Impact of Regression to the Mean on the Synthetic Control Method

Bias and Sensitivity Analysis

Nicholas A. Illenberger,[a] *Dylan S. Small,*[b] *and Pamela A. Shaw*[a]

Abstract: To make informed policy recommendations from observational panel data, researchers must consider the effects of confounding and temporal variability in outcome variables. Difference-in-difference methods allow for estimation of treatment effects under the parallel trends assumption. To justify this assumption, methods for matching based on covariates, outcome levels, and outcome trends—such as the synthetic control approach—have been proposed. While these tools can reduce bias and variability in some settings, we show that certain applications can introduce regression to the mean (RTM) bias into estimates of the treatment effect. Through simulations, we show RTM bias can lead to inflated type I error rates and bias toward the null in typical policy evaluation settings. We develop a novel correction for RTM bias that allows for valid inference and show how this correction can be used in a sensitivity analysis. We apply our proposed sensitivity analysis to reanalyze data concerning the effects of California's Proposition 99, a large-scale tobacco control program, on statewide smoking rates.

Keywords: Regression to the mean; Difference-in-difference; Matching; Sensitivity analysis; Synthetic control

(*Epidemiology* 2020;31: 815–822)

Panel studies are a type of longitudinal study that can be used to estimate the effect of an intervention on an outcome of interest by comparing outcome measurements collected pre- and posttreatment. Because treatment is not typically randomized, differences in outcomes cannot be attributed to intervention alone. If we consider the effect of a smoking cessation program on cigarette sales within a state, then state demographics, which may influence the likelihood of a cessation program being passed, can also effect sales trends. Additionally, temporal variation and outside events (such as natural disasters) can add noise to trends and affect estimates of the treatment effect. These features can occasionally create the illusion of a treatment effect where none exists. Given a set of treated and control units with outcomes measured pre- and postintervention, the difference-in-difference estimator is the difference in pretreatment outcomes between the two units subtracted from the difference in posttreatment outcomes.[5] Under the assumption that the treated and control groups would have parallel outcome trends in the absence of treatment, this estimator is unbiased for the average treatment effect on the treated (ATT). Ease of use and robustness to unmeasured confounding has made the difference-in-difference approach popular among epidemiologists.[10,19,21] However, because the estimator is not robust to deviations from the parallel trends assumption, control units must be selected with care.

To improve the selection of controls, Abadie[1] introduced the synthetic control approach. The method constructs a "synthetic control" unit using a weighted sum of donor controls. If donors are weighted such that they resemble the treated unit in the preintervention period, then the synthetic control should emulate how the treated unit would behave in the postintervention period in the absence of treatment. This is akin to matching, in that control units that are similar to the treated unit are weighted more heavily than those that are not. While matching may improve comparability between treated and control units, recent work by Daw and Hatfield[12] has shown that matching in difference-in-difference analyses can introduce regression to the mean (RTM) bias. Because of the similarities between matching and the synthetic control method, there is a need to better understand the effects of RTM when using synthetic control. In this article, we examine the effect of RTM on estimates of the ATT coming from the synthetic control and other matched difference-in-difference methods. Through simulations, we show that RTM can result in inflated type I error rates and, in some settings, decreased power. Compared

Submitted September 10, 2019; accepted July 29, 2020.

From the [a]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA; and [b]Department of Statistics, University of Pennsylvania, Philadelphia, PA.

The work of P.A.S. and D.S.S. was supported in part by R01 NIH grant R01-AI131771. The other authors have no conflicts to report.

Data and Code: All code for replicating the results of this article can be found in the eAppendix; http://links.lww.com/EDE/B721.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Nicholas Illenberger, Department of Biostatistics, University of Pennsylvania, 108/109 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104. E-mail: nillen@pennmedicine.upenn.edu.

ISSN: 1044-3983/20/3106-0815
DOI: 10.1097/EDE.0000000000001252

with other matching techniques, these effects are exaggerated in the synthetic control estimator. We also propose a novel sensitivity analysis, which can be used to check how robust inference may be to the effect of RTM bias. Sensitivity and quantitative bias analyses allow researchers to assess the potential effects of systematic error in an experiment[23] and are common in causal inference and missing data settings.[24] We apply our proposed sensitivity analysis to reanalyze data from Abadie et al,[2] estimating the effect of a large-scale tobacco control initiative on smoking sales in California. The data are from a publicly available dataset on state-level annual cigarette sales and are not subject to human subjects review.[3]

METHODS

Matched Difference-in-Difference

Consider a setting where observations are measured preintervention, $t = 0$, and postintervention, $t = 1$. Let $Y(t)$ represent the observed outcome at time t, A be an indicator of treatment status, and X be measure or unmeasured confounders. Define $Y^a(t)$ to be the potential outcome[25] which would be observed under treatment $A = a$ at time t. In this setting, $Y^1(0) = Y^0(0)$ because neither group receives treatment at time $t = 0$. Assume the linear model, $\mathbb{E}[Y^0(t)|X] = \beta X + \gamma t$, for the expected potential outcome under no treatment at time t. Because the distribution of X typically differs between the treatment groups, the potential mean under no treatment will differ as well. This model assumes that the effect of time, γ, does not depend on confounders and that the effect of confounders, β, does not depend on time. These are jointly known as the parallel trends assumption. If both are true and if the distribution of covariates within each group remains the same over time, then the expected difference between the potential untreated outcomes for the treated and control units in the pretreatment period is equivalent to that in the posttreatment period. Let $D(t)$ be the expected difference between the treatment and control groups at time t. If, alongside the parallel trends assumption, we also assume consistency ($Y^A(t) = Y(t)$) and random treatment conditional on X ($Y^0(t) \perp\!\!\!\perp A|X$), then we can show (see Appendix A):

$$\begin{aligned} D(t) &= \mathbb{E}[Y(t) \mid A=1] - \mathbb{E}[Y(t) \mid A=0] \\ &= \mathbb{E}[Y^1(t) - Y^0(t) \mid A=1] + \beta(\mathbb{E}[X \mid A=1] - \mathbb{E}[X \mid A=0]). \end{aligned}$$

For $t = 1$, the first term in this summation is the ATT, which we define as θ. Because $Y^1(0) = Y^0(0)$, it follows that θ is the difference between $D(1)$ and $D(0)$. A natural estimator of θ uses the empirical means within treatment groups at times $t = 0$ and $t = 1$. Specifically, $\hat{\theta} = \hat{D}(1) - \hat{D}(0)$, where

$$\hat{D}(t) = \frac{\sum_{i=1}^{n} Y_i(t)\mathbb{I}(A_i=1)}{\sum_{j=1}^{n} \mathbb{I}(A_j=1)} - \frac{\sum_{i=1}^{n} Y_i(t)\mathbb{I}(A_i=0)}{\sum_{j=1}^{n} \mathbb{I}(A_j=0)}.$$

In practice, it may be difficult to identify units such that β and γ are equivalent in the treated and control groups. Ryan et al[26] have shown that matching can decrease bias by improving the comparability of units. They consider the case where treated and control units are drawn from the same underlying population, but the probability of treatment depends on preintervention outcome levels or trend. In this setting, if preintervention outcomes are correlated with future observations, then matched difference-in-difference estimators of the ATT are less biased than their unmatched counterparts. However, as we will discuss, if treated and control units are pulled from populations with different outcome distributions, then matching can induce RTM bias into estimates of the ATT.

Regression to the Mean

Regression to the mean (RTM) is a statistical phenomenon in which extreme measurements of a random variable tend toward their expected value upon repeat measurement. Bias due to RTM is introduced when three conditions hold: (1) there is variability in outcome measures, (2) the population from which the treated unit is drawn differs from the control population, and (3) matching is done on pretreatment outcome levels. For example, suppose pretreatment outcome measurements are obtained from control and treatment populations with mean outcome levels μ_0 and μ_1, respectively. If $\mu_1 > \mu_0$, then the nearest-neighbor match for a treated unit is expected to be a control unit with outcome greater than μ_0. Because this control unit is expected to decrease upon repeat measurement in the posttreatment period (i.e., regress toward its mean), the differences in outcome levels between treated and matched control units are expected to be larger in the posttreatment period than in the pretreatment period even when there is no treatment effect. In this setting, matching results in a violation of the parallel trends assumption, leading to bias.

Matching Procedures

Synthetic Control Method

The method of synthetic controls is provided in detail elsewhere,[2] and so we provide only a brief overview. Suppose we collect data on a single treated unit and n_0 controls for a total of $n_0 + 1$ units. Let $i = 1$ index the treated unit and C denote the set of indices for the control units. Collect τ outcome measurements $Y_i = (Y_{i1}, \ldots Y_{i\tau})$ on each unit. Treatment is withheld until time τ_0, such that $j \in \{1, \ldots, \tau_0\}$ denote the pretreatment period and $j \in \{\tau_0 + 1, \ldots, \tau\}$ compose the posttreatment period. Select w_k, for $k \in C$ such that $Y_{1j} \approx \sum_{k \in c} w_k Y_{kj}$ for $j \in \{1, \ldots, \tau_0\}$ and $\sum_{k \in c} w_k = 1$. If weights are chosen so that these equalities approximately hold, then the weighted sum of the posttreatment control vectors can serve as a potential untreated outcome vector for the treated unit.

Nearest-Neighbor Matching

Let S be the set of pretreatment outcome measurements from the control group. If s are pretreatment measures for the treated unit, then we want to find the nearest-neighbor match for s over the set S. Given some distance metric, this match is the element of S which minimizes the distance from s.[7]

Different distance metrics can result in different matches. In this article, we consider two implementations of nearest-neighbor matching. The first method is based upon the distance between pretreatment outcome vectors as determined by the L_2 norm, while the second uses the L_1 distance between coefficients in an ordinary least squares regression of pretreatment outcome measurements on time (i.e., pretreatment trend).

SIMULATIONS

To examine the effect of RTM bias, we simulate a single treated unit alongside $n_0 = 40$ controls. For control units, outcome measurements are drawn from a multivariate normal distribution with mean μ_0, marginal variance $\sigma^2 = 1$, and first-order autoregressive (AR(1)) covariance structure with correlation $\rho^{|t_i - t_j|}$ between outcome measurements at t_i and t_j. The treatment unit is simulated similarly, with mean μ_1 rather than μ_0. For each simulated dataset, the treated unit is matched to controls using the synthetic control method, nearest neighbor based on the L_2 norm, and nearest neighbor based on pretreatment trend. For comparison, we provide an estimate of the treatment effect using the unmatched difference-in-difference. The situation we consider—that of one single unit and many controls—is typical for applications of synthetic control and unmatched difference-in-difference but is uncommon for 1:1 nearest-neighbor matching. Although it is not common under this setup, we implement the 1:1 nearest-neighbor matching approach to facilitate comparison with the other methods under study. Additionally, simulations by Daw et al[12] were based on 1:1 nearest-neighbor matching.

If we define $\overline{Y}_{0j} = n_0^{-1} \sum_{k \in c} Y_{kj}$ as the mean of the control units' outcomes at time j, then we calculate the unmatched difference-in-difference estimator as follows:

$$\hat{\theta} = \frac{1}{\tau - \tau_0} \sum_{j=\tau_0+1}^{\tau} \left\{ Y_{1j} - \overline{Y}_{0j} \right\} - \frac{1}{\tau_0} \sum_{j=1}^{\tau_0} \left\{ Y_{1j} - \overline{Y}_{0j} \right\}.$$

For the nearest-neighbor and synthetic control methods, the estimator for treatment effect simply replaces $\overline{Y}_{0j}$ with the value of the matched or synthetic control at time j. Because each unit is simulated multivariate normal with constant mean, the parallel trends assumption holds. Additionally, for μ_1 relatively close to μ_0, the synthetic control method should be able to find w_k such that $Y_{1j} \approx \sum_{k \in c} w_k Y_{kj}$ for $j \in \{1, \ldots, \tau_0\}$. If treated and control units are drawn from the same underlying distribution ($\mu_1 = \mu_0$), then all estimators would be unbiased for the ATT. However, when $\mu_1 \neq \mu_0$, both the matched difference-in-difference estimator and the synthetic control estimator will be biased from RTM.

Type I Error Rate

We first consider the effects of outcome level matching on type I error rates. For each method, we use permutation tests to test the null hypothesis of no treatment effect. For $i = 1, \ldots, n_0 + 1$, sequentially treat individual i as if they were the treated unit and estimate $\hat{\theta}_i$ for $i = 1, \ldots, n_0 + 1$. The P value for this test is given as $P = n^{-1} \sum \mathbb{I}\left(\left| \hat{\theta}_1 \right| \leq \left| \hat{\theta}_i \right| \right)$.

We consider settings with varying levels of μ_1 ($\mu_1 = 1, \ldots 5$), ρ ($\rho = 0.00, 0.25, 0.50, 0.75, 0.90$), and number of pretreatment observations (4 or 10). Each unit is simulated with four posttreatment observations. For each setting, 2000 simulations are performed. Results for simulations with four and 10 pretreatment observations are presented in Tables 1 and 2, respectively. In both settings, as μ_1 moves further from μ_0, the type I error rate for the synthetic control and outcome-level based nearest-neighbor approach increases. Additionally, as the correlation between repeat observations, ρ, increases, the type I error rate decreases. In all cases, the synthetic controls method leads to greater type I error rate inflation than the nearest-neighbor methods. Because the synthetic control uses information from all control units, there is less variance in the biased estimator. Matching on pretreatment linear trend does not appear to increase type I error rates in any scenario. This is consistent with findings from Daw et al.[12] By comparing Table 1 with Table 2, we see that the maximum type I error rate is greater when there are fewer pretreatment observations. As the number of preintervention observations grows, there will be less variability in the average preintervention outcome levels of each patient. Because of this, the maximum average among control units is expected to be closer to μ_0 when there are many preintervention observations than when there are fewer. This results in less bias due to RTM and lower type I error rates.

TABLE 1. Type I Error Rates for the Unmatched Difference-in-Difference, SC, NN_1, and NN_2

Type I Error Rate: Varying $\mu 1$					Type I Error Rate: Varying ρ				
μ_1	Unmatched	SC	NN_1	NN_2	ρ	Unmatched	SC	NN_1	NN_2
1.00	0.05	0.16	0.09	0.05	0.00	0.05	0.40	0.29	0.06
2.00	0.05	0.31	0.19	0.05	0.25	0.04	0.39	0.29	0.05
3.00	0.05	0.35	0.25	0.05	0.50	0.04	0.36	0.27	0.05
4.00	0.05	0.35	0.26	0.05	0.75	0.05	0.25	0.18	0.05
5.00	0.04	0.33	0.25	0.05	0.90	0.05	0.18	0.13	0.05

Data are simulated using 4 preintervention observations. In all simulation settings, $\mu_0 = 0$ and $\sigma^2 = 1$. For simulations varying μ_1, $\rho = 0.5$. For simulations varying ρ, $\mu_1 = 5$.
NN_1 indicates nearest neighbor using the L_2 norm; NN_2, nearest neighbor using linear trends; SC, synthetic control.

TABLE 2. Type I Error Rates for the Unmatched Difference-in-Difference, SC, NN_1, and NN_2

Type I Error Rate: Varying μ_1					Type I Error Rate: Varying ρ				
μ_1	Unmatched	SC	NN_1	NN_2	ρ	Unmatched	SC	NN_1	NN_2
1.00	0.05	0.16	0.07	0.06	0.00	0.05	0.26	0.08	0.05
2.00	0.04	0.25	0.12	0.04	0.25	0.06	0.22	0.08	0.05
3.00	0.05	0.24	0.15	0.04	0.50	0.05	0.16	0.07	0.05
4.00	0.04	0.25	0.15	0.05	0.75	0.05	0.11	0.07	0.05
5.00	0.05	0.26	0.17	0.04	0.90	0.05	0.07	0.06	0.03

Data are simulated using 10 preintervention observations. In all simulation settings, $\mu_0 = 0$ and $\sigma^2 = 1$. For simulations varying μ_1, $\rho = 0.5$. For simulations varying ρ, $\mu_1 = 1$.
NN_1 indicates nearest neighbor using the L_2 norm; NN_2, nearest neighbor using linear trends; SC, synthetic control.

TABLE 3. Type I Error Rate for the Different Estimators of the ATT

Type I Error Rate: Varying γ_1					Type I Error Rate: Varying βx				
γ_1	Unmatched	SC	NN_1	NN_2	β_x	Unmatched	SC	NN_1	NN_2
0.00	0.05	0.27	0.06	0.03	0.0	0.05	0.30	0.05	0.04
1.00	0.04	0.37	0.05	0.05	0.5	0.05	0.32	0.05	0.04
2.00	0.06	0.36	0.06	0.05	1.0	0.05	0.38	0.05	0.05
3.00	0.06	0.36	0.05	0.05	1.5	0.05	0.36	0.06	0.05
4.00	0.05	0.39	0.05	0.05	2.0	0.04	0.35	0.04	0.03

Data are simulated using 4 pretreatment observations and a covariate associated with the outcome level. Here, $\mu_0 = 0$, $\mu_1 = 2$, $\rho = 0.5$, and $\sigma^2 = 1$. When varying γ_1, $\beta_x = 1$. When varying β_1, $\gamma_1 = 1$.
NN_1 indicates nearest neighbor using the L_2 norm; NN_2, nearest neighbor using linear trends; SC, synthetic control.

Matching with Covariates

Because the motivating model for the synthetic control procedure includes covariates which are associated with the outcome of interest, we perform simulations to show the effects of regression to the mean in this setting. We simulate a covariate X from a multivariate normal distribution with mean γ_0 for control units and γ_1 for the treated unit. X is simulated with an AR(1) error covariance structure with variance, $\sigma^2_X = 0.25$, and correlation, $\rho_X = 0.4$. For each unit, Y_i is multivariate normal with mean $\mu_i + \beta_x X_i$, where $\mu_i = \mu_1$ for the treated unit and $\mu_i = \mu_0$ for control units. The synthetic control method now uses both preintervention levels of Y and X when constructing weights. We also consider unmatched and nearest-neighbor matching in our simulations. The first variation of nearest-neighbor matches on the preintervention levels of X, while the second matches on preintervention trend in X.

Table 3 contains type I error rates obtained while varying the value of γ_1 between 0 and 5 and β_x between 0 and 2. In these simulations, $\mu_0 = 0$, $\mu_1 = 5$, and $\rho = 0.5$. The type I error rate increases as the distributions of X in the treated and control groups move further apart and as the effect of X on the outcome level increases.

Bias Toward the Null

As evidenced by previous simulations, matching on pretreatment outcomes can lead to anticonservative bias. However, in some settings where there is a treatment effect, RTM can also result in bias toward the null. To illustrate this phenomenon, we perform 1000 simulations with $\mu_0 = 0$, $\mu_1 = 2$, and $\rho = 0.5$. We induce a treatment effect, θ. For each additional time point in the treatment period, the expected outcome for the treated unit increases by θ. For negative θ, the treatment effect and bias due to RTM are working in opposite directions resulting in bias toward the null and conservative rejection rates. Figure 1 provides rejection rates for the unmatched, nearest-neighbor, and synthetic control procedures when θ is between 0 and −1.5. As in the previous simulations,

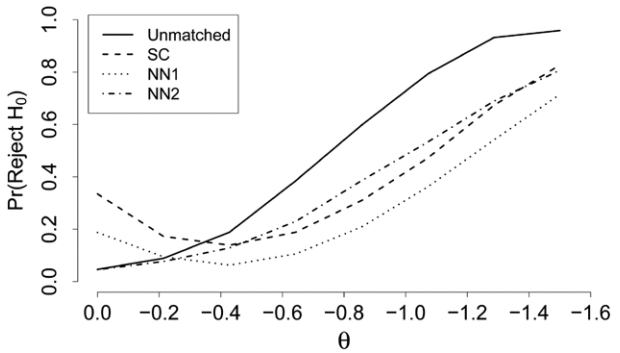

FIGURE 1. Empirical probability of rejecting the hypothesis that there is no treatment effect ($\theta = 0$) as a function of θ using the unmatched, synthetic control method (SC), nearest-neighbor matching on L_2 norm (NN_1), and nearest-neighbor matching on linear trend (NN_2).

we see that when there is no treatment effect, the synthetic control method exhibits inflated type I error rates, while the unmatched data have appropriate rejection rates.

As the treatment effect increases, the rejection rate of the unmatched estimator's power surpasses that of the synthetic control method. These results indicate that, depending on the direction of the treatment effect in relation to the direction of RTM, the synthetic control method can result in either conservative or anticonservative bias.

Correction and Sensitivity Analysis

Suppose $Y_1, \ldots, Y_T$ are jointly normal random variables. By properties of the multivariate normal distribution, for any given i and j, we have $\mathbb{E}[Y_i \mid Y_j] = \mu_i + \Sigma_{ij}\Sigma_{ij}^{-1}(Y_j - \mu_j)$. If we know the mean vectors and the covariance structure for each unit, then we can use this representation to account for RTM bias in matched difference-in-difference estimates of the ATT. To illustrate, suppose we have a single treated unit and a sample of control units. Using a 1:1 matching technique, the treated unit Y_1 is matched with a control unit, Y_m, and an estimate of the ATT, $\hat{\theta}_{obs}$, is obtained. This estimate can be conceptualized as the sum of the effect due to RTM bias and the effect due to treatment. Our correction technique subtracts the estimated effect of RTM, $\hat{\theta}_{rtm}$, from the observed effect to obtain a bias-adjusted estimate of the ATT, $\hat{\theta}_{adj} = \hat{\theta}_{obs} - \hat{\theta}_{rtm}$.

If we assume that observations are normally distributed and follow a Markov process (as would be true for an AR(1) error structure), then we can predict postintervention observations using only the most recent preintervention observation. Define $\hat{Y}_{ij} = \mu_{ij} + \sum_{j\tau_0}\sum\tau_0\tau_0^{-1}\left(Y_{i\tau_0} - \mu_{i\tau_0}\right)$ for $j > \tau_0$ and $i \in \{1, m\}$, where m is the index of the matched control unit. Here, τ_0 is the final pretreatment observation time and $\mu_{ij} = \mathbb{E}[Y_i^0(j)]$ is the expected potential outcome level under no treatment for unit i at posttreatment time j. $\hat{Y}_{ij}$ is the expected observation for unit i at posttreatment time j conditional on the pretreatment observations assuming no treatment effect. Estimating the ATT using these expected values in place of observed posttreatment values for the treated and matched control units provides $\hat{\theta}_{rtm}$, which can be used to find $\hat{\theta}_{adj}$. To generalize this adjustment for use with synthetic controls, first obtain the synthetic control weights w_k for $k \in C$. Using these weights, construct a synthetic outcome vector Y_S, where $Y_{Sj} = \sum_{k \in c} w_k Y_{kj}$, and corresponding estimate of the ATT, $\hat{\theta}_{obs}$. To obtain the expected estimate of ATT under RTM, construct an augmented synthetic control by replacing post treatment control measurements with the previously defined values $\hat{Y}_{ij}$s and incorporating the fit weights, $W k$. Note that $\hat{Y}_{ij}$ replacing posttreatment control measurements must be found for $j \in \{1, \ldots, n\}$. Call this augmented control $\hat{Y}_S$ and calculate $\hat{\theta}_{rtm}$ by subtracting the mean difference in observed pretreatment outcomes of the treated unit and the synthetic control unit from the mean difference in expected posttreatment outcomes. The covariance matrices used to construct the $\hat{Y}_{ij}$ in this correction procedure are those of the unmatched and unweighted observations. The goal of this procedure is to use preintervention observations to estimate the expected outcomes of each unit in the postintervention period under the assumption of no treatment effect. Applying the different estimators of the ATT to those projections allows us to determine what portion of the original estimate of the ATT is explainable by RTM bias.

As a proof of concept, we perform 2000 simulations with outcomes drawn from a multivariate normal distribution with AR(1) error structure. Here, $\mu_0 = 0$, $\mu_1 = 1$, $\sigma^2 = 1$, $\rho = 0.5$, and there is no treatment effect. For each simulation, we test the null hypothesis of no treatment effect using the permutation test described earlier, replacing $\hat{\theta}_i$ with $\hat{\theta}_{i,adj}$. To test if this adjustment is robust to deviations from this assumption, we also determine type I error rates when outcomes are drawn from a multivariate t distribution. Simulation results are given in Table 4. When errors are normally distributed, the adjusted synthetic control estimate of the ATT attains nominal type I error rates. Error rates are inflated for t distributed outcomes, particularly for highly correlated outcomes. However, observed error rates are lower than those obtained in Table 1 using the unadjusted synthetic control approach with normally distributed errors.

In practice, estimating $\hat{\theta}_{adj}$ is not possible without assuming the values of μ_{1j}, μ_{0j}, ρ, or σ. We propose treating these as sensitivity parameters. By positing a range of values for these parameters and calculating $\hat{\theta}_{adj}$ under each set, we can quantify how much an estimate of the ATT is affected by RTM.

Reanalysis of Smoking Cessation Data

To further understanding of our proposed sensitivity analysis, we reanalyze data from Abadie et al[2] concerning the effect of California's Proposition 99 on smoking cessation. The act added a 25 cents per pack tax on the sale of cigarettes and earmarked tax revenue for use in health care programs and antitobacco advertisements. The original analysis concluded that the initiative decreased cigarette consumption in California by approximately 20 packs per capita annually. Because this analysis was based upon the synthetic control method, we aim to determine if these findings are robust to RTM bias using our proposed sensitivity analysis.

TABLE 4. Type I Error Rates for the Adjusted Difference-in-Difference Estimator When Normality Assumption Is Not Satisfied

Degrees of Freedom	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$
∞ (Normal)	0.05	0.05	0.05
50	0.05	0.05	0.06
10	0.05	0.06	0.08
3	0.05	0.08	0.12

Errors come from a t distribution with degrees of freedom described.

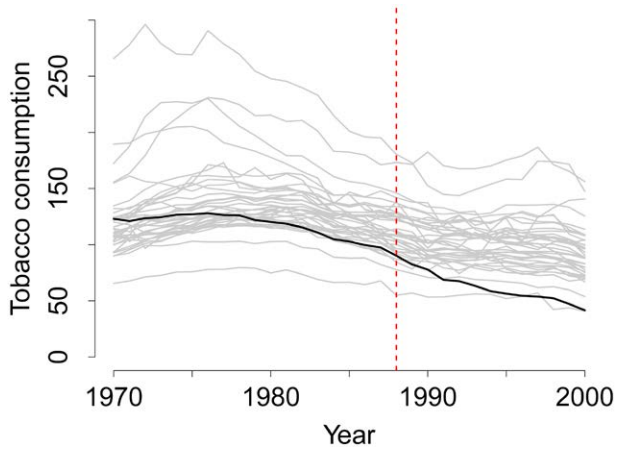

FIGURE 2. Tobacco consumption (per capita cigarette consumption) in a subset of states between 1970 and 2000. California highlighted in black, treatment initiation indicated by dashed vertical line.

Following the original study, this analysis was based upon cigarette consumption rates in California and 38 control states. Figure 2 provides a plot of cigarette consumption rates between 1970 and 2000 for the included states. To construct the synthetic control unit, we use logged per capita GDP, the average retail price of cigarettes within each state, beer consumption per capita, the percentage of the population aged 15 to 24, and cigarette sales in the pretreatment years 1975, 1980, and 1988. Without adjustment, we estimate that Proposition 99 reduced consumption by 18.9 packs per capita annually between the years 1989 and 2000.

To perform the sensitivity analysis we must (1) propose a set of reasonable models for the distribution of outcomes under no treatment effect for both the treated and control groups, (2) calculate the expected value of outcomes in the posttreatment period given our assumed outcome models and pretreatment observations, and (3) calculate adjusted estimates of the ATT using these values. Proposed distributions of outcomes under no treatment effect can be pulled from domain knowledge or from statistical modeling. For illustration, we employ generalized estimating equations (GEE) with an AR(1) working correlation matrix to regress per capita cigarette sales on the covariates used to construct the synthetic control. Because this model is meant to estimate the ATT under the assumption of no treatment effect, we fit the model using all states. If there is no treatment effect, then even California's outcomes are representative of outcomes under no treatment. The estimated residual standard deviation is 11.6 and the sample correlation is 0.72. Let $g_i(t)$ denote the predicted value from the fit model for unit i at time t. Define $\hat{Y}_{ij} = \hat{\rho}^{j-1988}\left(Y_{i,1988} - g_i(1988)\right) + g_i(j)$ for $j > 1988$ and $i \in \{1, \ldots, 39\}$. If W_k for $k \in C$ are the weights for our synthetic control, then $\hat{\theta}_{rtm} = \frac{1}{12}\sum_{j=1989}^{2000}\left(\hat{Y}_{1j} - \hat{Y}_{sj}\right) - \frac{1}{19}\sum_{j=1970}^{1988}\left(Y_{1j} - Y_{sj}\right)$, where $Y_{sj} = \sum_{k \in c} w_k Y_{kj}$ and $\hat{Y}_{sj} = \sum_{k \in c} w_k \hat{Y}_{kj}$. This expected estimate under no treatment effect is then subtracted from the observed estimate to obtain the adjusted estimate of the ATT. Using this procedure, the adjusted estimate of the ATT is 12.1 with P value 0.10. We also consider the set of outcome models indexed by Δ: $g_i(j; \Delta) = g_i(j) + \Delta\mathbb{I}(i = 1)$. This is the same null outcome model considered above except we shift the mean of the treated unit by Δ. For $\Delta = -1$ and 1, the adjusted estimates of the ATT is 11.3 ($P = 0.1$) and 12.9 ($P = 0.05$). Likewise, if we look at $\Delta = -5$ and 5, then the estimates of the ATT become 8.14 ($P = 0.3$) and 16.0 ($P = 0.05$). Because the estimated treatment effects and associated P values vary over relatively similar null outcome models, there is evidence that RTM may play a large role in our estimate of the ATT and suggest that further research be done to determine the effect of the tobacco tax.

Applying multiple models for the null outcome distribution when performing this analysis can help better characterize the sensitivity of results. If, instead of an AR(1) structure, we had chosen an unstructured error model then the adjustment would proceed as described except we would change our calculation of $\hat{Y}$. In this case, calculate $\hat{Y}_{ij} = \sum_{j,pre}\sum_{\text{pre,pre}}^{-1}\left(Y_{i,\text{pre}} - g_{i,\text{pre}}\right) + g_i(j)$, where $Y_{i,pre}$ is the vector of preintervention observations, $g_{i,pre}$ is the vector of predicted preintervention outcomes obtained from our fit model, and $\Sigma_{\text{pre,pre}}$ is the estimated covariance matrix of the preintervention outcomes. Because the unstructured model does not have the Markov property, we must condition on all pretreatment observations when calculating our expected values of Y under the null distribution.

DISCUSSION

In this article, we have illustrated the effects of RTM bias on matched difference-in-difference estimators. This builds upon work done by Daw and Hatfield[12] showing the bias induced by 1:1 nearest-neighbor matching. Here, we have shown how the synthetic control approach can also introduce bias and have provided simulations showing the effect of this bias on type I error rates and power. Our results suggest that synthetic control approaches are more prone to bias than nearest-neighbor matching. Added "confidence" in the model, gained from utilizing information from all of the control units, can increase the type I error rate by a factor of two over the nearest-neighbor approach. We also developed an approach to determine the sensitivity of matched estimators of the ATT to RTM bias. Using our approach, we showed that previous results concerning the effect of Proposition 99 on cigarette consumption in California[2] may be overstated. The results we obtained differ from those of Ryan et al,[26] which showed that matching can be beneficial in settings where the probability of treatment is associated with preintervention outcomes, but both treated and control units are drawn from the same underlying distribution. In our simulations and in those of Daw and Hatfield,[12] treated and control units are drawn from populations with different outcome distributions. Because it

is difficult to know which of these two settings hold, we cannot know whether matching will be protective against bias or if it will induce bias. There is a need for methodology which performs well in either settings. In recent work, Doudchenko and Imbens[13] and Arkhangelsky et al,[6] propose novel synthetic control approaches that weight control units based on how similar their preintervention outcome trends are to that of the treated unit. Because our results show that matching on trend did not induce bias, these implementations of synthetic control could be helpful. Further research is needed to determine how these and other recent approaches (such as that of Ben-Michael et al[9]) are affected my RTM bias.

Because the goal of this article is to examine the effects of RTM on the simpler and more popular variants of synthetic controls and matched difference-in-difference, we do not consider these approaches. Additionally, we do not consider how k:1 matching techniques are affected by RTM bias. Stuart[28] suggests that whether k:1 matching is superior to 1:1 matching depends on the setting. Thus, comparisons with k:1 matching may be more nuanced and deserving of further research.

In the future, it may be worthwhile to look for ways to correct for RTM bias when we cannot assume normality of errors. For t distributed errors, we noticed that type I error rates were slightly greater than desired α levels. While the adjustment still performed better than the unadjusted synthetic control estimator, we believe the method could be improved upon. As a whole, we believe that when researchers apply matched difference-in-difference estimators, they should also provide evidence that their results are robust to RTM bias, either by using our adjusted difference-in-difference estimator or by providing a sensitivity analysis.

REFERENCES

1. Abadie A. Semiparametric difference-in-differences estimators. *Rev Econ Stud*. 2005;72:1–19.
2. Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *J Am Stat Assoc*. 2010;105:493–505.
3. Abadie A, Diamond A, Hainmueller J. Synth: stata module to implement tsynthetic control methods for comparative case studies. https://EconPapers.repec.org/RePEc:boc:bocode:s457334. 2014
4. Angrist JD, Pischke J-S. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press, 2008.
5. Arkhangelsky D, Athey S, Hirshberg DA, Imbens GW, Wager S. Synthetic difference in differences. Technical report. National Bureau of Economic Research Cambridge, MA; 2019.
6. Arya S, Mount DM, Netanyahu N, Silverman R, Wu AY. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. In *Proc. 5th ACM-SIAM Sympos. Discrete Algorithms*. 1994;573–582.
7. Ben-Michael E, Feller A, Rothstein J (2019) The augmented synthetic control method. University of California Berkeley, Mimeo. https://arxiv.org/pdf/1811.04170.pdf. Accessed Mar 2020.
8. Branas CC, Cheney RA, MacDonald JM, Tam VW, Jackson TD, Ten Have TR. A difference-in-differences analysis of health, safety, and greening vacant urban space. *Am J Epidemiol*. 2011;174:1296–1306.
9. Daw JR, Hatfield LA. Matching and regression to the mean in difference-in-differences analysis. *Health Serv Res*. 2018;53:4138–4156.
10. Doudchenko N, Imbens GW. Balancing, regression, difference-in-differences and synthetic control methods: a synthesis. Technical report. National Bureau of Economic Research; Cambridge, MA 2016.
11. Hamad R, Batra A, Karasek D, et al. The impact of the revised WIC food package on maternal nutrition during pregnancy and postpartum. *Am J Epidemiol*. 2019;188:1493–1502.
12. Kagawa RMC, Castillo-Carniglia A, Vernick JS, et al. Repeal of comprehensive background check policies and firearm homicide and suicide. *Epidemiology*. 2018;29:494–502.
13. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol*. 2014;43:1969–1985.
14. Robins JB, Rotnitzky A, Scharfstein DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. Springer; 2000, New York:1–94.
15. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *J Am Stat Assoc*. 2005;100:322–331.
16. Ryan AM, Burgess JF Jr, Dimick JB. Why we should not be indifferent to specification choices for difference-in-differences. *Health Serv Res*. 2015;50:1211–1235.
17. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25:1.

APPENDIX

We wish to prove:

$$D(t) = \mathbb{E}[Y^1(t) - Y^0(t) \mid A = 1] - \beta(\mathbb{E}[X \mid A = 1] - \mathbb{E}[Y \mid A = 0])$$

Consider the following:

$$\begin{aligned} D(t) &= \mathbb{E}[Y(t) \mid A{=}1] - \mathbb{E}[Y(t) \mid A{=}0] \\ &= \mathbb{E}[Y^1(t) \mid A{=}1] - \mathbb{E}[Y^0(t) \mid A{=}0] \\ &= \mathbb{E}[Y^1(t) - Y^0(t) \mid A{=}1] + \mathbb{E}[Y^0(t) \mid A{=}1] - \mathbb{E}[Y^0(t) \mid A{=}0] \end{aligned}$$

Here, the second line follows from the consistency assumption. Next note, for $A = 0$ or 1, the expected value of the potential distribution can be rewritten as follows:

$$\begin{aligned} \mathbb{E}[Y^0(t) \mid A] &= \mathbb{E}\{\mathbb{E}[Y^0(t) \mid X, A] \mid A\} \\ &= \mathbb{E}\{\mathbb{E}[Y^0(t) \mid X] \mid A\} \\ &= \mathbb{E}[\beta X + \gamma t \mid A] \\ &= \beta \mathbb{E}[X \mid A] + \gamma t \end{aligned}$$

The second line is true because we assume $Y^0(t) \perp\!\!\!\perp A|X$. Plugging this into the expression for $D(t)$ we can

$$D(t) = \mathbb{E}[Y^1(t) - Y^0(t) \mid A{=}1] + \beta(\mathbb{E}[X \mid A{=}1] - \mathbb{E}[X \mid A{=}0])$$