

WORKING PAPER · NO. 2019-144

Randomization Tests in Observational Studies with Staggered Adoption of Treatment

Azeem M. Shaikh and Panos Toulis

DECEMBER 2019

Randomization Tests in Observational Studies with Staggered Adoption of Treatment*

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

Panos Toulis
Booth School of Business
University of Chicago
ptoulis@chicagobooth.edu

December 22, 2019

Abstract

This paper considers the problem of inference in observational studies with time-varying adoption of treatment. In addition to an unconfoundedness assumption that the potential outcomes are independent of the times at which units adopt treatment conditional on the units' observed characteristics, the main restriction underlying our analysis is that the time at which each unit adopts treatment follows a Cox proportional hazards model. This assumption permits the time at which each unit adopts treatment to depend on the observed characteristics of the unit, but imposes the restriction that the probability of multiple units adopting treatment at the same time is zero. In this context, we study Fisher-style randomization tests of a “sharp” null hypothesis that there is no treatment effect for all units and all time periods. We first show that an infeasible test that treats the parameters of the Cox model as known has rejection probability no greater than the nominal level in finite samples. We then establish that the feasible test that replaces these parameters with consistent estimators has limiting rejection probability no greater than the nominal level. These tests rely upon an important implication of the Cox model that provides a parametric expression for the probability that a particular unit is the first to adopt treatment conditional on both the observed characteristics and the time of first treatment. If these probabilities are equal across all units, then our testing procedure reduces to the randomization test proposed by [Abadie et al. \(2010\)](#). In a simulation study, we examine the practical relevance of our theoretical results, including robustness to misspecification of the model for the time at which each unit adopts treatment. Finally, we provide an empirical application of our methodology using the synthetic control-based test statistic and tobacco legislation data found in [Abadie et al. \(2010\)](#)

KEYWORDS: Randomization test, synthetic controls, time-varying treatment adoption, sharp null hypothesis, Cox proportional hazards model, observational data

JEL classification codes: C12, C14

*The research of the first author is supported by NSF Grant SES-1530661.

1 Introduction

This paper considers the problem of inference in observational studies in which units adopt treatment at varying times and remain treated once adopting treatment. The widespread availability of data with this type of structure has led to the development of several different methods for its analysis, including difference-in-differences (Snow, 1855; Card and Krueger, 1993) and, more recently, synthetic controls (Abadie and Gardeazabal, 2003; Abadie et al., 2010; Abadie, 2019). For a modern overview, see Section 5 of Abadie and Cattaneo (2018); further references are provided below. In contrast to the literature on difference-in-differences, which has focused largely on the estimation of certain average effects of the treatment on the outcome of interest, we study Fisher-style randomization tests of the “sharp” null hypothesis that there is no effect of the treatment on the outcome of interest for all units and all time periods. This type of null hypothesis has appeared previously in the literature on synthetic controls (Abadie et al., 2010; Firpo and Possebom, 2018), but, in contrast to this literature, which has focused primarily on settings in which there is only a single unit that adopts treatment, our testing procedure exploits the availability of multiple units that adopt treatment at different times in a novel way that we describe further below.

In addition to an unconfoundedness assumption that the potential outcomes are independent of the times at which units adopt treatment conditional on observed characteristics, the main restriction underlying our analysis is a survival model for the time at which each unit adopts treatment. We require, in particular, that the time at which each unit adopts treatment follows a Cox proportional hazards model. This assumption permits the time at which each unit adopts treatment to depend on the observed characteristics of the unit, but restricts the probability of multiple units adopting treatment at the same time is zero. For this reason, as discussed further in Remark 2.1 below, we view our methodology as being best suited for settings in which time is measured with sufficient granularity to make this probability small. Even though the Cox proportional hazards model is semiparametric, it further implies a parametric restriction on the distribution of the identity of the unit that first adopted treatment conditional on observed characteristics and the time of such first treatment adoption. As in Athey and Imbens (2018), our analysis is “design-based” in the sense that it exploits heavily this more subtle implication of our assumption on the distribution of the times at which each unit adopts treatment.

In order to motivate our proposed testing procedure, our first result shows that an infeasible test that treats the parameters of this conditional distribution of the identity of the unit that first adopted treatment — more succinctly, the “first adopter” — as known has rejection probability no greater than the nominal level in finite samples. We then establish that the feasible test that replaces these parameters with consistent estimators has limiting rejection probability no greater than the nominal level. We emphasize that consistent estimation of these parameters relies upon the

availability of multiple units that adopt treatment at different times. As mentioned previously, this feature distinguishes our analysis from some analyses found in the literature on synthetic controls in which there is only a single treated unit. We also note that our method for establishing this limiting result is novel in that it relies upon results in [Romano and Shaikh \(2012\)](#) to argue that the difference in the rejection probabilities of the feasible and infeasible tests tends to zero as the number of units becomes large.

Our analysis is most closely related to randomization tests that have appeared in the literature on synthetic controls. If the conditional distribution of the identity of the first adopter in our analysis is uniform, then our testing procedure reduces to the randomization test proposed by [Abadie et al. \(2010\)](#). In the same spirit as the literature on sensitivity analysis in observational studies ([Rosenbaum, 1987, 2002](#)), [Firpo and Possebom \(2018\)](#) parameterize this conditional distribution in order to explore the sensitivity of the testing procedure to deviations from the assumption that it is uniform. The parameterization used by [Firpo and Possebom \(2018\)](#) resembles our expression for the conditional distribution of the identity of the unit that first adopted treatment. In this way, a by-product of our analysis is an alternative viewpoint on their specific parameterization. [Chernozhukov et al. \(2019\)](#) have also recently proposed randomization tests for synthetic controls, but their analysis exploits exchangeability assumptions across the temporal dimension, whereas we, as explained previously, exploit a cross-sectional restriction of our survival model on the conditional distribution of the identity of the first adopter. Other recent proposals for inference in the context of synthetic controls include [Cattaneo et al. \(2019\)](#), who propose methods based on concentration inequalities from the high-dimensional statistics literature, and [Li \(2019\)](#), who propose methods based on subsampling that are applicable whenever both the number of pre-treatment and number of post-treatment time periods are large.

The remainder of the paper is organized as follows. In [Section 2](#), we describe our setup and notation. In particular, there we describe the Cox proportional hazards model that we use to model the time at which units adopt treatment and the resulting parametric restriction on the distribution of the identity of the first adopter unit conditional on observed characteristics, and the time of first treatment adoption. [Section 3](#) contains our main results, beginning with the finite-sample result for the infeasible test before presenting the large-sample result for the feasible test. In [Section 4](#), we explore the practical relevance of our theoretical results, including robustness to misspecification of the model for the time at which each unit adopts treatment. Finally, in [Section 5](#), we provide an empirical application of our methodology using the synthetic control-based test statistic and tobacco legislation data found in [Abadie et al. \(2010\)](#). Proofs of all results can be found in the [Appendix](#).

2 Setup and Notation

We index units by $i \in \mathbb{N} = \{1, \dots, n\}$ and time by $t \in \mathbb{T} = \{1, \dots, t_{\max}\}$. Denote by X_i the time-invariant characteristics of the i th unit. We assume that the treatment status of unit i at time t is given by

$$I\{t \geq T_i\}, \quad (1)$$

where $T_i \in [0, \infty)$ is the time at which unit i adopts treatment. In other words, treatment can not be “rolled back” once adopted. We further assume that there is no interference between treatment statuses, so we denote by $Y_{i,t}(1)$ and $Y_{i,t}(0)$ the potential outcomes of unit i at time t under treatment and control, respectively. The observed outcome of unit i at time t is denoted by $Y_{i,t}$, which satisfies

$$Y_{i,t} = Y_{i,t}(1)I\{T_i \leq t\} + Y_{i,t}(0)I\{T_i > t\}. \quad (2)$$

It will be convenient below to adopt the following shorthand notation:

$$\begin{aligned} Y^{(n)}(0) &= (Y_{i,t}(0) : i \in \mathbb{N}, t \in \mathbb{T}) \\ Y^{(n)}(1) &= (Y_{i,t}(1) : i \in \mathbb{N}, t \in \mathbb{T}) \\ Y^{(n)} &= (Y_{i,t} : i \in \mathbb{N}, t \in \mathbb{T}) \\ X^{(n)} &= (X_i : i \in \mathbb{N}) \\ T^{(n)} &= (T_i : i \in \mathbb{N}). \end{aligned}$$

Note that $Y^{(n)}$ and $X^{(n)}$ are observed, but adoption times may be censored. Indeed, T_i is not observed whenever $T_i > t_{\max}$, so only $T_i \vee t_{\max}$ is observed. In what follows, we will also make use of

$$T_{(1)} = \min_{i \in \mathbb{N}} T_i,$$

the time at which a unit first adopts treatment, as well I_1 , the (random) index corresponding to the first adopter. Note that Assumption 2.2 below will impose that ties occur with probability zero, so I_1 may be defined to be the unique index satisfying $T_{I_1} = T_{(1)}$.

The null hypothesis of interest is the “sharp” null hypothesis that specifies that the treatment has no effect for all units and all time periods. Using the notation introduced above, we may express this null hypothesis formally as

$$H_0 : Y_{i,t}(0) = Y_{i,t}(1) \text{ for all } i \in \mathbb{N} \text{ and } t \in \mathbb{T}. \quad (3)$$

As mentioned previously, this type of null hypothesis has appeared in the literature on synthetic controls (Abadie et al., 2010; Firpo and Possebom, 2018).

Our analysis will require the following assumption:

Assumption 2.1. (*Unconfoundedness*) $(Y^{(n)}(0), Y^{(n)}(1)) \perp\!\!\!\perp T^{(n)} | X^{(n)}$.

A central object in our analysis will be the conditional distribution

$$I_1 | T_{(1)}, X^{(n)}. \quad (4)$$

This distribution will be governed by the following additional assumption:

Assumption 2.2. (*Proportional Hazards Model*) $(T_i, X_i), i = 1, \dots, n$ are i.i.d. and the distribution of $T_i | X_i$ has density w.r.t. Lebesgue measure such that

$$\lim_{\delta \downarrow 0} \frac{1}{\delta} P \{t \leq T_i \leq t + \delta | T_i \geq t, X_i\} = \lambda(t) \exp(X_i' \beta),$$

where $\beta \in \mathbb{B} \subseteq \mathbf{R}^d$.

Finally, in order to facilitate estimation of β , we further impose the following additional conditions:

Assumption 2.3. (*Regularity*) (i) \mathbb{B} is open, convex and bounded; (ii) $0 < \int_0^{t_{\max}} \lambda(t) dt < \infty$; (iii) $\text{Var}[X_i]$ is positive definite; and (iv) $\text{supp}(X_i)$ is bounded.

Remark 2.1. As mentioned previously, an implication of Assumption 2.2 for our purposes is that the probability of multiple units adopting treatment at the same time is zero. Since in most empirical applications the time of treatment adoption is only measured discretely, we view our methodology as being suited to settings in which treatment adoption times are measured with sufficient granularity so that this probability is small. In our empirical application in Section 5, for example, T_i denotes the time at which a state adopts tobacco legislation and our data permit measurement up to the month or even the day of adoption. ■

3 Main Result

In order to motivate our proposed testing procedure, it is useful first to describe an infeasible test of (3) that assumes β is known and is level α in finite samples. To this end, denote by

$$S_n = S_n \left(I_1, Y^{(n)}, (T_{(1)} \vee t_{\max}), X^{(n)} \right) \quad (5)$$

a test statistic such that large values of S_n provide evidence against (3). While (5) imposes some restrictions on the form of the test statistic, it accommodates many test statistics used in the literature, including choices in Abadie et al. (2010) and Firpo and Possebom (2018). See Remark

3.1 below for further discussion. In order to describe a suitable critical value with which to compare S_n , we begin with a straightforward implication of the null hypothesis (3) and Assumption 2.1.

Lemma 3.1. *Under Assumption 2.1,*

$$Y^{(n)} \perp\!\!\!\perp I_1 | T_{(1)}, X^{(n)}, \quad (6)$$

whenever the null hypothesis (3) holds.

We also note the following characterization of (4), the conditional distribution of the identity of the first adopter, I_1 , under Assumption 2.2:

Lemma 3.2. *Under Assumption 2.2, $P\{I_1 = i | T_{(1)}, X^{(n)}\} = \omega_{i,n}(\beta)$, where*

$$\omega_{i,n}(\beta) = \frac{\exp(X'_i \beta)}{\sum_{1 \leq k \leq n} \exp(X'_k \beta)}. \quad (7)$$

Importantly, the right-hand side of (7) does not depend on $T_{(1)}$. It is therefore possible to use these two lemmas to calculate the distribution of S_n conditional on $Y^{(n)}, T_{(1)}, X^{(n)}$ exactly using only the observed data whenever the null hypothesis (3) holds. Indeed, whenever (3) holds,

$$P \left\{ S_n \leq s | Y^{(n)}, T_{(1)}, X^{(n)} \right\} = \sum_{1 \leq i \leq n} \omega_{i,n}(\beta) I \left\{ S_n \left(i, Y^{(n)}, (T_{(1)} \vee t_{\max}), X^{(n)} \right) \leq s \right\}, \quad (8)$$

where $\omega_{i,n}(\beta)$ is given by (7). We again emphasize that the right-hand side of (8) may be computed using only the observed data. A suitable critical value with which to compare S_n is therefore given by

$$\hat{c}_n(1 - \alpha, \beta) = \inf \left\{ s \in \mathbf{R} : \sum_{1 \leq i \leq n} \omega_{i,n}(\beta) I \left\{ S_n \left(i, Y^{(n)}, (T_{(1)} \vee t_{\max}), X^{(n)} \right) \leq s \right\} \geq 1 - \alpha \right\}. \quad (9)$$

By construction, the test of (3) that rejects the null hypothesis if and only if S_n exceeds $\hat{c}_n(1 - \alpha, \beta)$ is level α in finite samples. Our feasible test is given by replacing β with

$$\hat{\beta}_n = \arg \max_{\beta \in \mathbb{B}} \prod_{1 \leq i \leq n} \left(\frac{\exp(X'_i \beta)}{\sum_{j \in \mathcal{R}_i} \exp(X'_j \beta)} \right)^{1 - \delta_i}, \quad (10)$$

where $\delta_i = I\{T_i > t_{\max}\}$ is an indicator for censorship of T_i and $\mathcal{R}_i = \{j \in \mathbb{N} : T_j \geq T_i\}$ is, in the language of the survival analysis literature, the “risk set” for unit i . We note that the maximand on the right-hand side of (10) is the celebrated partial likelihood of Cox (1975). The following

theorem shows that, under our assumptions, the resulting test, i.e.,

$$\phi_n = I \left\{ S_n > \hat{c}_n(1 - \alpha, \hat{\beta}_n) \right\}, \quad (11)$$

has limiting rejection probability under the null hypothesis no greater than the nominal level.

Theorem 3.1. *If Assumptions 2.1–2.3 hold, then the test ϕ_n defined in (11) satisfies*

$$\limsup_{n \rightarrow \infty} E[\phi_n] \leq \alpha$$

whenever the null hypothesis (3) holds.

The proof of Theorem 3.1 involves showing that the difference in the rejection probabilities of the infeasible and feasible tests tends to zero as n tends to infinity. The desired result then follows immediately since it is known that the infeasible test has rejection probability no greater than the nominal level in finite samples. A key step in the argument is to use Lemma A.1 in [Romano and Shaikh \(2012\)](#) to show that the difference in these rejection probabilities may be linked to the sum of the differences in $\omega_{i,n}(\beta)$ and $\omega_{i,n}(\hat{\beta}_n)$. Even though this bound involves a growing number of terms, we show that it is possible to control it using a combination of the boundedness of the support of X_i and the consistency of $\hat{\beta}_n$.

Remark 3.1. We emphasize that our theory applies to any choice of test statistic S_n that can be written as in (5), though some choices of test statistics may be preferable in terms of power of the resulting test. [Abadie et al. \(2010\)](#) suggests, for example, a test statistic of the form

$$\frac{\sum_{t \in \mathbb{T}: t \geq T_{(1)} \vee t_{\max}} (Y_{I_1, t} - \hat{Y}_{I_1, t}^N)^2}{\sum_{t \in \mathbb{T}: t < T_{(1)} \vee t_{\max}} (Y_{I_1, t} - \hat{Y}_{I_1, t}^N)^2}, \quad (12)$$

where $\hat{Y}_{I_1, t}^N$ is a linear combination of $\{Y_{i, t} : i \in \mathbb{N} \setminus \{I_1\}\}$, where the weights are chosen so that $Y_{I_1, t} \approx \hat{Y}_{I_1, t}^N$ for $t \in \mathbb{T}$ with $t < T_{(1)} \vee t_{\max}$. Under the sharp null hypothesis, $\hat{Y}_{I_1, t}^N$ may therefore be viewed as an estimator of $Y_{I_1, t}(0)$ for $t \in \mathbb{T}$ with $t \geq T_{(1)} \vee t_{\max}$. The weights in the construction of $\hat{Y}_{I_1, t}^N$ represent the combination of the other units that are intended to be used as a control for I_1 , i.e. the “synthetic control.” Different choices for these weights have been suggested by a variety of authors in the literature. Section 2 of [Cattaneo et al. \(2019\)](#) provides a succinct summary of proposals by [Abadie et al. \(2010\)](#), [Hsiao et al. \(2012\)](#), [Doudchenko and Imbens \(2016\)](#), [Chernozhukov et al. \(2018\)](#), [Ferman and Pinto \(2019\)](#), and [Arkhangelsky et al. \(2019\)](#). For further discussion of and contributions to the synthetic control literature, see [Abadie and L’Hour \(2017\)](#), [Amjad et al. \(2018\)](#), [Athey et al. \(2018\)](#) and [Ben-Michael et al. \(2019\)](#). Of course, other choices for S_n are possible, including a simple difference-in-differences test statistic, which we employ in our simulations in Section 4. ■

Remark 3.2. A p -value corresponding to the test in (11) may be defined as

$$\inf\{\alpha \in (0, 1) : S_n > \hat{c}_n(1 - \alpha, \hat{\beta}_n)\} . \quad (13)$$

In order to facilitate computation and our subsequent discussion, it is useful to define

$$\hat{p}_n(\omega) = \sum_{1 \leq i \leq n} \omega_i I\{s_{n,i} \geq S_n\} , \quad (14)$$

where S_n is defined as in (5), $s_{n,i} = S_n(i, Y^{(n)}, (T_{(1)} \vee t_{\max}), X^{(n)})$, and ω is an element of the n -dimensional simplex. In this notation, the p -value of our test defined in (13) is simply given by (14) with $\omega = (\omega_{i,n}(\hat{\beta}_n) : i \in \mathbb{N})$. ■

Remark 3.3. As mentioned previously, if (4), the conditional distribution of the identity of the first adopter, I_1 , is uniform, then the test defined in (11) equals the test proposed by [Abadie et al. \(2010\)](#). [Firpo and Possebom \(2018\)](#) discuss the properties of this test further and explore its sensitivity to deviations from the assumption that it is uniform. For a certain choice of S_n , they first parameterize the weights ω_i in (14) as

$$\frac{\exp(v_i \phi)}{\sum_{1 \leq k \leq n} \exp(v_k \phi)} , \quad (15)$$

where $v_i \in \{0, 1\}$ is unobserved for all $i \in \mathbb{N}$ and $\phi \in \mathbb{R}$ and, for an adversarial choice of $v = (v_i : i \in \mathbb{N})$, find the smallest value of ϕ that results in a p -value that differs in a meaningful way from the same p -value when $\phi = 0$. Despite the apparent similarity between the weights in (15) with those in (7), we emphasize that [Firpo and Possebom \(2018\)](#) do not derive the weights in (15) from more primitive assumptions on the time at which each unit adopts treatment like we do here. In this way, our results provide an alternative viewpoint on their specific parameterization. ■

Remark 3.4. By considering randomized tests, it is possible to construct a test that has limiting rejection probability exactly equal to the nominal level under the null hypothesis. This may be desirable in terms of power, especially when n is small. In particular, the test defined in (11) should be augmented so that it additionally rejects with probability q when $S_n = \hat{c}_n(1 - \alpha, \hat{\beta}_n)$, where

$$q = (1 - \alpha) - \sum_{1 \leq i \leq n} \omega_{i,n}(\hat{\beta}_n) I\{s_{n,i} \leq \hat{c}_n(1 - \alpha, \hat{\beta}_n)\}$$

with $s_{n,i} = S_n(i, Y^{(n)}, (T_{(1)} \vee t_{\max}), X^{(n)})$. Similar modifications are often used in the context of randomization tests to achieve exactness; see, for example, [Lehmann and Romano \(2006, Section 15.2.1\)](#). ■

4 Simulations

In this section, we explore the finite-sample behavior of our proposed testing procedure with a small simulation study. We first consider in Section 4.1 a situation in which the model for the time at which units adopt treatment is correctly specified; we then consider in Section 4.2 a situation in which this model is incorrectly specified.

4.1 Correct Specification

For $i \in \mathbb{N}$ and $t \in \mathbb{T}$, we assume that $Y_{i,t}(1) = \tau + Y_{i,t}(0)$, where

$$Y_{i,t}(0) = \rho Y_{i,t-1}(0) + \delta \sqrt{t} + \gamma X_i + \epsilon_{i,t} \quad (16)$$

with $X_i, i \in \mathbb{N}$ i.i.d. with

$$P\{X_i = c\} = \begin{cases} 0.7 & \text{if } c = -1 \\ 0.2 & \text{if } c = 0 \\ 0.1 & \text{if } c = 1 \end{cases},$$

$\epsilon_{i,t}, i \in \mathbb{N}, t \in \mathbb{T}$ i.i.d. as $N(0, 1)$. Additionally, we set $Y_{i,0}(0) = 0$ for all $i \in \mathbb{N}$, $\rho = 0.8$, $\delta \in \{0, 0.5, 1, 2\}$ and $\gamma \in \{0, 0.5, 1, 2\}$. The observed outcome, $Y_{i,t}$, is given by (2), and treatment status is given by (1), where $T_i|X_i$ distributed as $\text{Exp}(\lambda_i)$ with $\lambda_i = \exp(X_i\beta)$ and $\beta = 1$. It is straightforward to verify that this distribution of T_i satisfies Assumption 2.2 with baseline hazard equal to the hazard function of an exponential distribution with parameter equal to one. We set $n = 100$ and $t_{\max} = 100$.

We specify that S_n in (5) is a difference-in-differences test statistic:

$$\begin{aligned} & \frac{1}{t_{\max} - \lceil T_{(1)} \rceil} \sum_{T_{(1)} < t \leq t_{\max}} \left(Y_{I_1, t} - \frac{1}{n-1} \sum_{i \in \mathbb{N}: i \neq I_1} Y_{i, t} \right) \\ & - \frac{1}{\lceil T_{(1)} \rceil} \sum_{1 \leq t < T_{(1)}} \left(Y_{I_1, t} - \frac{1}{n-1} \sum_{i \in \mathbb{N}: i \neq I_1} Y_{i, t} \right). \end{aligned} \quad (17)$$

As mentioned in Remark 3.1, another possible choice would have been a synthetic control test statistic in (12), but this simpler choice of test statistic facilitates computation as well as some analytical calculations we present below in Remark 4.1. We emphasize that in our empirical application in Section 5 we employ a synthetic control test statistic.

In our simulations below, we present rejection probabilities computed using 10,000 replications for different values of δ and γ under the null hypothesis (3), i.e., with $\tau = 0$. In addition to our

proposed feasible testing procedure described, we also consider the infeasible test which treats β as known as well as the test proposed by [Abadie et al. \(2010\)](#). In our discussion, we refer to this test as the uniform test since it corresponds to our testing procedure with $\omega_i = 1/n$ for all $i \in \mathbb{N}$. The nominal level of all tests is $\alpha = 0.05$.

The results of our simulations are presented in [Table 1](#) below. As expected, after accounting for simulation error, the infeasible test has rejection probability no greater than the nominal level for any values of δ and γ . In accordance with [Theorem 3.1](#), the feasible test has rejection probability that is close to that of the infeasible test and therefore does not exceed the nominal level by a meaningful amount for any values of δ and γ . When $\gamma = 0$, i.e., when X_i does not enter [\(16\)](#), the uniform test also has rejection probability no greater than the nominal level, but, for all other values of δ and γ , the test exhibits rejection probabilities that exceed the nominal level by a considerable amount. Indeed, for such values of δ and γ , the rejection probability always exceed 0.08 and in some cases exceeds twice the nominal level. In [Remark 4.1](#) below, we discuss the performance of the uniform test when $\gamma = 0$ further.

Remark 4.1. A modest amount of calculation shows that under the distributional assumptions described above the test statistic in [\(17\)](#) may be written as

$$\frac{n}{n-1}\gamma(\rho^+ - \rho^-)(\bar{X}_n - X_{I_1}) + \zeta \tag{18}$$

under the null hypothesis, where

$$\begin{aligned} \bar{X}_n &= \frac{1}{n} \sum_{1 \leq i \leq n} X_i, \\ \rho^+ &= \frac{1}{t_{\max} - \lceil T_{(1)} \rceil} \sum_{T_{(1)} < t \leq t_{\max}} \rho^t / (1 - \rho), \\ \rho^- &= \frac{1}{\lfloor T_{(1)} \rfloor} \sum_{1 \leq t < T_{(1)}} \rho^t / (1 - \rho), \end{aligned}$$

and ζ is distributed as a normal random variable with mean zero. We note, in particular, that the difference-in-differences statistic eliminates any dependence on δ , but the effect of γ persists. When $\gamma = 0$, the effect of I_1 and X_i are both eliminated from [\(18\)](#), and tests corresponding to any choice of weights ω_i , including the uniform test, will have rejection probability under the null hypothesis no greater than the nominal level. Of course, when $\gamma \neq 0$, this need not be the case, as shown by the simulation results presented above. ■

δ	γ	uniform	feasible	infeasible
0.00	0.00	5.00	4.95	4.80
0.50	0.00	5.09	5.15	5.18
1.00	0.00	4.98	4.78	4.96
2.00	0.00	5.10	5.34	5.02
0.00	0.50	8.42	5.21	5.12
0.50	0.50	8.10	5.14	4.96
1.00	0.50	10.04	4.68	4.74
2.00	0.50	8.62	4.91	4.92
0.00	1.00	8.67	4.78	4.78
0.50	1.00	10.59	5.42	4.53
1.00	1.00	9.69	4.64	4.53
2.00	1.00	9.05	5.15	4.81
0.00	2.00	8.70	5.03	4.74
0.50	2.00	8.82	4.92	4.67
1.00	2.00	9.76	5.67	5.05
2.00	2.00	8.46	5.26	4.98

Table 1: Null rejection probabilities for three tests: uniform, feasible, infeasible. Rejection probabilities are computed using 10,000 replications

4.2 Incorrect Specification

In this section, we explore the robustness of our proposed test to misspecification of Assumption 2.2, i.e., the model for the times at which units adopt treatment. To this end, we first generate observed data using the model described in Section 4.1 under the null hypothesis. We additionally compute the weights $\omega_n(\hat{\beta}_n)$ used in implementing the feasible test discussed in that section. In order to examine the performance of this test to a wide variety of different types of misspecification of Assumption 2.2 in a succinct way, we implement the test with weights $\tilde{\omega}$ that are misspecified in the following way:

$$\tilde{\omega} = (1 - \epsilon)\omega_n(\hat{\beta}_n) + \epsilon\nu ,$$

where $\epsilon \in [0, 1]$ now indexes the amount of misspecification and ν is given by either $\nu_i = 1/n$ for all $i \in \mathbb{N}$ or $\nu_{(i)} = \omega_{(n-i),n}(\hat{\beta}_n)$ with indices $(1), \dots, (n)$ corresponding to

$$\omega_{(1),n}(\hat{\beta}_n) \leq \dots \leq \omega_{(n),n}(\hat{\beta}_n) .$$

Below we refer to these two choices of ν as “uniform” and “worst”, respectively, where the latter terminology reflects the fact that such a choice of ν is the most adversarial possible. Following the discussion in Remark 4.1, we see that δ is irrelevant for the performance of these tests; we therefore

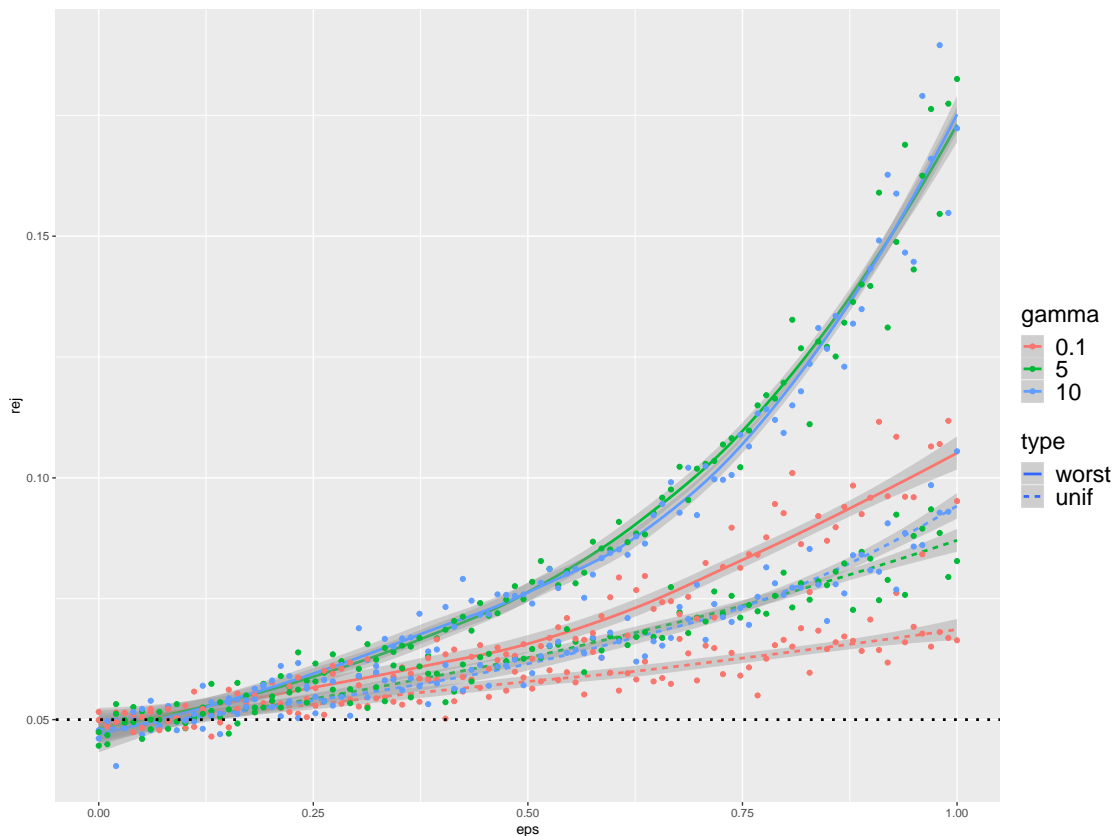


Figure 1: Null rejection probabilities (y-axis) for the feasible test with different amounts of misspecification indexed by ϵ (x-axis) and two different types of misspecification: uniform and worst. Rejection probabilities are computed using 10,000 replications.

focus on the effect of γ and ϵ . As before, the nominal level of all tests is $\alpha = 0.05$ and all rejection probabilities are computed using 10,000 replications.

These results are summarized in Figure 1. For values of ϵ as high as 0.25, we see that the rejection probabilities of our feasible test do not exceed the nominal level by an appreciable amount for any of the values of γ that we consider and either choice of ν . For larger values of ϵ , we see that the rejection probabilities may be significantly above the nominal level, but they remain remarkably robust. For instance, when $\gamma = 0.1$, they never greatly exceed 0.10 for any choice of ϵ and ν . The deterioration is worse for larger values of γ , which agrees with our discussion in Remark 4.1.

5 Empirical Application

In this section, we apply our proposed test to revisit the analysis in Abadie et al. (2010) of the effect of tobacco legislation on smoking prevalence. We recall that Abadie et al. (2010) was motivated by California’s adoption in 1989 of Proposition 99, a large-scale tobacco control program. A main

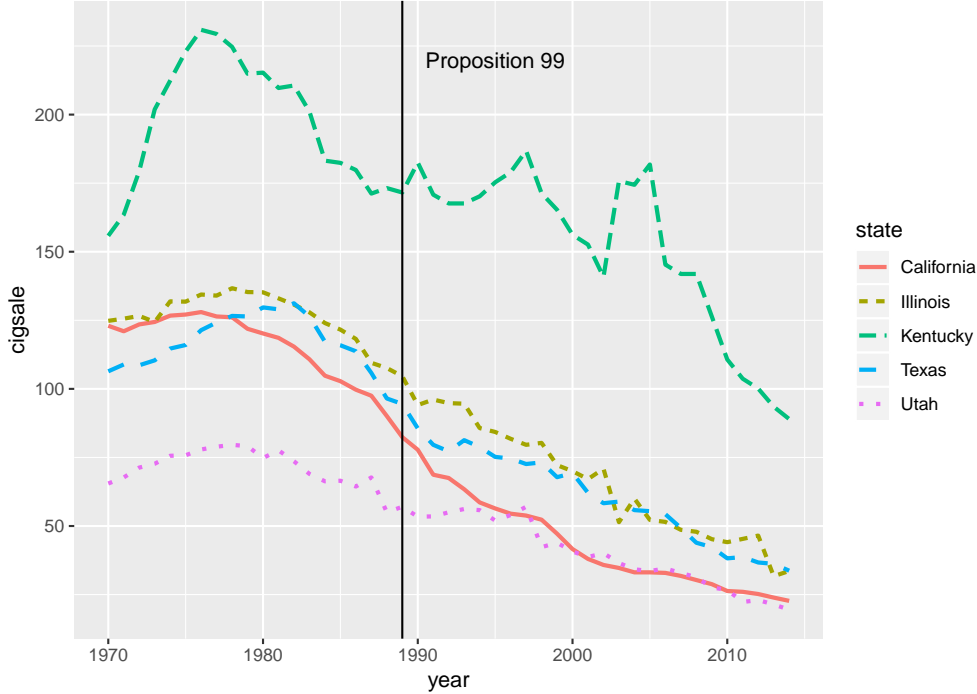


Figure 2: Annual cigarette packet sales per capita from 1970 to 2015 in select states. Proposition 99 was adopted by California in 1989.

component of this legislation was a steep increase in cigarette packet tax by 25 cents, representing an increase in taxes of 250%. While smoking prevalence declined after this legislation was adopted, it is important to emphasize that this decline happened in the backdrop of nationwide declining smoking prevalence dating at least as far back as the late 1970s. See Figure 2 for a graphical depiction of these trends for several different states, including California. In this context, [Abadie et al. \(2010\)](#) used his synthetic control methodology to test the null hypothesis that *none* of decline in smoking prevalence observed in California after 1989 can be attributed to the effects of the Proposition 99. Their analysis rejects this null hypothesis with a p -value of 0.026.

We now describe how we apply our methodology to this setting. To facilitate comparison with the results in [Abadie et al. \(2010\)](#), we restrict attention to the same $n = 39$ states in their analysis. These states are indexed by $i \in \mathbb{N}$. We index time by $t \in \mathbb{T} = \{“01/1971”, “02/1971”, \dots, “12/2014”\}$, where we have adopted the “month/year” format and identify 1 with “01/1971” and t_{\max} with “12/2014”. Denote by $Y_{i,t}$ the number of cigarette packets sold in state $i \in \mathbb{N}$ at time $t \in \mathbb{T}$. Finally, let T_i denote the time at which state $i \in \mathbb{N}$ adopts tobacco legislation. [Orzechowski and Walker \(2014\)](#) provide a comprehensive record of tobacco tax increases across states during this time period. In order to resolve any ambiguities, we define this to be the first time taxes on cigarette packets are increased by at least 50%. Every state except for Missouri adopt such tobacco legislation in our sample period. In Appendix B, we examine the robustness of our results

to different ways of defining T_i . In addition, while the data permit measurement of T_i up to the day of adoption, we simply record the month of adoption. In particular, $T_{(1)} = \text{“01/1989”}$. We emphasize, however, that no two states adopted tobacco legislation during the same month, so this is immaterial. Finally, treatment status is then computed as in (1).

Following [Abadie et al. \(2010\)](#), we employ a test statistic like that given by (12) in Remark 3.1. The weights in the construction of $\hat{Y}_{I_i,t}^N$ are computed as follows

$$w^* = \arg \min_w \sum_{t \in \mathbb{T}: t < T_{(1)}} \left(Y_{I_1,t} - \sum_{j \in \mathbb{N}: j \neq I_1} w_j Y_{j,t} \right)^2, \quad (19)$$

where the possible values of w are understood to be in the n -dimensional simplex with w_{I_1} restricted to be equal to zero.

In order to complete the description of our testing procedure, we assume that the times at which states adopt tobacco legislation is governed by a Cox proportional hazards model as described in Assumption 2.2, which requires, in particular, a specification of the covariates. We include the following covariates: per capita income (in logs), average price levels, fraction of population that are youth (with youth defined to be people with ages between 15 and 24), the unemployment level, and the fraction of state legislators that are Democrats. We note that these variables are not time-invariant and change on a yearly basis. While we did not allow for time-varying covariates in our formal analysis, it is straightforward to include them in a Cox proportional hazards model. See, for example, the discussion in Chapter 8 of [Cox and Oakes \(1984\)](#). In order to account for the systematic increase in some of these variables over time, we first de-trend these variables using a common linear trend across states. We emphasize that our test remains valid because the main requirement underlying its validity is simply the consistency of $\hat{\beta}_n$. Before proceeding, we note that in Appendix B we examine the robustness of our findings to more parsimonious specifications of the covariates.

For our baseline specification defined by the choices above, we compute a p -value of 0.044. We therefore reject the null hypothesis (3) at conventional significance levels, such as $\alpha = 0.05$. In order to gain some further insight into this result, it is worthwhile to examine the estimated conditional distribution of the identity of the state that first adopted tobacco legislation, i.e., the distribution of $I_1 | T_{(1)}, X^{(n)}$. This is presented in Table 2 for our baseline specification. We see that the model of treatment adoption implies that there are nine states that had higher probability of being the first to adopt treatment (conditional on the first adoption occurring in January 1989), namely, Nevada, Connecticut, Rhode Island, North Dakota, Maine, Illinois, Wisconsin, Texas, and Nebraska. In contrast, states such as Kentucky, Missouri or North Carolina, had much lower probability of being the first to adopt to adopt treatment (conditional on the first adoption occurring in January

State	Prob. of treatment adoption	State	State
Nevada	0.1037	Arkansas	0.0250
Connecticut	0.1018	Pennsylvania	0.0225
Rhode Island	0.0681	Louisiana	0.0201
North Dakota	0.0617	Ohio	0.0200
Maine	0.0605	Delaware	0.0186
Illinois	0.0580	Minnesota	0.0171
Wisconsin	0.0517	Tennessee	0.0163
Texas	0.0491	Montana	0.0151
Nebraska	0.0460	Idaho	0.0139
California	0.0440	Indiana	0.0134
New Hampshire	0.0360	Kansas	0.0124
Wyoming	0.0291	Georgia	0.0124
New Mexico	0.0279	Colorado	0.0111
		Virginia	0.0086
		Alabama	0.0081
		West Virginia	0.0066
		Oklahoma	0.0037
		South Carolina	0.0033
		South Dakota	0.0027
		Vermont	0.0026
		Utah	0.0023
		Iowa	0.0018
		North Carolina	0.0016
		Missouri	0.0014
		Kentucky	0.0013
		Mississippi	0.0006

Table 2: Estimated conditional distribution of the identity of the state that first adopted tobacco legislation, i.e., the distribution of $I_1|T_{(1)}, X^{(n)}$.

1989). Indeed, California was more than 34 times as likely to be the first adopter than Kentucky (conditional on the first adoption occurring in January 1989). These features reflect differences in both the characteristics of these states as well as disparities in the time at which different states adopted treatment. For instance, Mississippi only introduced such tobacco legislation as late as 2009.

We conclude the discussion of our empirical results by noting that had we implemented the test with $\omega_i = 1/n$ we would have computed, like [Abadie et al. \(2010\)](#), a p -value of 0.026. This phenomenon simply reflects the fact that for our specification above $S_n(i, Y^{(n)}, (T_{(1)} \vee t_{\max}), X^{(n)})$ is largest for i corresponding to California. As mentioned previously, in [Appendix B](#), we further examine the robustness of this finding to different ways of defining T_i as well as more parsimonious choices of covariates in the Cox proportional hazards model. We repeat the analysis for more than 30,000 different resulting specifications and find that in the vast majority of specifications, we still reject the null hypothesis at the $\alpha = 0.05$ significance level and in all specifications we reject at the $\alpha = 0.10$ significance level. In this sense, we believe our analysis largely confirms the findings in [Abadie et al. \(2010\)](#).

A Appendix

A.1 Proof of Lemma 3.1

Assumption 2.1 and (3) implies that

$$Y^{(n)} \perp\!\!\!\perp T^{(n)} | X^{(n)} .$$

The desired conclusion (6) now follows immediately upon noting that I_1 and $T_{(1)}$ are functions of $T^{(n)}$.

A.2 Proof of Lemma 3.2

Note that

$$\begin{aligned} P\{I_1 = i | T_{(1)} = t, X^{(n)}\} &= P\{T_i = t | T_{(1)} = t, X^{(n)}\} \\ &= \lim_{\delta \downarrow 0} P\{t \leq T_i \leq t + \delta | t \leq T_{(1)} \leq t + \delta, X^{(n)}\} \\ &= \lim_{\delta \downarrow 0} \frac{P\{t \leq T_i \leq t + \delta | X_i\} \prod_{1 \leq j \leq n: j \neq i} P\{T_j \geq t | X_j\}}{\sum_{1 \leq k \leq n} P\{t \leq T_k \leq t + \delta | X_k\} \prod_{1 \leq j \leq n: j \neq k} P\{T_j \geq t | X_j\}} \\ &= \lim_{\delta \downarrow 0} \frac{P\{t \leq T_i \leq t + \delta | T_i \geq t, X_i\}}{\sum_{1 \leq k \leq n} P\{t \leq T_k \leq t + \delta | T_k \geq t, X_k\}} \\ &= \lim_{\delta \downarrow 0} \frac{\frac{1}{\delta} P\{t \leq T_i \leq t + \delta | T_i \geq t, X_i\}}{\sum_{1 \leq k \leq n} \frac{1}{\delta} P\{t \leq T_k \leq t + \delta | T_k \geq t, X_k\}} \\ &= \frac{\lambda(t) \exp(X_i' \beta)}{\sum_{1 \leq k \leq n} \lambda(t) \exp(X_k' \beta)} , \end{aligned}$$

where the first equality follows by inspection, the second equality is understood to be by definition, the third through fifth equalities follow from Bayes' rule, and the sixth equality follows from Assumption 2.2. The desired conclusion now follows immediately.

A.3 Proof of Theorem 3.1

We first note that, under Assumptions 2.2–2.3, it follows from the discussion on page 55 of Kosorok (2008) that $\hat{\beta}_n \xrightarrow{P} \beta$ as $n \rightarrow \infty$. In what follows, make use of the notation $\hat{\Delta}_n = \hat{\beta}_n - \beta$ and $a \lesssim b$ for $a \leq cb$ for some constant c . Before proceeding, note that Assumption 2.3 implies that

$$0 < \inf_{x \in \text{supp}(X_i)} \exp(x' \beta) \leq \sup_{x \in \text{supp}(X_i)} \exp(x' \beta) < \infty . \quad (20)$$

Similarly, Assumption 2.3 implies that

$$\sup_{x \in \text{supp}(X_i)} \left| \exp(x' \hat{\Delta}_n) - 1 \right| = o_P(1) . \quad (21)$$

Now, note that

$$\begin{aligned}
& \sup_{s \in \mathbf{R}} \left| P \left\{ S_n \leq s \mid Y^{(n)}, (T_{(1)} \vee \tau), X^{(n)} \right\} - \frac{1}{n} \sum_{1 \leq i \leq n} \omega_{i,n}(\hat{\beta}_n) I \left\{ S_n \left(i, Y^{(n)}, (T_{(1)} \vee \tau), X^{(n)} \right) \leq s \right\} \right|, \\
& \leq \sum_{1 \leq i \leq n} \left| \omega_{i,n}(\beta) - \omega_{i,n}(\hat{\beta}_n) \right| \\
& \leq \sup_{1 \leq i \leq n} \left| \frac{\exp(X'_i \beta)}{\frac{1}{n} \sum_{1 \leq k \leq n} \exp(X'_k \beta)} - \frac{\exp(X'_i \hat{\beta}_n)}{\frac{1}{n} \sum_{1 \leq k \leq n} \exp(X'_k \hat{\beta}_n)} \right| \\
& = \sup_{1 \leq i \leq n} \left| \frac{\exp(X'_i \beta) \left(\frac{1}{n} \sum_{1 \leq k \leq n} \exp(X'_k \hat{\beta}_n) \right) - \exp(X'_i \hat{\beta}_n) \left(\frac{1}{n} \sum_{1 \leq k \leq n} \exp(X'_k \beta) \right)}{\left(\frac{1}{n} \sum_{1 \leq k \leq n} \exp(X'_k \beta) \right) \left(\frac{1}{n} \sum_{1 \leq k \leq n} \exp(X'_k \hat{\beta}_n) \right)} \right| \\
& = \sup_{1 \leq i \leq n} \left| \frac{\exp(X'_i \beta) \left(\left(\frac{1}{n} \sum_{1 \leq k \leq n} \exp(X'_k \hat{\Delta}_n) \exp(X'_k \beta) \right) - \exp(X'_i \hat{\Delta}_n) \left(\frac{1}{n} \sum_{1 \leq k \leq n} \exp(X'_k \beta) \right) \right)}{\left(\frac{1}{n} \sum_{1 \leq k \leq n} \exp(X'_k \beta) \right) \left(\frac{1}{n} \sum_{1 \leq k \leq n} \exp(X'_k \hat{\Delta}_n) \exp(X'_k \beta) \right)} \right| \\
& \lesssim \sup_{1 \leq i \leq n} \left| \frac{\left(\left(\frac{1}{n} \sum_{1 \leq k \leq n} \exp(X'_k \hat{\Delta}_n) \exp(X'_k \beta) \right) - \exp(X'_i \hat{\Delta}_n) \left(\frac{1}{n} \sum_{1 \leq k \leq n} \exp(X'_k \beta) \right) \right)}{\frac{1}{n} \sum_{1 \leq k \leq n} \exp(X'_k \hat{\Delta}_n)} \right| \\
& = \sup_{1 \leq i \leq n} \left| \frac{\left(\left(\frac{1}{n} \sum_{1 \leq k \leq n} \left(\exp(X'_k \hat{\Delta}_n) - 1 \right) \exp(X'_k \beta) \right) - \left(\exp(X'_i \hat{\Delta}_n) - 1 \right) \left(\frac{1}{n} \sum_{1 \leq k \leq n} \exp(X'_k \beta) \right) \right)}{\frac{1}{n} \sum_{1 \leq k \leq n} \left(\exp(X'_k \hat{\Delta}_n) - 1 \right) + 1} \right| \\
& = o_P(1).
\end{aligned}$$

where the first inequality exploits (8), the second inequality exploits (7), the first and second equalities follow by inspection, the third inequality exploits (20), the third equality follows by inspection, and the final equality exploits (20)–(21). The desired result now follows by applying Lemma A.1 in [Romano and Shaikh \(2012\)](#).

B Additional Specifications for Empirical Application

In this section, we examine the robustness of our empirical findings in Section 5. Specifically, we re-compute our p -value for different ways of defining T_i as well as exclusion of some of the six covariates we include in our Cox proportional hazards model.

As mentioned in Section 5, we define T_i to be the first time at which a state increased taxes on cigarette packets by at least 50%. While this eliminates any ambiguity, we identify nine states for which an alternative choice of T_i seems reasonable based subjectively on the magnitude or timing of the increase. These nine states are indicated in bold face in Table 3. For each of those states, we indicate in the column labeled ‘Specification B’ the choice of T_i corresponding to our specification in Section 5 and the alternative choice in the column labeled ‘Specification A’. For all other states, we simply repeat in these two columns the single choice of T_i that we consider. By considering all possible choices of T_i for these nine states, we obtain $2^9 = 512$ possible specifications of T_i . For each of these specifications, we additionally consider each of the $2^6 = 64$ possible subsets of the six covariates to include in the Cox proportional hazards model. We

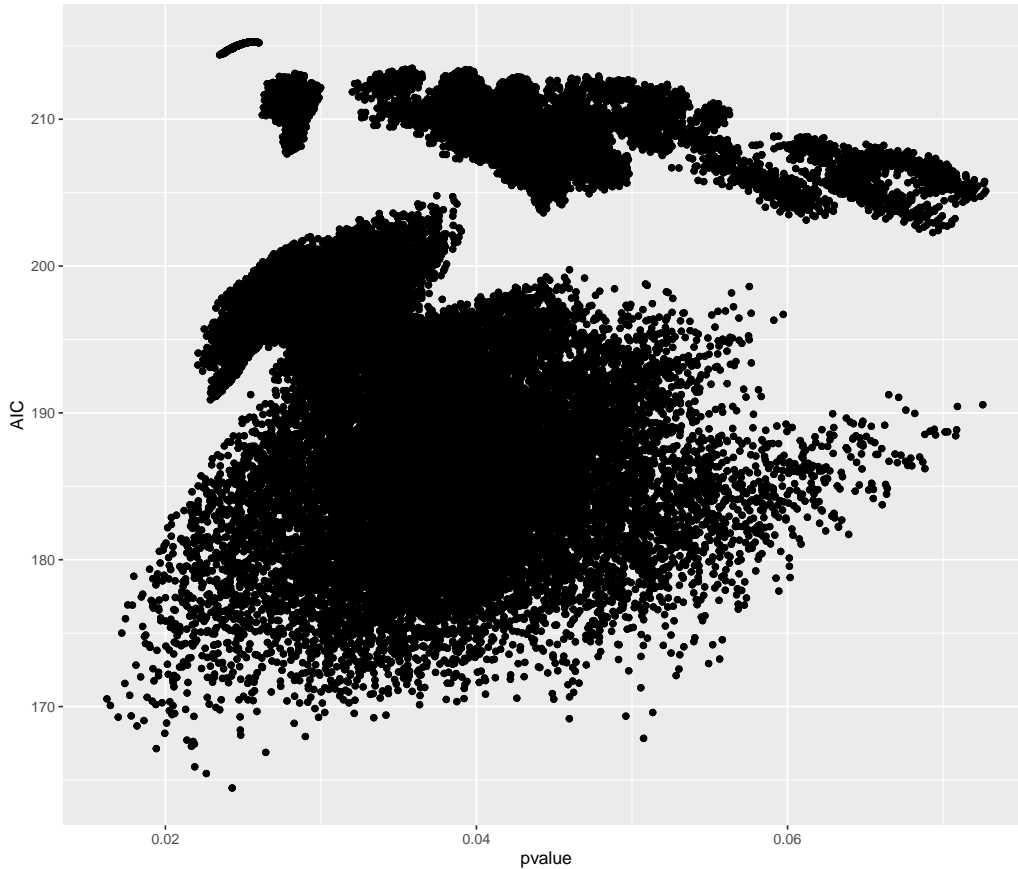


Figure 3: Scatter plot of (p -value, AIC) across all specifications.

therefore obtain in total $512 \times 64 = 32,768$ possible specifications. In order to facilitate our discussion below, we compute, in addition to the p -value for our test, the Akaike Information Criteria (AIC) for each Cox proportional hazards model.

In Figure 3, we plot for each of these specifications the (p -value, AIC). We see that in the vast majority of cases, we continue to reject the null hypothesis at the $\alpha = 0.05$ significance level. This conclusion is further strengthened if we restrict attention to specifications with better (i.e., lower) values of AIC. Finally, in all specifications, we reject the null hypothesis at the $\alpha = 0.10$ significance level. Indeed, the maximum p -value across all specifications is 0.072. In this sense, we find that our findings in Section 5 and, by extension, those of [Abadie et al. \(2010\)](#) are remarkably robust.

	State	Specification A	Specification B
1	Alabama	05/2004	05/2004
2	Arkansas	06/2003	02/1993
3	California	01/1989	01/1989
4	Colorado	01/2005	01/2005
5	Connecticut	04/1989	04/1989
6	Delaware	01/1991	08/2003
7	Georgia	07/2003	07/2003
8	Idaho	07/1994	06/2003
9	Illinois	07/1989	07/1989
10	Indiana	07/2002	07/2002
11	Iowa	04/2007	04/2007
12	Kansas	07/2002	07/2002
13	Kentucky	06/2005	06/2005
14	Louisiana	08/2002	08/2002
15	Maine	11/1997	07/1991
16	Minnesota	06/1991	08/2005
17	Mississippi	05/2009	05/2009
18	Missouri	12/2014	12/2014
19	Montana	05/2003	05/2003
20	Nebraska	10/2002	10/2002
21	Nevada	07/1989	07/1989
22	New Hampshire	02/1990	02/1990
23	New Mexico	07/2003	07/2003
24	North Carolina	09/2005	09/2005
25	North Dakota	05/1989	05/1989
26	Ohio	07/2002	07/2002
27	Oklahoma	01/2005	01/2005
28	Pennsylvania	08/1991	08/1991
29	Rhode Island	07/1997	07/1993
30	South Carolina	07/2010	07/2010
31	South Dakota	03/2003	07/1995
32	Tennessee	07/2002	07/2002
33	Texas	07/1990	07/1990
34	Utah	07/1991	07/1997
35	Vermont	07/1995	07/1995
36	Virginia	09/2004	09/2004
37	West Virginia	05/2003	05/2003
38	Wisconsin	05/1992	05/1992
39	Wyoming	07/2003	07/1989

Table 3: Different specifications for T_i .

References

- ABADIE, A. (2019). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*.
- ABADIE, A. and CATTANEO, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, **10** 465–503.
- ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California’s tobacco control program. *Journal of the American Statistical Association*, **105** 493–505.
- ABADIE, A. and GARDEAZABAL, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review*, **93** 113–132.
- ABADIE, A. and L’HOUR, J. (2017). A penalized synthetic control estimator for disaggregated data. *Work. Pap., Mass. Inst. Technol., Cambridge, MA*.
- AMJAD, M., SHAH, D. and SHEN, D. (2018). Robust synthetic control. *The Journal of Machine Learning Research*, **19** 802–852.
- ARKHANGELSKY, D., ATHEY, S., HIRSHBERG, D. A., IMBENS, G. W. and WAGER, S. (2019). Synthetic difference-in-differences. Tech. rep., National Bureau of Economic Research.
- ATHEY, S., BAYATI, M., DOUDCHENKO, N., IMBENS, G. and KHOSRAVI, K. (2018). Matrix completion methods for causal panel data models. Tech. rep., National Bureau of Economic Research.
- ATHEY, S. and IMBENS, G. W. (2018). Design-based analysis in difference-in-differences settings with staggered adoption. Tech. rep., National Bureau of Economic Research.
- BEN-MICHAEL, E., FELLER, A. and ROTHSTEIN, J. (2019). Synthetic controls and weighted event studies with staggered adoption. *arXiv preprint arXiv:1912.03290*.
- CARD, D. and KRUEGER, A. B. (1993). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. Tech. rep., National Bureau of Economic Research.
- CATTANEO, M. D., FENG, Y. and TITIUNIK, R. (2019). Prediction intervals for synthetic control methods. Tech. rep.
- CHERNOZHUKOV, V., WUTHRICH, K. and ZHU, Y. (2018). Practical and robust t -test based inference for synthetic control and related methods.
- CHERNOZHUKOV, V., WUTHRICH, K. and ZHU, Y. (2019). An exact and robust conformal inference method for counterfactual and synthetic controls. *arXiv preprint arXiv:1712.09089*.
- COX, D. and OAKES, D. (1984). *Analysis of Survival Data*, vol. 21. CRC Press.
- COX, D. R. (1975). Partial likelihood. *Biometrika*, **62** 269–276.

- DOUDCHENKO, N. and IMBENS, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Tech. rep., National Bureau of Economic Research.
- FERMAN, B. and PINTO, C. (2019). Synthetic controls with imperfect pre-treatment fit. *arXiv preprint arXiv:1911.08521*.
- FIRPO, S. and POSSEBOM, V. (2018). Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference*, **6**.
- HSIAO, C., STEVE CHING, H. and KI WAN, S. (2012). A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong Kong with mainland China. *Journal of Applied Econometrics*, **27** 705–740.
- KOSOROK, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer.
- LEHMANN, E. L. and ROMANO, J. P. (2006). *Testing Statistical Hypotheses*. Springer Science & Business Media.
- LI, K. T. (2019). Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association* 1–16.
- ORZECZOWSKI, W. and WALKER, R. (2014). The tax burden on tobacco. *Historical Compilation*, **39**.
- ROMANO, J. P. and SHAIKH, A. M. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *The Annals of Statistics*, **40** 2798–2822.
- ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, **74** 13–26.
- ROSENBAUM, P. R. (2002). Observational studies. In *Observational Studies*. Springer, 1–17.
- SNOW, J. (1855). *On the mode of communication of Cholera*. John Churchill.