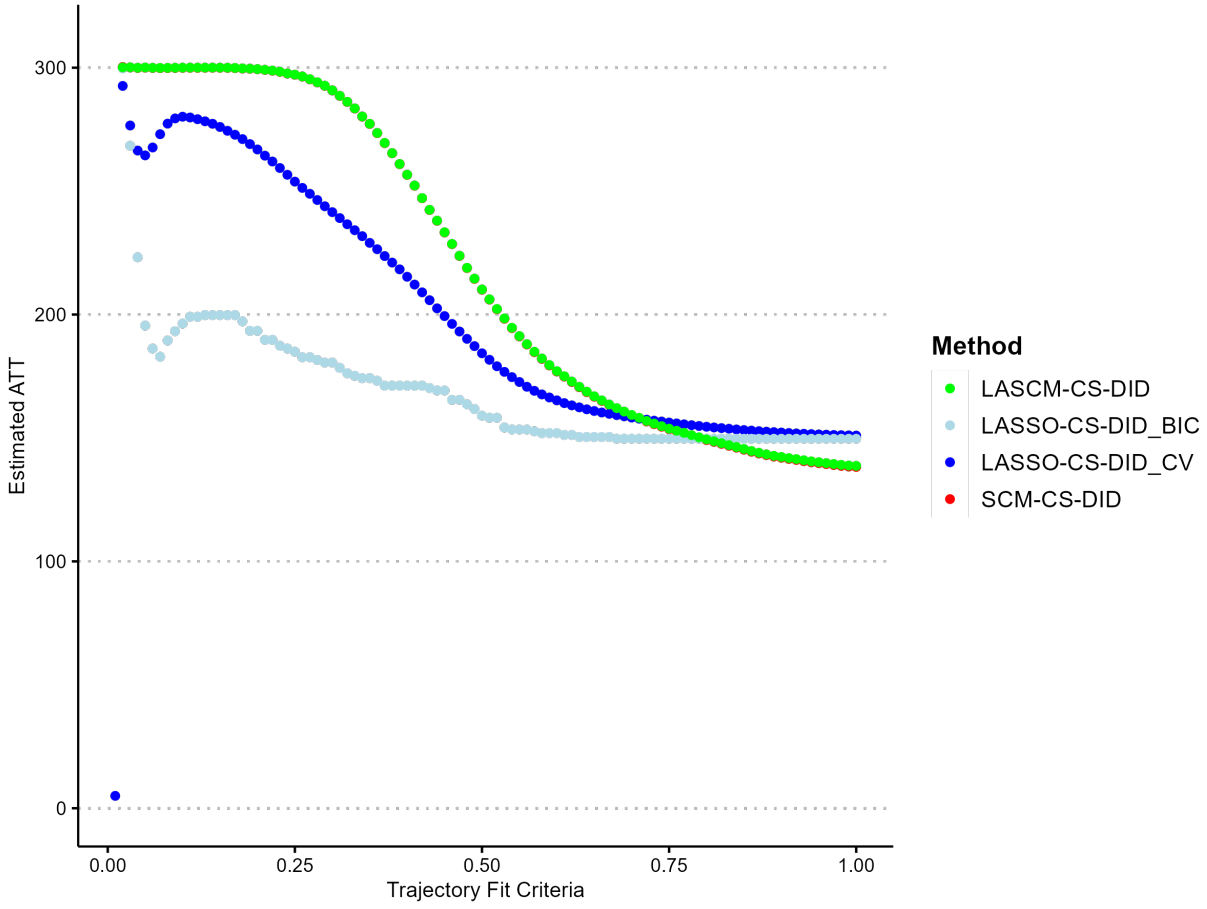
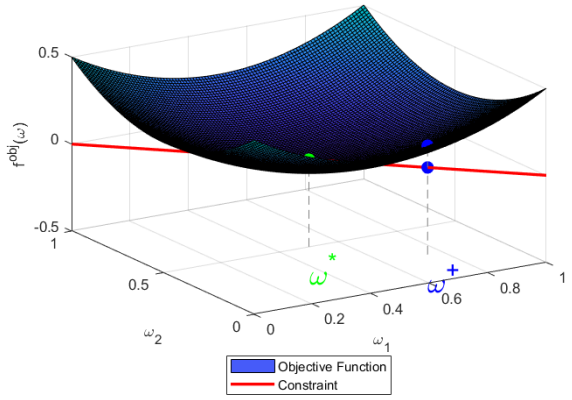
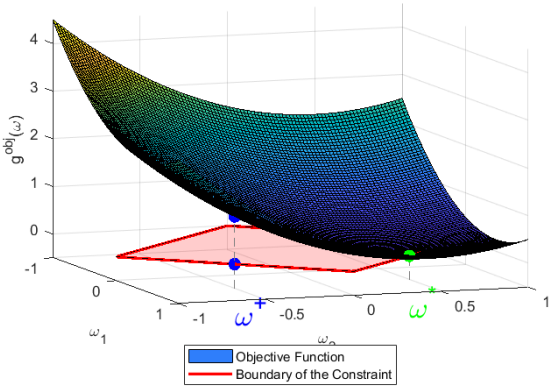


Figure 9: Graphical Illustration of the SCM, LASSO, and LCR

(a) The SCM



(b) The LASSO



(c) The LCR

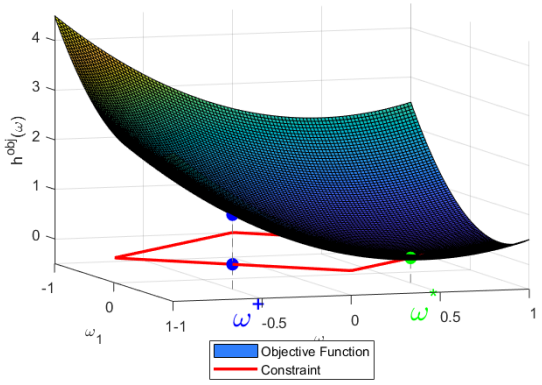


Figure 10: Stagggered Adoption of The 1984 National Minimum Drinking Age Act from 1980 to 1988

This figure shows the information of states included in our dataset. States marked with grey are ones that are removed from the dataset; States with orange are never-treated states (NV); States marked from light-green to blue are states adopted the Act on a continuous scale from 1980 to 1988.

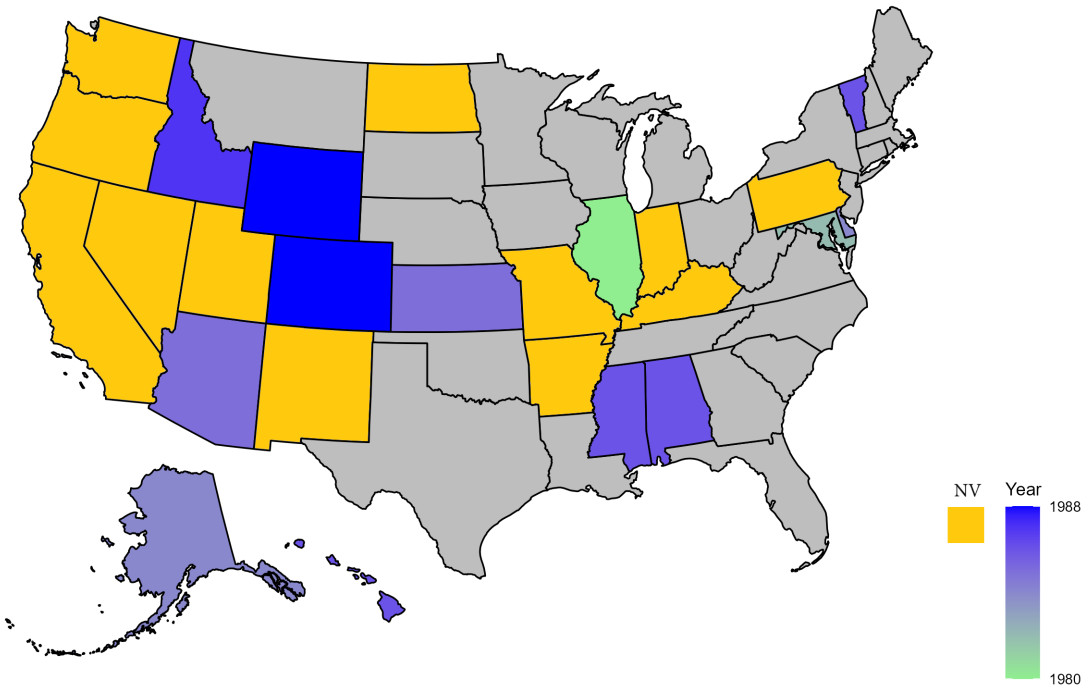


Figure 11: Outcome Path of Beer Consumption of Alabama, Arizona, and Indiana

This figure depicts the outcome path of gallons of beer consumption of Alabama (AL), Arizona (AZ), and Indiana (IN) which are in group 1986, 1985, and NV respectively. We observe no apparent parallel trends/conditional parallel trends between any two of these three states, and this violates the assumption of CS-DID.