Matching on Noise: Finite Sample Bias in the Synthetic Control Estimator

Joseph Cummins
University of California, Riverside

Douglas L. Miller
Cornell University & NBER

Brock Smith
Montana State University

David Simon
University of Connecticut & NBER

July 11, 2023

Abstract

We investigate the properties of a systematic bias that arises in the synthetic control estimator in panel data settings with finite pre-treatment periods, offering intuition and guidance to practitioners. The bias comes from matching to idiosyncratic error terms (noise) in the treated unit and the donor units' pre-treatment outcome values. This in turn leads to a biased counterfactual for the post-treatment periods. We use Monte Carlo simulations to evaluate the determinants of the bias in terms of error term variance, sample characteristics and DGP complexity, providing guidance as to which situations are likely to yield more bias. We also offer a procedure to reduce the bias using a direct computational bias-correction procedure based on re-sampling from a pilot model that can reduce the bias in empirically feasible implementations. As a final potential solution, we compare the performance of our corrections to that of an Interactive Fixed Effects model. An empirical application focused on trade liberalization indicates that the magnitude of the bias may be economically meaningful in a real world setting.

Keywords: Synthetic Control, Over-fitting

JEL Codes: C23; C52

1 Introduction

The synthetic control method (Abadie and Gardeazabal, 2003; Abadie et al., 2010) is a data-driven approach used to construct a counterfactual for a treatment unit from a pool of candidate untreated donor units. In large part due to the rhetorical power of its visual evidence—a synthetically-weighted time-series that often closely matches the treated unit's time series—the method has become a staple for economists and policy analysts conducting comparative case studies. The logic used to justify synthetic controls is simple and compelling: if the synthetic control group outcomes match those of the treated unit across the entire pre-intervention period, it should project a valid counterfactual for the treated unit into the post-intervention period.

We use Monte Carlo simulations to investigate a specific way the synthetic control method can systematically fail to project a valid counterfactual in social science applications. In a typical panel data generating process (DGP), the estimator systematically "matches on noise" at the expense of matching on structural parameters, and as a consequence projects a distorted counterfactual into the post-treatment periods. This bias is a form of "over-fitting" to pre-period observed outcomes that, combined with mean reversion in error terms in the post-period, leads to a systematic distortion in the post-period counterfactual estimate[2]. This bias occurs even when the synthetic control and treatment group outcomes are well matched in the pre-treatment period. The problem depends on the location of the treated unit among the control units. It is most severe when the treated unit's unobserved structural parameters (e.g. intercept and trend) are away from the center of the distribution of the donor units' parameters.

Prior studies have examined bias in the synthetic control estimator arising from matching on error terms. Abadie et al. (2015) brings up the point that synthetic controls does not do well when there are a small number of pre-periods, using the rationale that the estimator may end up matching on error. Our paper, building off insights developed in Smith (2013) in a less flexible and less realistic data generating environment, develops that intuition for practitioners and demonstrates the bias as a function of a larger set of determinants. Beyond those works, Ferman and Pinto (2021) is the study most closely related to our own, and provides an analytical proof of the existence of an asymptotic bias due to matching on error terms. Several other papers (reviewed in in Section 2.2) have explored alternative concerns with synthetic control models and developed alternative estimators. However, the focus of these works is on proving the existence of an asymptotic bias in various theoretical data

[2] Here we think of over-fitting as when synthetic control weights are estimated not using information from the true factor loadings but also from noise. In many contexts the effect of noise is symmetric around the true parameter and therefore over-fitting only inflates the variance of the estimator but does not increase bias. One of the points we make in our paper is that this does not always hold with synthetic controls in finite-sample conditions. We discuss this in section 3.1.

environments. Our focus is on exploring the bias in ways that we hope will inform practice.

Using Monte Carlo methods, our primary goals in this work are to characterize the nature and determinants of bias, to provide intuition and guidance to practitioners regarding the bias, and to propose corrections for addressing the problem. We investigate the determinants of the bias in finite sample settings similar to those commonly found in social science applications. Of particular relevance to practitioners, we show how empirical analogs of unobserved parameters (such as level and trend differences) can reveal potential risks in a researcher's specific application. In addition, the correction procedure we propose is feasibly implementable in real-world applications, and points towards future potential approaches for bias correction.

We first provide background on the synthetic control estimator itself and its implementation in social science research. We then illustrate the bias in a graphical example, and provide mathematical intuition as to how the bias is generated. Next, we use a series of Monte Carlo simulations to evaluate how various DGP parameters affect the magnitude of the bias. We show how, within the context of our simulations and assumed DGP, the bias varies with: (i) the degree to which the treated unit's unobserved structural parameters are in the tails of the sample distribution; (ii) the variance of the error term; (iii) the number of donor units; (iv) the pre-treatment period length; and (v) the complexity of the DGP (measured in our simulations by the polynomial order of group-specific time trends).

We then propose an intuitive correction procedure for practitioners to employ in robustness tests or informal model checking procedures. Using a parametric bootstrap, we estimate the magnitude of bias, given estimates of the functional form of the DGP and observable features of the data. We then subtract this estimated bias from the main estimate. This methods performs reasonably well at reducing bias, but at a cost of increased variance[3].

We also compare the performance of our corrections to that of an Interactive Fixed Effects model. This approach performs well in our simulations and does not appear to be subject to the same "matching on noise" bias that affects synthetic control methods.

Finally, we investigate the bias in a real-world application. We conduct a partial re-analysis of a recent empirical paper (Billmeier and Nannicini (2013)) which uses synthetic control methods to estimate the macroeconomic effects of free trade agreements at the national level. Using the dates of free trade agreements and data from the world bank on 17 countries, we estimate both baseline synthetic control models that match on all pre-period Y values, and models employing our alternative, corrected estimates. We find that the bias corrections are often large enough to be economically meaningful, can change qualitative interpretation, and can even lead to estimates of

[3] In the simulations we consider the increased variance typically dominates the lowered bias, leading to higher mean squared error (MSE).

differing sign.

In summary, in this paper we analyze a particular form of bias in the synthetic control estimator, one we believe is not yet appreciated by applied researchers. The bias arises whenever the treated unit's unobserved parameters are away from the middle of the distribution of control unit parameters, and there is idiosyncratic noise in the DGP. Our Monte Carlo simulations of the determinants of bias can provide guidance on when synthetic controls are more or less likely to generate biased estimates. Additionally, our correction procedures can provide alternative specifications and robustness checks when synthetic controls estimates are central to an empirical analysis.

2 Synthetic Controls: Background, Estimation and Graphical Representation

Synthetic control techniques were developed primarily as a tool for quantitative comparative case studies that exploit panel variation (e.g. state and year) to estimate the impacts of an external event (e.g. policy changes) that affected a single "treatment unit" (Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015). The synthetic control method compares the outcome of interest for the unit affected by the intervention or policy to an algorithmically-determined weighted average of donor (control) units that were not affected. In this way it reduces researcher degrees of freedom over choice of control groups and allows the data itself to determine appropriate matches. As with matching methods in general, the approach works when the treated unit, absent the intervention, would have evolved in a similar way to the (weighted) mean of the matched units, in this case synthetic control group. This means that selecting the correct weights for the control units is of central importance.

2.1 Estimation

The estimation of treatment effects using synthetic control methods involves three steps. First, a weight (w_i) is chosen for each control unit i, numbered from 1...G. These weights are derived from an optimization procedure that minimizes the mean squared distance between a weighted average of control group variables (X) and their corresponding values in the treated unit. Second, these weights are applied to control group observations to create a "synthetic control unit" with period-specific weighted average outcomes defined as $Y_{S,t} = \sum_{1}^{G} w_i * Y_{it}$ (where G is the number of control units). The "fit" of the model is generally assessed either informally and visually by the ability of the synthetic control outcome time-series to match that of the treatment unit in the pre-treatment periods, or by the magnitude of the real mean squared prediction error over the pre-period. Conditional on

an adequate pre-treatment fit, the treatment effect is then calculated as the difference between the treated unit and the synthetic control unit in the post-treatment periods. Finally, inference is typically performed through a permutation-based procedure. These steps are outlined in detail below.

2.1.1 Model of Unit Level Time-Series

Abadie et al (2010) propose the following model for outcome Y in period t to justify the use of synthetic controls.

$$Y_{it}^{N} = \theta_t Z_i + \lambda_t \mu_i + \epsilon_{it} \tag{1}$$

Y_{it}^{N} is the value of outcome Y_{it} absent any treatment. θ_t is a vector of time-varying coefficients on observed variables Z_i, which are constant across time within a unit (i). μ_i are the unobserved (and time-invariant) factor loadings for unit i, and the vector λ_t contains period-specific common factors.

Next, let the value of Y_{it} conditional on having been exposed to the treatment be defined as:

$$Y_{i=treated,t} = Y_{it}^{N} + \delta_t D_{it} \tag{2}$$

Where D_{it} is an indicator equal to one if unit i has been exposed to the treatment in time t, and zero otherwise, and δ estimates the effect of the policy or treatment. We observe a balanced panel of data with G control units and a single treatment unit, $Y_{i=treated,t}$ [4], each observed for T periods. There are T_0 pre-treatment periods, and in period $T_0 + 1$ the treated unit is exposed to the treatment and immediately experiences the (potentially period-specific) treatment effect.

2.1.2 Choosing Weights

At the core of the synthetic control method is an algorithm that chooses a set of weights for the control units. These weights are chosen to make the weighted average of control group variables X match their treated-unit counterparts as closely as possible. A critical step in this process is for the researcher to choose the predictor variables (X) for the synthetic control machinery to match on. Predictors are generally linear combinations of pre-period covariates and outcomes, and it is typical to choose outcomes in different parts of the pre-treatment period in order to capture secular time trends driven by unobservable factors.

[4] It is possible to run this procedure for more than one treated unit. One approach is to take averages of observables to combine all treated units into a single unit (e.g.Kreif et al., 2016).

Let W be a (G x 1) vector of non-negative weights w_i that sum to one. If the data has the factor structure in equation (1) above, Abadie et al (2010) show that any optimal weight w_i^* satisfies:
$\sum_{i=1}^{G} w_i^* Y_{i,1} = Y_{i=treated,1}$, ..., $\sum_{i=1}^{G} w_i^* Y_{i,T_0} = Y_{i=treated,T_0}$, and $\sum_{i=1}^{G} w_i^* Z_i = Z_{i=treated}$
Further, if the number of pre-periods is large relative to the scale of the errors, then for $t > T_0$ the following is an unbiased estimate of δ_t:

$$\hat{\delta_t} = Y_{i=treated,t} - \sum_{i=1}^{G} w_i^* Y_{i,t} \tag{3}$$

Here $Y_{i=treated,t}$ indicates the outcome for the treated unit in time period t. Control unit weights W are selected such that the average distance between the resulting synthetic control and the treated unit for each predictor is minimized. An optimization procedure selects weights that minimize the following function:

$$(X_{i=treated} - X_0 W)' V (X_{i=treated} - X_0 W) \tag{4}$$

Where $X_{i=treated}$ is a (K x 1) vector of predictors (chosen from Z and $Y_{t \leq T_0}$) for the treatment unit, X_0 is a (K x G) matrix of the predictors for the control units, and V is a (K x K) diagonal matrix with the diagonal elements representing the importance of each predictor. W is a (G x 1) vector of time-invariant, non-negative weights to be estimated for each control unit and constrained to sum to one. In this way, the weights are chosen to make the distance in the X variables between the treated and synthetic control units small. This is done in the expectation that reducing the distance between $X_{i=treated}$ and X_s (the synthetically weighted average, computed as $X_s = \sum_{i=1}^{G} w_i^* * X_i$) will lead to a counterfactual Y_s that is also appropriate for $Y_{i=treated}$ in the post-treatment period. As Abadie et al. (2015) show, this result holds as T_0 goes to infinity, and there is a potential for bias in the presence of finite pre-periods. However, our work is the first the characterize the dynamics and determinants of this bias, and relate these features to mismatches on unobserved factor loadings.

2.1.3 Inference

The typical inference procedure for synthetic control estimation involves a permutation test based on the null hypothesis of no effect. The procedure uses the donor unit pool as a set of "placebo" treatment states and estimates a "placebo treatment effect" for each control unit by assigning that unit as the placebo-treatment group and the original treatment year as the placebo treatment year. The distribution of placebo estimates then acts as an estimate of the sampling distribution of the estimator under the null hypothesis of no effect. A p-value can then be computed, based on the rank of the magnitude of the actual estimated treatment effect relative to the distribution of placebo

estimates. Abadie et al. (2010) recommend selecting only those placebo estimates which have a good pre-treatment period fit.

2.2 The Development of Synthetic Control Methods

In a series of foundational papers, Abadie and Gardeazabal (2003) Abadie et al. (2010) and Abadie et al. (2015) introduced social scientists to the synthetic control method. The method was first proposed in Abadie and Gardeazabal (2003) where it was applied to study the effects of political conflict in the Basque region on economic development. Abadie et al. (2010), which focused on the effects of a California anti-smoking measure on cigarette sales, further developed the econometric grounding and introduced the permutation-based inference procedure described above. Following these papers, there has been a growth of interest in using the method on a range of topics in economics and other social sciences as well as a series of methodological improvements for implementation and theoretical advances on the underlying theory.

2.2.1 Recent Empirical Work Using Synthetic Control Methods

Synthetic controls have most commonly been employed to evaluate state or region-level policy changes affecting aggregate production, health and labor market outcomes. Courtemanche and Zapata (2014) evaluate self-reported health outcomes after the implementation of universal health care in Massachusetts; Bohn et al. (2014) study the immigrant labor market using the 2006 Legal Arizona Workers Act in Arizona; Eren and Ozbeklik (2016) study the passage of right-to-work laws in Idaho and Oklahoma and their effect on union membership, FDI and manufacturing employment; and Jones and Marinescu (2018) investigate the labor market impacts of universal cash income from the Alaska Permanent Fund. Recently, researchers have expanded the domain of analyses to include the effects of a broader range of policies and outcomes, such as the impact of educational reforms on preschool (Fitzpatrick, 2008) and college (Klasik, 2013) performance, as well as criminal justice issues ranging from prostitution (Cunningham and Shah, 2017) to gun control (Donohue et al., 2019). A number of papers have alternatively used synthetic controls to study country-level events such as the effect of natural disasters (Cavallo et al., 2013), resource discovery (Smith, 2015), political systems (Nannicini and Ricciuti, 2010) and monetary policy on economic growth (Lee, 2010). Other analyses examine alternative group level units, such as universities (Hinrichs, 2014), financial firms (Acemoglu et al., 2016), grocery stores (Kiesel and Villas-Boas, 2013), neighborhoods (Gautier et al., 2009), and even physician specialty classifications (Chen et al. 2018).

2.2.2 Bias in Synthetic Control Methods

Abadie et al. (2010) proved the consistency of the synthetic control estimator given a long pre-intervention period, by showing that a bounding function for bias asymptotically approaches zero as the pre-intervention length T_0 grows. Their bounding function bounds their term R_{1t}, which indirectly comes from biases in the weighted synthetic control factor loadings (see appendix B of Abadie et al. (2010) for a full definition of R_{1t}).

Ferman and Pinto (2021) show that the bound on this bias does not go to zero if the pre-period fit is imperfect. This may especially be the case when the number of donor units G is fixed, and small compared to T_0[5]. The framework in Ferman and Pinto (2021) involves a fixed number of donor units, and considers asymptotics as the number of pre-treatment periods grows, $T_0 \to \infty$.

For the model in (1) and (2) without covariates, the bias in the synthetic control estimator will come from biased estimates of the factor loadings of the treatment unit. We distinguish between our potentially biased estimated weights $\widehat{w_i}$, and w_i^{tgt}: the target weight which correctly recovers the treated unit's factor loadings. Then the formula for bias is in post-period t is: $\lambda_t \cdot E_\epsilon[\mu_{treated} - \sum_{i=1}^{G} (\mu_i \cdot \widehat{w_i})]$. The expectation is taken over the error terms (ϵ), a random variable which determines $\widehat{w_i}$. μ_i is the vector of unobserved factor loadings for that donor unit. Bias in period t occurs when the $\widehat{w_i}$ weights fail to recover the factor loadings of the treated unit. This formula for the bias is found in Ferman and Pinto (2021), and also applies in our paper.

Similar to our paper, the bias in Ferman and Pinto (2021) is based on the synthetic control algorithm trying to match in part on model errors. In Ferman and Pinto (2021), in a DGP with non-trending unobserved random factors, the bias can be corrected by demeaning the pre-treatment average of the outcome for each unit. In other important papers, Ferman and Pinto show that that visual displays of inference by placebo can be misleading (Ferman and Pinto, 2017), and that in applications with few pre-periods there is substantial room for specification hunting and cherry picking of results (Ferman et al. (2020)). Another paper that discusses bias in synthetic controls is Kaul et al. (2022) which shows that bias arises when matching on all pre-intervention periods. The formulation of bias in our paper is different than in Kaul et al. (2022). In that paper, the weights can neglect matching on important (observable) covariates. In our paper, we are concerned about weights mis-matching on the unobservable factor loadings. The approach we take in this paper is focused on using Monte Carlo simulations to provide intuition about the sources of bias to practitioners, and is complementary to these existing papers.

In contrast with earlier work, in this paper we think of the bias as conditional on the locations

[5] Ferman and Pinto (2021) is an updated version of Ferman and Pinto (2016), and was developed concurrently with this paper.

of the factor loadings μ_i. In the models we consider, the factors λ_t are fixed polynomial series, and so are not random. Below we show the case for when the dimensonality of the factors loadings is one ($D = 0$). In this case, μ_i is simply a group specific intercept.

$$Bias = \lambda_t \cdot E_\epsilon[\mu_{i=treated} - \sum_{i=1}^{G} (\mu_i \cdot \widehat{w_i})|\mu_{i=treated}] \tag{5}$$

We show results for the bias in equation 5 in our illustrative model of section 3.1, and in Online Appendix A. Most of the Monte Carlo results in this paper additionally treats the factor loadings μ_i as random, and conditions on the rank of the factor loadings for the treated unit. When there is more than one factor, we condition on the rank of the sum of the factor loadings across the $D + 1$ dimensions of unobserved factor loadings:

$$Bias = \lambda_t \cdot E_\epsilon[\mu_{i=treated} - \sum_{i=1}^{G} (\mu_i \cdot \widehat{w_i})|Rank(\sum_{d=0}^{D} \mu_{d,i=treated}) = r] \tag{6}$$

The bias we describe in this paper is a *conditional* bias. We show that the bias depends on the location of the treated unit among the control units. What we exactly condition on varies across simulation specification but is often either the intercept (when $D = 0$) or the sum of the intercept and trend (when $D = 1$). We consider the $\epsilon_{i,t}$ as random, and the rank of $\mu_{i=treated}$ in the distribution of μ_i as fixed. That is, our bias is conditional on the location of the treated group's μ within the distribution of donor units: holding the relative location of the treated unit fixed.

Although we conceptualize the bias as depending on the relative locations of the factor loadings, in our simulations we additionally model the μ_i as randomly distributed, and choose the treatment group as the group with appropriately ranked μ_i. That is, our simulations condition on the rank-location of the factor loadings. In this way we measure the average bias across a range of possible realizations of the μ_i. In Appendix E, we re-run a simulation with fixed factor loadings, where the only randomness comes from the model errors ϵ. This exercise gives results very similar to our main simulations.

If we were to additionally consider random assignment across units to "treated unit" status, then the conditional biases could average out to zero. However, in any given application the location of the treated unit among the donor units is fixed. We believe that this conditional bias should be an important consideration in applied work. This is particularly true for case-study approaches for which the synthetic control estimator was developed. In these applications the research is focused on a particular, already realized "treated" unit.

We use Monte Carlo simulations to provide a direct examination of finite sample (fixed T_0) bias in synthetic control estimates. In order to explore the applicability of these findings, we investigate

a number of dimensions of our simulations. We consider settings with: (1) a fixed (and moderate or small) pre-treatment-period T_0; (2) more donor units G than T_0; and (3) allowing the synthetic control unit to provide a perfect pre-treatment fit (and in fact requiring a reasonably good fit). We contribute to this setting by documenting a bias that applies to mismatching trending variables.

Our focus on settings with relatively few pre-intervention time periods has particular relevance for empirical policy analyses. Short pre-periods are common in real-world applications. For example: Chen et al. (2018), uses $T_0 = 3$ pre-period years before the implementation of a medicare fee bump under the ACA. Gurantz (2020) employs synthetic controls to evaluate the impact of free community college in Oregon with $T_0 = 4$ pre-period years. Larramona and Sanso-Navarro (2016) use $T_0 = 3$ pre-period years total. Other studies use longer pre-periods but with relatively long post periods. For example, Hu et al. (2018) use $T_0 = 12$ pre-period quarters. Pinotti (2015) uses $T_0 = 10$ pre-period years to predict outcomes over 40 post-period years. Smith (2015) and Billmeier and Nannicini (2013) both perform case studies separately for a number of individual countries, several of which have five or fewer pre-treatment years. Our simulations offer substantive guidance with regard to T_0. For example, we show how, within the context of our assumed DGP, bias decreases as the length of the pre-period increases.

We offer similar guidelines to practitioners on the the importance of donor group size, and perhaps more importantly on the importance of where the treated unit is in the distribution of potential matches. For example, Sun et al. (2019) use seventeen donor cities when looking at the impact of environmental credit regulations on firms in a single treated city, substantially smaller than the 30 or so control states used in most applications. Our work suggests that this situation with relatively few donor units is not as much of a concern for bias related to matching on noise, and that a greater concern might be warranted if pre-treatment rank of $Y_{i=treated}$ is in the tails of the distribution of the donor Y_i's.

2.2.3 Extensions and Refinements to Synthetic Control Methods

A number of papers relax assumptions in the original synthetic control algorithm. For example, Doudchenko and Imbens (2016) propose a method that allows synthetic weights to be negative and to not sum to 1, while Kreif et al. (2016) propose a version allowing for multiple treated units. Other papers contextualize synthetic control within a broader framework of matching and difference estimators, including Xu (2017), which frames synthetic control in a fixed-effects framework, and Athey and Imbens (2017), which places the synthetic control estimator within the broader potential outcomes framework and treatment-effects estimation literature.

Ben-Michael et al. (2021) propose an "augmented" synthetic control estimator that is viable in settings where a good pre-treatment fit is infeasible. They combine a direct modeling of the relationship (using a ridge regression) between the pre-treatment variables and the post-treatment outcome, with synthetic control matching. Arkhangelsky et al. (2021) propose a "synthetic difference-in-differences" estimator that is a generalization of both traditional synthetic control models and also a generalization of panel fixed-effects difference in difference models. Their estimator has a double-robust property with potentially reduced bias compared to the traditional synthetic control model. They prove consistency of their estimator as both $G \to \infty$ and $T \to \infty$, and in simulations show improvements over traditional synthetic control model with G and T as small as 50.

Closer to our own work, a few papers have attempted to evaluate specification issues directly using simulations similar in spirit to those employed here. Particularly focused on concerns about matching on noise, Peri and Yasenov (2019) show that choosing small sub samples of potential donor groups can cause inflated type 1 error rates due to measurement error in the outcome variable. However, unlike Peri and Yasenov (2019), our documented bias applies even in cases where there are a relatively large number of control groups, demonstrating that matching on noise continues to be problematic even when large donor pools are available for matching

In terms of specifications, our baseline estimates follow the potentially problematic but widely-used specification of matching on all pre-period outcome values, as discussed Kaul et al. (2022). That work argues both theoretically and through a replication analysis that matching on all pre-period outcome variable values guarantees that weights of 0 are assigned to all non pre-period outcome variables, thus forcing the fit to rely entirely on the pre-intervention outcome time series. However, this restricts the ability of underlying structural factors to predict changes in the time-series path and for their influence to be replaced by noise in the outcome measurement. While Kaul et al. (2022) argue there is potential for bias from such a specification, it remains common in practice, and so appropriate as a benchmark for our purposes. Importantly, the bias we describe persists whether we control for all pre-period outcomes or only some periods. Furthermore, while Kaul et al. (2022) demonstrate empirical instances of bias, their focus is on specification and choice of matching variables, not the nature or mechanics of the bias that we highlight here.

2.3 Graphical Demonstration and Data Simulation

Much like regression discontinuity and event study analyses, synthetic control methods derive rhetorical power from the elegance and (seeming) transparency of their graphical presentation. In order to bridge the background discussion above with the simulation-based results to come, we generate and analyze one realization of our simulation model and display the results.

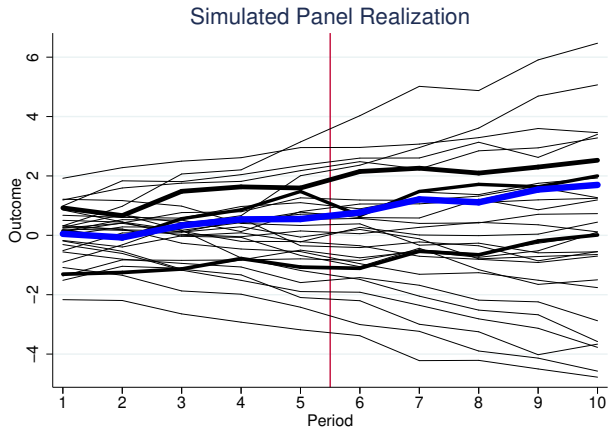
We generate time-series data from a random-intercepts and random-slopes DGP:

$$Y_{it} = \mu_{0,i} + \mu_{1,i} * t + \delta * Treated_{it} + \epsilon_{it} \tag{7}$$

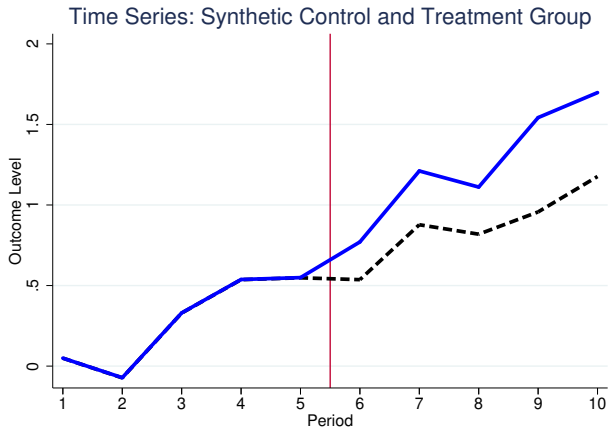
We draw intercepts $\mu_0 \sim N(0, 1)$ and trends $\mu_1 \sim N(0, 0.25)$ independently for each of 30 donor units and one treatment unit. There are 10 time periods ($T = 10$), 5 of which are pre-period ($T_0 = 5$). ϵ_{it} is drawn independently for all observations from a $N(0, 0.2)$ distribution. The "treatment unit" has the 8th ranked intercept and the 26th ranked trend from the realized sample of 31 values of μ_0. and μ_1. For this and all further simulations, we set the treatment effect to zero ($\delta = 0$).[6]

The raw data from a single realization of the DGP described above are presented as a panel of unit-level time-series graphs in Figure 1a. The thick blue line represents the treated unit. The gray lines represent the 30 donor units from which the procedure generates a synthetically-weighted counterfactual estimate. We then fit a synthetic control model to the data, matching on all pre-period outcome values, and shade the donor units accordingly, with their thickness varying according to the weight they are assigned by the algorithm (including light gray lines for zero weight). The blue line is squarely within the distribution of both potential and realized donor units for the entirety of the pre-period, and gently, but noisily, trends upwards with no apparent or programmed break in the trend or level when the placebo "treatment" occurs in period 6.

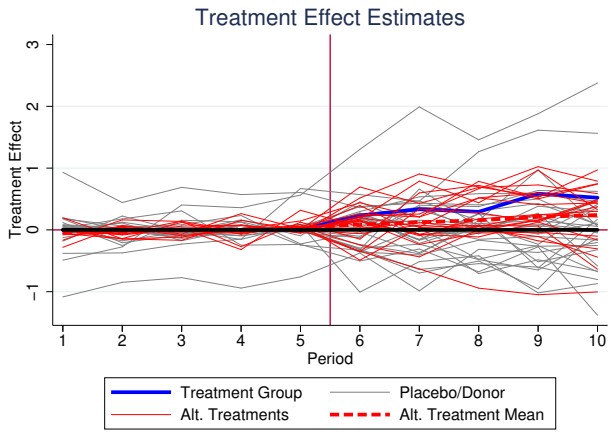
[6] These variance, group number and time period parameters are the baseline values used throughout Section 3. From Section 4 on, our baseline model is similar to Eq. 7 but we increase the trend function complexity and set $T_0 = 10$ in order to allow for higher dimensional trend functions to reveal their shape over the pre-period.



(a)



(b)



(c)

Figure 1: Graphical Demonstration of Bias.

Data are from one realization of the baseline DGP. The top panel depicts the raw time-series by group. The middle panel depicts synthetic control fit. The bottom panel provides placebo-based inference with grey lines denoting estimates from remaining "control" groups, and red lines from simulated "treatment" groups with the same rank of the unobserved parameters ($\mu_{0,i}$ and $\mu_{1,i}$).

We then calculate $Y_{S,t}$ and graph it in a dashed black line in panel (b), repeating the treatment group blue line from panel (a). The synthetic control estimate is perfectly matched in the pre-period and then diverges from the synthetic control group in the post period, despite the lack of any true treatment effect in the DGP. The distance between the treated unit and control group increases nearly linearly over the post-period. Taken at face value, the estimate would suggest a positive treatment effect that increases in magnitude over the post-period. But as will be shown, the divergence between the treated unit and the synthetic control is driven by an underlying mismatch between the structural trends of the treated unit and the synthetic control.

We then contextualize the magnitude of the implied treatment effect in the bottom panel (c) using the standard inference strategy and a second, simulation based approach. Panel 1c preserves the data from panels (a) and (b) but graphs the difference between the treatment and synthetic control groups instead of the levels of both groups separately (blue line). The gray lines are the "placebo" estimates proposed in Abadie et al (2010), which are the synthetic control estimates for the 30 donor units in the top panel. These trace out an approximation of a "sampling distribution" of synthetic control estimates when the null effect of 0 is assumed to be true, treatment is assumed to be exchangeable and the estimate is unbiased. The first two conditions hold by construction, but, as we show in red in panel (c), the third condition does not.

The thin red lines in panel (c) represent results from 20 additional simulation runs with the same DGP settings and the treated unit again chosen to be the one with the 8th ranked μ_0 and 26th ranked μ_1 in a sample of 30 control units from this DGP. These "alternative treatment" estimates are not centered on zero. The mean of 500 similarly-constructed alternative treatment simulation runs is traced out by the dashed thick red line, which begins the post-period at 0 but then increases linearly over the post-period. The slope of the dashed red line can be interpreted as a (somewhat noisy) estimate of the average degree to which the synthetic control estimator systematically under-estimates the underlying trend of the treatment group when the treatment group happens to have the 26th (of 31) ranked trend and 8th ranked intercept from this DGP. A simple interpretation of these results is that about half of the treatment effect seen in Figure 1 is due to bias and half due to sampling and fit variation (and none due to any real treatment effect).

3 Noise Induced Bias

What is going wrong in the example of Figure 1 above? We illustrate below that the bias arises from asymmetrically matching on errors when the "true value" (before errors) of the closest control units on one side of the treated group are nearer than on the other side of the treated unit. This is likely

to occur when the treated unit is far from the mass of control units. Both noise and asymmetry are key determinants of the problem. In terms of noise, the match in the pre-period ($t \leq t_0$) of $Y_{S,t}$ to $Y_{i=treated,t}$, is based in part on matching noise in the treatment and control units' outcome variables (that is, matching to $\epsilon_{i,t}$). This induces a systematic deviation between $\mu_{i=treated}$ and the "synthetic μ", defined as $\mu_S = \sum_{i=1}^{G} \widehat{w}_i \cdot \mu_i$, which is the underlying trend component of Y_S. And this in turn leads to the realized bias in the post-treatment periods when, by construction, ϵ averages 0 in the post period.

As discussed in Section 2, the synthetic control algorithm is designed to match on X between the treatment group's values ($X_{i=treated}$) and the synthetically weighted average X_s. When X includes both predictor variables Z and pre-treatment values of the outcome, matching on the outcomes can help to capture μ_i. By matching on levels and secular trends in the outcome in addition to known determinants, the model implicitly matches on the unobserved determinants that drive those trends. A good match on both Z and $Y_{i,t}$ (for $t \leq T_0$), can then be hopefully interpreted as evidence that $\mu_{i=treated} = \mu_S$ (Abadie et al., 2010). This is how the synthetic control weights allow researchers to control for a rich unobserved factor structure.

The problem arises when the pre-treatment outcomes embody not only the unobserved factor loadings, but also idiosyncratic noise, ϵ_{it}. In this setting matching on both the noise and the true parameter is a type of over-fitting that (unlike typical cases of over-fitting) generates not only estimation variability but also bias. In the commonly modeled empirical settings we explore, the algorithm will regularly match in part on this idiosyncratic variation in the process of optimizing model fit in the pre-period. This process of matching to noise induces a systematic relationship between the error terms in the donor unit observations and the synthetic weights assigned to those units. Mean reversion in the post-period ensures that the synthetic control group then projects a distorted counterfactual to the treatment group into the post-period.

The magnitude of the bias depends on several factors and we investigate these in turn in our Monte Carlo simulations in Section 4. In this section, we first provide intuition into the nature of the bias using a stylized thought-experiment that demonstrates the fundamental role of noise in the matching process. We then use simulation techniques to demonstrate how the bias depends on the variance of idiosyncratic error terms and the location of the treated unit's unobserved factor loadings among the distribution of the control unit factor loadings. Finally, we connect the bias to the magnitude of the difference between $\mu_{i=treated}$ and μ_S.

3.1 Illustrative Example

We consider a simple, stylized case-study to provide mathematical intuition for understanding the bias. Let the factors λ_t be a single constant $\lambda_t = 1$, for all t. This implies that the unobserved factor loadings μ_i are i-level fixed effects and Eq. 1 reduces to:

$$Y_{it} = \mu_i + \delta * Treated_{it} + \epsilon_{it} \tag{8}$$

In this subsection, we consider a setting where there is exactly one pre-treatment time period, and there are exactly two "donor" units, whose true factor loadings μ_i are on either side of the treated unit, and with small enough error realizations that their observed pre-treatment $Y_{i,t=1}$ are also on the same side of the treated unit as μ_i.

The weights produced by matching the pre-treatment observation $Y_{i=treated,t=1}$ will be based on how close the treated unit is to each of the two donor units. The ideal "target" weights depend on the true locations of the factor loadings:

$$w_0^{tgt} = \frac{\mu_1 - \mu_{treated}}{\mu_1 - \mu_0}$$

$$w_1^{tgt} = \frac{\mu_{treated} - \mu_0}{\mu_1 - \mu_0}$$

The estimated SC weights depend on the realized locations in period 1. In the formulas below the model errors ϵ refer to first period errors only.

$$\widehat{w_0} = \frac{\mu_1 - \mu_{treated} + \epsilon_1 - \epsilon_{treated}}{\mu_1 - \mu_0 + \epsilon_1 - \epsilon_0}$$

$$\widehat{w_1} = \frac{\mu_{treated} - \mu_0 + \epsilon_{treated} - \epsilon_0}{\mu_1 - \mu_0 + \epsilon_1 - \epsilon_0}$$

We consider the factor loadings $(\mu_0, \mu_1, \mu_{treated})$ as fixed, and we condition on these when we compute the bias. We consider the model errors (ϵ's) to be random. The bias in the estimated treatment effect will depend on the bias in the estimated synthetic control weights:

$$Bias = E[(w_0^{tgt} - \widehat{w_0}) \cdot \mu_0 + (w_1^{tgt} - \widehat{w_1}) \cdot \mu_1]$$

with the expectation taken over the distribution of the first period model errors.

For a given realization, if $\epsilon_{i=1,t=1}$ is positive then donor unit 1 will be farther away from the treated unit, and so will get less weight than it should. This will give a downward error to the synthetic counterfactual, and an upward error to the estimated treatment effect. We might then

think that in expectation this would balance out between donor units above and below the treated unit, leading to no bias. However, this is not necessarily the case. This is because the impact of an error term ϵ on the error in the SC weight depends on the distance between the treated unit and the donor unit. For example, if among the donor units, unit 0 is closer to the treated unit than is unit 1, this will lead to an under-weighting of unit 1, a downward-biased counterfactual for the treated unit, and thus an upward-biased estimated treatment effect.

To illustrate this phenomenon, we use numeric integration to calculate the bias for a setting where $\mu_{i=0} = 0, \mu_{i=1} = 1$. In this calculation these are fixed. We place the treated unit in between, considering fixed values $\mu_{i=treated} \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, and assume its observed value has no error. We model the errors for the control units as randomly drawn from a truncated normal $\epsilon \sim N(0, \sigma^2)$, truncating the distribution of ϵ symmetrically at $\pm min(\mu_{treated}, 1 - \mu_{treated})$ so that no draw of errors can change the rank ordering of observed values (ensuring that $y_{i=0} < y_{i=treated} < y_{i=1}$). We then calculate the bias in the estimated treatment effect. This bias comes from mistakes in the weights assigned to the control units, such mistakes arise from noise in the measured outcomes of those control units.

For this setting, the bias can be expressed as:

$$Bias = E(\frac{\mu_{treated} \cdot (\epsilon_1 - \epsilon_0) + \epsilon_0)}{1 + \epsilon_1 - \epsilon_0}) \tag{9}$$

with the expectation taken over the distribution of the first period model errors, ϵ_1 and ϵ_0. The derivation of this formula and further details are provided in Online Appendix A. Appendix A also justifies a key diagnostic of the bias used throughout the paper, the Synthetic Control Average Error (SCAE: see section 3.2.2), which reduces to the the bias formula presented in Eq. 9.

Results are presented in Figure 2. This shows that unless the treated unit is exactly mid-way between the donor units ($\mu_{i=treated} = 0.5$), there is a systematic bias. When the treated unit is closer to $\mu_{i=1}$, then $\mu_{i=1}$ will on average get too much weight, and the counterfactual $\widehat{y}_{i=treated,t=1}$ will be too high. This counterfactual being too high would in turn lead to a negative bias for estimated treatment effect in a subsequent period. The reverse holds when the treated unit is closer to $\mu_{i=0}$. The amount of bias is increasing in the magnitude of noise, and in the distance from the treated unit away from the middle of the range. These results are echoed below in our richer synthetic control simulations in Section 3.2.

What would cause an asymmetric distribution of donor units around the treated unit? This will result from an asymmetric distribution of the density of μ_i around the treated unit. So, for a typical symmetric and unimodal distribution (such as Gaussian) this will happen whenever the treated unit is not located at the mode. More generally, if the distribution of μ_i is such that the density decreases

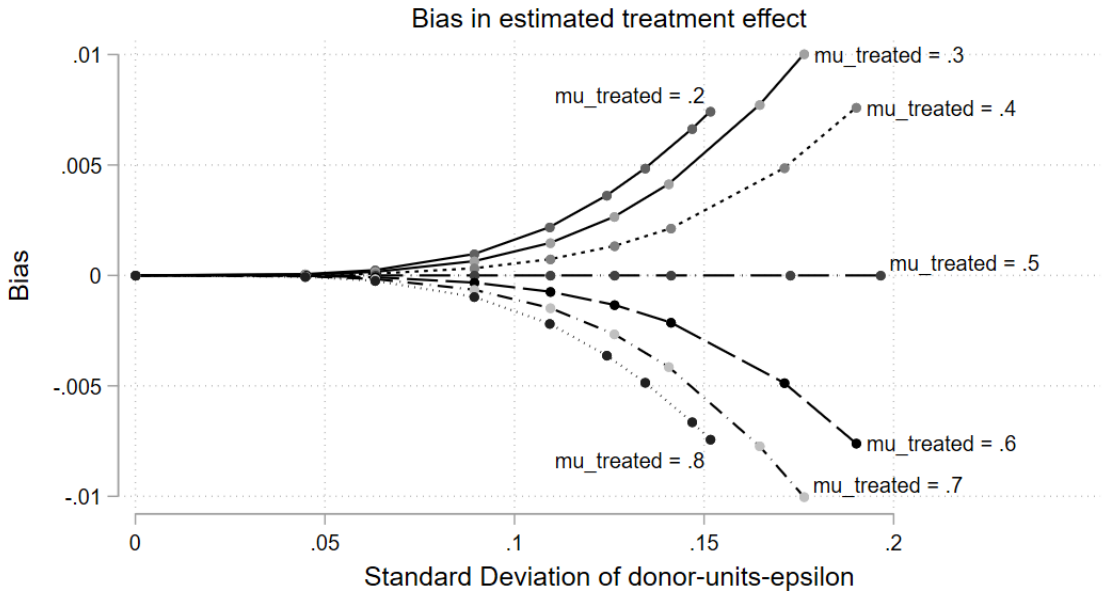


Figure 2: Computational Demonstration of Mathematical Intuition

This figure shows the bias for the estimated treatment effect, as a function of noise of the donor units, and of the true location of the treated unit. There are two donor units whose intercepts are: $(\mu_0 = 0, \mu_1 = 1)$. The location of the treated unit is given by $\mu_{i=treated}$ (measured without error). The error distribution for donor units is truncated Normal, truncated so that no epsilon could change the rank ordering of observed values. See Section 3.1 text for more details.

towards the boundary, and if the treated unit is not in the center of the distribution, then there will be more density of donor units on the side toward the center, and less density on the other side. This will mean that on average the donor units are closer on the side toward the center, which will result in a systematic bias in estimated treatment effects away from zero.

To summarize, the bias comes from a combination of three factors. First, it depends on (1) whether there is a difference in how close (on average) donor units are on one side of the treated unit versus the other. This interacts with (2) the importance of distance on the impact of "how noise in donor units influences their weights and thus the counterfactual." Finally, in many reasonable settings we expect that (3) the treated unit being further from the center of the distribution leads to an asymmetric density of donor units, and thus a difference in expected distance above versus below the treated unit. The combination of these factors is illustrated in the impact of the rank of the treated unit on bias, in our simulations in sections 3.2.2 and 3.2.3 below.

The thought experiment extends to additional pre-periods and alternative choices of matching variables. Consider the case where there are $T_0 > 1$ pre-treatment periods of data, and we are trying to match on pre-treatment average Y. In this case the same problem will apply, only based on $\bar{\epsilon}_t$ averaged over $t \leq T_0$. Because of the averaging over the error term over the pre-periods, the bias should be decreasing as T_0 gets bigger and thus the variance of $\bar{\epsilon}_t$ gets smaller. In the next section we demonstrate the implications of this thought experiment empirically, showing that both the variance of ϵ and the number of pre-periods, as well as the location of $\mu_{i=treated}$ in the distribution of μ_i, affect the magnitude of bias.

3.2 Simulation Evidence of Noise-Induced Bias

In this section we use Monte Carlo simulations to empirically explore the bias described above in a panel data setting. We show results from 3 sets of Monte Carlo simulations and demonstrate the bias visually using simple figures graphing bias across various specifications of the model parameters and hyper-parameters. We first show that the bias in a basic random-intercepts DGP (Eq. 7) exists only when there is a positive variance on the period-specific error term (noise). We then demonstrate that the biased matches generated by the synthetic control algorithm persist into DGPs with trends, where the method again partly matches on noise and fails to accurately match on the unobserved components of the DGP. We show that in both cases the bias can be explained by the mismatch between $\mu_{i=treated}$. and μ_S.

All three simulations share the same basic structure. In each of a series of realizations of simulated data from Eq. 7, we use synthetic controls to estimate a treatment effect on the simulated data, matching on all pre-period values of Y. We generate the factor loadings as independent draws from

$\mu_i \sim N(0, I)$, and then assign the treated unit to be the unit at a fixed rank of the distribution of $\mu_{0,i} + \mu_{1,i}$. We then assign $\epsilon_{it} \sim_{iid} N(0, \sigma_\epsilon^2)$, and generate observed data according to Eq. 7. The true treatment effect in the DGP is set to 0, and the average estimated treatment effect across Monte Carlo realizations is our estimate of the bias. We design the simulation (and all simulations performed in this paper) so that the treatment group is always within the convex hull of the donor group distribution by rejecting simulation runs where the treated unit's outcome variable is the largest or smallest value among all control units within any particular pre-intervention period. We do this because we are not interested in cases where the treated unit is an outlier with no possible pre-treatment match, which then trivially produces poor estimates.

3.2.1 Noise and Bias

We first empirically demonstrate the relationship between the bias and variance of the error term using the DGP used to generate Figure 1 and described by Eq. 7, with 30 donor units and 10 periods. We then run the entire Monte Carlo simulation for 4 different values of σ_ϵ and compute the period-specific average estimated treatment effect across realizations.

As discussed in section 2.2.2. the bias we demonstrate comes from the synthetic control algorithm trying to match in part on model errors, and because of this, failing to perfectly recover w_i^{tgt}. We measure this bias conditional on the sum of the factor loadings (For this DGP that is the sum of $\mu_{0,i}$ and $\mu_{1,i}$). As discussed above, for this DGP we have $G = 30$ donor units and $D = 1$ factor loadings.

We choose as the treatment group the unit with the 70th percentile value of the sum of $\mu_{0,i}$ and $\mu_{1,i}$ (over all i) from the realized distribution in the simulated data[7]. This generates an asymmetry in the distribution of μ around the treatment group[8]. This choice of treated unit is illustrative; an alternative region of μ would produce a different magnitude of bias, but that bias would stem from the same cause of matching on noise even as the sign and magnitude of the bias change. The specific formula of the bias in post period t is now: $Bias = \lambda_t \cdot E_\epsilon[\mu_{treated} - \sum_{i=1}^{30} (\mu_i \cdot \widehat{w_i}) | Rank(\sum_{d=0}^{1} \mu_{d,i=treated}) = 0.7]$.

Figure 3 graphs the bias in our simulations by period. The solid line averages from simulations where σ_ϵ is set to 0 (removing idiosyncratic errors from the DGP). The dashed lines represent variances of 0.2, 0.5 and 2, respectively.

[7] This is a compromise mode of treatment assignment that balances competing interests. Enforcing that the treatment unit contain the 70th percentile on every value of μ_d would ensure better comparability of treatment groups across realizations, but would generate a highly selected sample of potential realizations and treatment groups. Instead, under our treatment assignment rule, in any given data realization some group will indeed be assigned "treated" status. That group effectively represents units with relatively positively influencing unobserved characteristics.

[8] Averaging our simulation results over all ranks of μ would average out the bias. In that sense, unconditional on any particular group being designated the treatment group, the estimator is in fact unbiased. However, conditional on the treatment group having a particular location, the estimator is biased unless the unit is in the center of the realized donor unit distributions of μ.

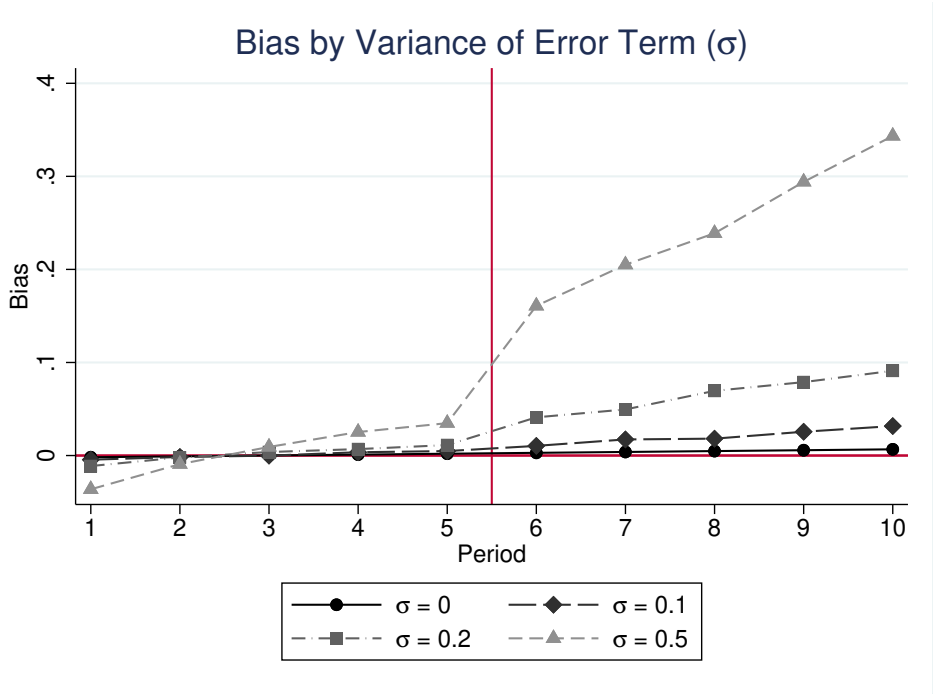


Figure 3: Mean bias estimates by period across Var(ϵ) (σ).

This figure shows the average bias in our Monte Carlo simulations across different pre-periods. To select the treatment group we take the sum of the intercept and slope coefficients for each group and choose the treated group to be the one at the 70th percentile of that distribution. We specify the synthetic control estimator to match on all pre-period values of Y. The DGP includes a linear time trend, 10 pre-periods, and 30 donor groups. Results are averaged over 3000 Monte Carlo replications.

There is no bias when σ_ϵ is set to zero. The synthetic control group matches only on the structural components of the underlying DGP, and so produces an unbiased counterfactual. Increasing the variance of ϵ_{it} a relatively small amount (0.2) generates bias: a mismatch between the synthetic control and treatment group trends leads to a bias that increases over time in the post-period. This comes despite a very small loss of precision in the pre-period fit. Increasing the variance further increases the bias. The higher variances also produce a slight average trend in in the pre-period, despite the fact that we are matching on all pre-treatment periods of the outcome variable. This occurs because in some cases no perfect match is available. This is distinct from the phenomenon we focus on in this paper, which is that a bias occurs even when there is a perfect pre-treatment match. This issue is discussed and illustrated further in Appendix Figure A3, where we show that the (mean-0) trend over pre-periods disappears while the bias persists when you focus on the realizations that produced low real mean square prediction error (RMSPE). We interpret this as a warning to practitioners that choosing a matching specification based purely on pre-period fit does not guarantee a good counterfactual projected into the post-period.

In this paper, we conceptualize the bias as depending on the relative locations of the factor loadings μ_i, and as coming from the randomness of the model errors ϵ. In this and subsequent simulations we additionally model the μ_i as randomly distributed, and condition on the rank of the treated unit. In this way we measure the average bias across a range of possible realizations of the μ_i, conditioning on the rank-location of the factor loadings. In Appendix E, we re-run a simulation in which the factor loadings are fixed across simulation replications. In that simulation the only randomness comes from the model errors ϵ. That exercise gives results very similar to our main simulations presented in Figure 3.

3.2.2 Intercept Rank and SCAE

To illustrate the mechanism of matching on noise and how it relates to the treated unit's location in the distribution of unobserved parameters, we next use simulations to show that synthetic control weights selectively load onto units with systematically non-zero average error terms over the pre-periods (conditional on our assumed GDP). Here we employ the simplified DGP from Eq. 7 where the unit-specific intercepts are the only factor loadings, are random, and are drawn from a N(0,1) distribution. We then compute the mean bias as a function of the rank of μ_0. The changing density of donor groups around the treatment unit (with more density on one side than the other) at the edges of the outcome distribution (high or low ranks of μ_0) implies that the influence of error terms on the matching should be largest towards the tails of the distribution of μ and smaller towards the center, with no influence at the center where the density is balanced on both sides.

To clarify the relationship between bias and matching on noise, we define the *Synthetic Control Average Error* (SCAE) as $\frac{1}{T_0}\sum_{t=1}^{T_0}\sum_{i=1}^{G}\widehat{w_i}\cdot\epsilon_{it}$, the synthetically-weighted mean error term averaged over the donor units for the entire pre-period[9]. In usual empirical practice, this cannot be calculated since ϵ_{it} are by definition unobserved, but in our simulations we know each ϵ_{it}. Since these errors are on average mean 0 in the post-period by construction (they are not included as matching variables), an $SCAE \neq 0$ in the pre-period will generate bias in the post period.

In our illustrative example of section 3.1.1, the exact formula for the SCAE can be worked out, and is derived in Online Appendix section A. For that setting, the SCAE is the same as the formula (9) for the bias of the synthetic control estimator. This underscores the usefulness of the SCAE to understanding the mean bias.

We plot the average (across realizations) SCAE, as well as the mean bias, by treated unit rank

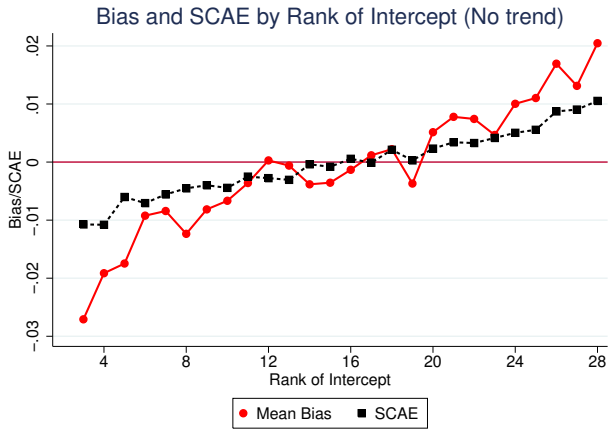
[9] The SCAE is related to the term for bias in Abadie et al. (2010). In Appendix B of that paper, the bias is based on the expected value of their term R_{1t}. That term is the synthetic control weighted average of the unit-specific predicted pre-treatment model error, with predictions coming from a regression of the pre-treatment model errors on the unobservable factors.

of μ_0 in Figure 4a. The solid red line plots the mean bias for each rank of the treated unit intercept $\mu_{0,i=treated}$ and for each rank the bias is estimated across 3000 Monte Carlo runs. The dashed black line overlays a graph of mean SCAE on top of the mean bias graph across the rank of μ_0. Here we are only conditioning on the intercept as shown in the formula for the bias: $Bias = E_\epsilon[\mu_{treated} - \sum_{i=1}^{30} (\mu_i \cdot \widehat{w_i})|rank(\mu_{0,i=treated}) = r]$

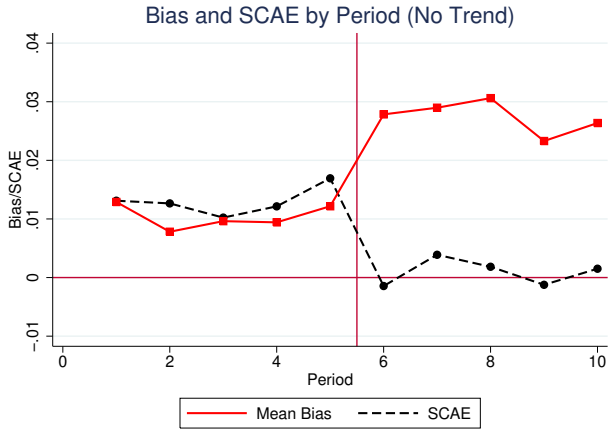
The two curves follow a similar pattern. When the treated unit's rank is in the middle of the distribution of μ_0, the density of control units is symmetric around the treated unit, and there is little to no bias when the treated unit's rank is near the median. However, the bias increases in magnitude for ranks closer to the extremes as the asymmetry of control units around the treated unit increases. The figure suggests that the bias in the synthetic control estimates can be explained by the algorithm matching on a string of realized error terms at the expense of the unobserved structural features of the DGP. The resulting effect on the difference between $Y_{i=treated}$ and Y_S is shifted from a match on errors in the pre-period to a bias in the post-period and explains a large fraction of the deviation of the synthetic control treatment effect estimate from 0. We speculate that the bias that is not accounted for by SCAE is driven by the fact that model errors in the pre-treatment values for the treated unit leads to distorted weights on the donor units, as discussed above in section 3.1.

Figure 4b further clarifies the role of matching on error terms in causing bias. Here we plot the mean SCAE and bias by period for the simulations where the treated unit intercept rank is 29th out of 31. The matching algorithm tends to select control units with positive errors during the pre-period, distorting the counterfactual. Synthetic control average error then reverts to roughly mean zero in the post-period, transferring the pre-period SCAE into post-period bias. The small pre-period bias is driven by realizations with poor fit in the pre-period, and much like in Appendix Figure A.1, the pre-period bias disappears when we restrict our analysis to matches with relatively low MSE in the pre-period.

One possible concern is that the results we report here are dependent on the choice of distribution of μ_0. The specific shape of the bias across treatment rank does depend on the distribution from which unit-level intercepts are drawn. However the presence of bias exists for alternative distributions of μ_0. We show this in Figure A1, which replicates the bias results in Figure 4a, additionally including simulations that draw μ_0 from two other distributions (a uniform (0,1) distribution and a $\chi^2(1)$ distribution). While the exact shape of the bias function across rank of μ_0 depends on the distribution of μ_0's, the presence of bias persists across all three distributions we examine.



(a)



(b)

Figure 4: Bias and Synthetic Control Average Error (SCAE)

The top panel graphs mean bias and SCAE by treatment group rank of $\mu_{0,i}$. The bottom panel graphs mean bias and SCAE across periods, choosing $\mu_{0,i}$ to be rank 29th (of 30) in each simulation. Results are averaged over 3000 Monte Carlo simulations of the baseline DGP with no time trend (random intercepts only).

3.2.3 Trend Rank and Synthetic Trends

We next return to including unit-specific linear trends, to show how those impact the results of our Monte Carlo Simulations (as in Figure 1). The presence of unit-level trends in our simulations induces a second source of bias relative to the intercept-only DGP above - there can be a mismatch on levels, as in Figure 4a, and there can be a mismatch on trends if $\mu_{1,S}$ doesn't match $\mu_{1,i=treated}$. To focus on the bias arising from mismatched trends, for this exercise we fix $\mu_{0,i=treated}$ at the median of the realized donor pool.

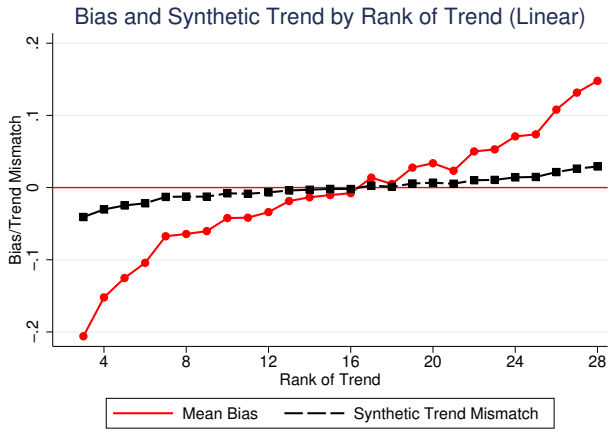
Figure 5 is similar to Figure 4, but here the red line plots the average bias across treated

unit *trend* ($\mu_{1,i=treated}$) rank. In this case the formula for the bias is: $Bias = \lambda_t \cdot E_\epsilon[\mu_{treated} - \sum_{i=1}^{30}(\mu_i \cdot \widehat{w_i})|rank(\mu_{1,i=treated}) = r])$. Again, there is little to no bias in the middle of the trend rank distribution, but lower ranked trend ranks are biased downwards and vice-versa. In this case, the bias is generated not as a level-mismatch (as above), but as a trend-mismatch with an increasing bias across the post-period, as we saw in Figure 1.

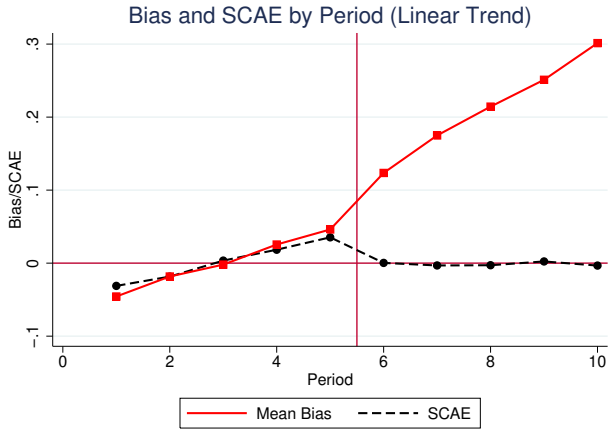
To demonstrate that this bias is the result of a mismatch between the treated unit's trend and the trend in the synthetic control group, for each Monte Carlo run we also calculate the "synthetic trend" $\mu_{1,S} = \sum_{i=1}^{G} w_i * \mu_{1,i}$. We then subtract $\mu_{1,S}$ from $\mu_{1,i=treated}$ to find the "trend bias", the mismatch between the true trend of the treatment group and the true trend of the synthetic control group. Mirroring Figure 4, the black line in Figure 5 graphs the synthetic trend bias ($\mu_{1,i=treated} - \mu_{1,S}$). When the treated unit trend rank is relatively low, the synthetic trend tends to be too high (the trend bias is negative), and vice-versa.

The mismatch on trends is itself a result of the algorithm matching on noise instead of structural parameters. Figure 5b shows the mean SCAE by period for the case where the treated unit trend rank is 29th. The matching algorithm in this case selects control units where the errors happen to trend upwards, and again mean errors revert to roughly zero in the post-treatment period. This type of bias becomes more severe (in level terms) longer into the post-treatment period, as the mismatched trends diverge further over time.

The previous figures demonstrate that a key determinant of bias, in the context of our simulations, is the degree to which the treated unit is in the tails of the distribution of the control units. In practice, the treated unit's structural parameters are unobserved, making our findings more difficult to implement; a policy analyst cannot directly assess whether this is likely to be a problem. She can, however, examine whether the observed realizations of the treated unit outcome variable are in the middle of the distribution of the control units, or whether they are toward the tails. If they are not in the middle, then there should be concern about bias from matching to noise.



(a)



(b)

Figure 5: Bias and Synthetic Linear Trend (D=1)

The top panel graphs mean bias and synthetic trend mismatch by rank of the treatment group. Rank is determined by the relative position of a group across the sum of each group's value of their intercept and trend ($\mu_{0,i} + \mu_{1,i}$ for each i). The bottom panel graphs mean bias and synthetic trend across periods, choosing the treated group to be ranked 29th (of 30) in each simulation. Results are averaged over 3000 Monte Carlo simulations of the baseline DGP with linear time trends.

4 Determinants of Magnitude of Bias

The preceding section generates two key insights regarding the data generating process and bias. First, the bias we describe is driven by matching to the error terms. The matching on noise comes at the expense of matching the structural features. In many settings such "over-fitting" does not cause bias. However, our example shows that in synthetic control estimation it can cause bias. This happens in finite samples when the donor units are not uniformly distributed around the treated

unit. The second insight is that the rank of unobserved parameters for the treatment unit relative to the distribution of control units affects the average bias. These two features - noise and asymmetry - are key determinants of the problem.

In this section we use Monte Carlo simulations to examine the impacts of additional features of the sample and DGP on the *magnitude* of the bias: DGP polynomial order (D), the number of potential donor units (G), and the number of pre-treatment time periods (T_0).

Up to this point we have restricted the DGP to include only random intercepts and/or unit-specific linear trends. In this section we extend this to consider higher-order polynomial trend functions. For a polynomial time trend function of degree D, we generate Y according to the DGP:

$$Y_{it} = \sum_{d=0}^{D} \lambda_{d,t} * \mu_{d,i} + \delta * Treated_{it} + \epsilon_{it} \tag{10}$$

with $\lambda_{d,t} = t^d$. Then $\mu := (\mu_{0,i}, \mu_{1,i}, ..., \mu_{D,i})$ is a vector interpreted as including (intercept, linear trend slope, ... , coefficient on D'th degree polynomial).

The choice of polynomial trends is motivated by the fact that most empirical settings using synthetic control have relatively smooth but non-linear pre-period outcome time-series. This is explicitly shown and discussed in Abadie et al. (2010) as one of the main benefits of synthetic controls in their application. Figure 1 in their paper graphs the highly non-linear trends in cigarette sales and the authors note how poorly these polynomial trends are captured by the consumption patterns of the rest of the United States: smoking follows a quadratic trend, increasing through the mid 1970s and then begins to decline. The synthetic control group is then shown to successfully capture these dynamics. Empirical work following Abadie et al. (2010) has likewise built off this insight and employed synthetic controls in cases where there are non-linear trends. We use polynomial-based time trends to generate such non-linear trends in our simulations, as higher order polynomials can flexibly mimic smooth trends across time. We note that our results are based on this modeling assumption, and this may limit the external validity of those results. Another limitation readers should note is that because in Eq. (10) what we condition on varies across simulation specifications, it is not clear that our bias is always comparable between simulations.

One computational complication with such polynomial trends is that the higher-order terms may dominate in the later periods and lead to extreme outcome values. To address this, we generate λ_t through a D-degree orthogonal polynomial transformation of the time period variable[10]. Additionally, we now set $T_0 = 10$ so that there is an adequately long pre-period for higher-order DGPs to reveal their underlying trends.

[10] We use the "orthpoly" command in Stata, defining the orthogonal polynomial transformations over the full time period.

For all simulations that follow, we draw each component of $\mu_{d,i}$ from an independent standard normal distribution $N(0,1)$, and ϵ_{it} is drawn from $N(0,0.2)$ unless otherwise specified (such as when we vary the error variance in Figure 6). We choose the treated unit by calculating the sum of $\mu_{d,i}$ for each unit individually and then selecting the unit with the 70th percentile of $\sum_{d=0}^{D} \mu_{d,i}$ as treated. We choose the 70th percentile so that the treated unit is some distance away from the center of the control unit distribution, but the algorithm is still typically able to find reasonable pre-treatment matches.

4.1 Noise (σ_ϵ)

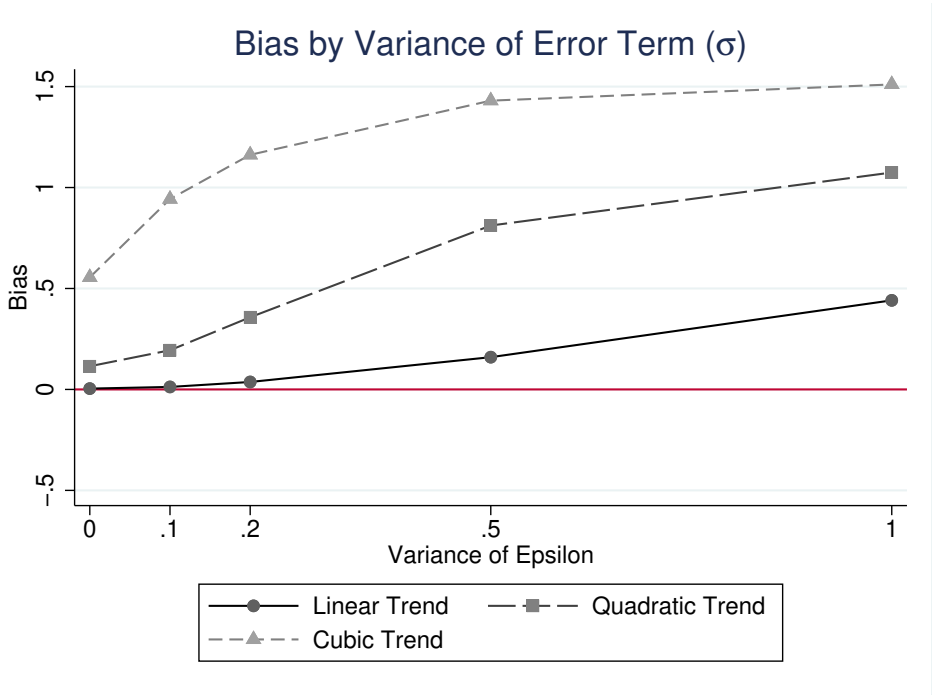


Figure 6: Bias by Noise

This figure shows bias estimates across variance of ϵ and by the trend polynomial degree (D). To select the treatment group we take the sum of the intercept and slope coefficients for each group and choose the treated group to be the one at the 70th percentile of that distribution. We specify the synthetic control estimator to match on all pre-period values of Y. The DGP includes 10 pre-periods and 30 donor groups. Results are averaged over 3000 Monte Carlo replications of the baseline DGP with varying D and σ.

Figure 3 showed simulation results when the DGP had an underlying linear trend ($D = 1$). In this section we expand the scope of the exercise and consider DGPs from Eq. 10 with D=1, 2 and 3 (linear, quadratic, and cubic trends). We graph the average bias in the final post-period (5 periods after treatment) across 4 variance levels of σ_ϵ (0, .5, 2, 5) separately for each degree polynomial.

Specifically: $Bias = \lambda_t \cdot E_\epsilon[\mu_{i=treated} - \sum_{i=1}^{30}(\mu_i \cdot \widehat{w_i})|Rank(\sum_{d=0}^{D}\mu_{d,i=treated}) = 0.7]$; where the exact value of D changes across simulations.

The results are shown in Figure 6. Bias increases with σ_ϵ, though the marginal impact of additional variance diminishes at higher values. In addition, the bias increases in D (the degree of the highest polynomial trend term). These are basic themes of the subsequent exercises as well; increases in D and σ_ϵ, tend to increase bias[11].

While our primary concern in this paper is bias, the factors that influence bias can also impact the statistical uncertainty of the estimates. We also examine the sampling variability for the synthetic control estimates, and report these results in Figure A2. The top panel of Figure A2 replicates Figure 6, reporting on the standard deviation of estimates in the final post-period. Results are as expected: increasing the variance of the error term increases sampling variability, as does the increasing the dimensionality of the unobserved parameters.

4.2 Size of Donor Pool (G)

One candidate approach to mitigate the bias given the assumed DGP is to increase the number of potential matches, that is, the number of donor units, G. This might increase the chance of matching on a structurally similar set of units and decrease the likelihood of matching on noise.

[11] Note that there is a non-zero bias even with zero error terms for the higher-order DGPs. This occurs because for more complex DGPs, in some cases there is no perfect match available. This is again similar to the issue discussed in Appendix A3.

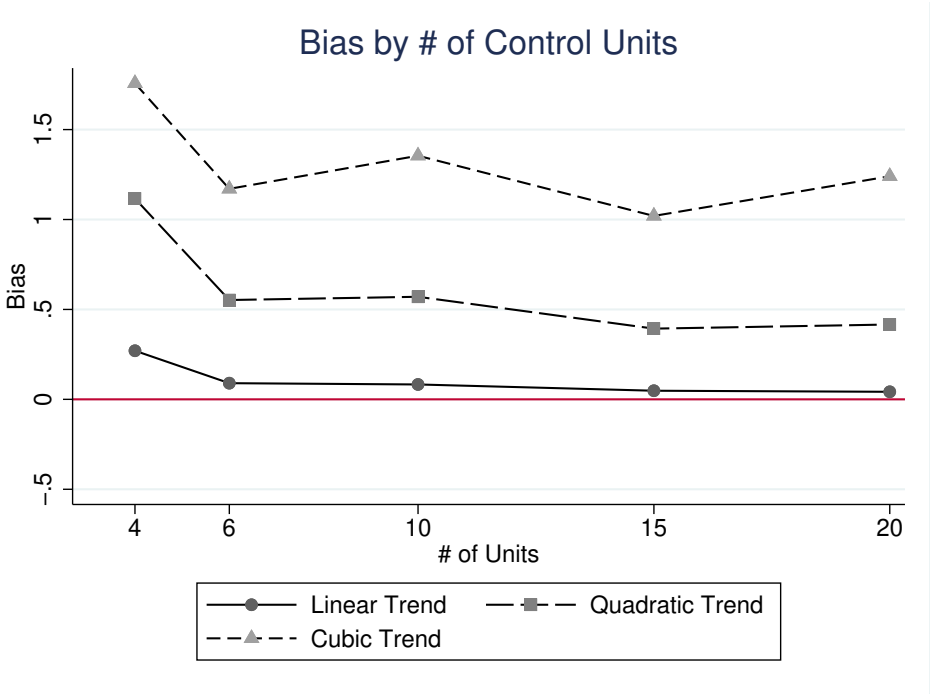


Figure 7: Bias and Size of Donor Pool

This figure shows mean bias estimates across number of pre-periods (T_0) and the trend polynomial degree (D). To select the treatment group we take the sum of the intercept and slope coefficients for each group and choose the treated group to be the one at the 70th percentile of that distribution. We specify synthetic controls to match on all pre-period values of Y. The DGP includes 10 pre-periods and a varying number of donor groups. Results are averaged over 3000 Monte Carlo replications of the baseline DGP with varying D and σ.

This turns out to be less effective at reducing bias than we had expected. Figure 7 graphs mean bias across the number of donor units, separately by D. While increasing the number of groups can reduce bias when there are very few units, in general the gains beyond 10 or so units are modest if they exist at all. Higher dimensionality of D dwarfs the impact of increased donor pool size.

We again provide estimates of sampling variability across the number of control units in the middle panel of Appendix Figure A2. As would be expected, increasing the number of control units decreases the sampling variability of the estimates. However, this effect is relatively modest, and the dimensionality of unobserved parameters again largely dominates the number of control units in terms of affecting estimate variability.

4.3 Pre-Period Length (T_0)

We next consider the role of the number of pre-treatment periods, T_0. Figure 8 graphs mean bias in the final-post period against the number of pre-periods. In the Monte Carlo simulation, the bias decreases as the length of the pre-period panel increases for all values of D. This result reflects the

fact that with longer matching periods the importance of the trend parameters increases relative to idiosyncratic error terms, so that matching on outcomes yields a better match to the underlying trend function. When (in terms of pre-period length) the gains are realized depends on the the dimensionality of the DGP (D). The relative bias relative reduction going from 10 to 25 periods with a cubic DGP (D=3) is smaller, percentage-wise, than going from 4 to 6 periods with a linear DGP (D=1).

As above, we also provide estimates of sampling variability in the Appendix Figure A2. Following Figure 8, we graph the standard deviation of estimates across the number of pre-periods for each values of dimensionality D in the bottom panel of Appendix Figure A2. Increasing the number of pre-periods leads to lower sampling variability, but only after sufficient pre-periods to identify the polynomial, which depends on D.

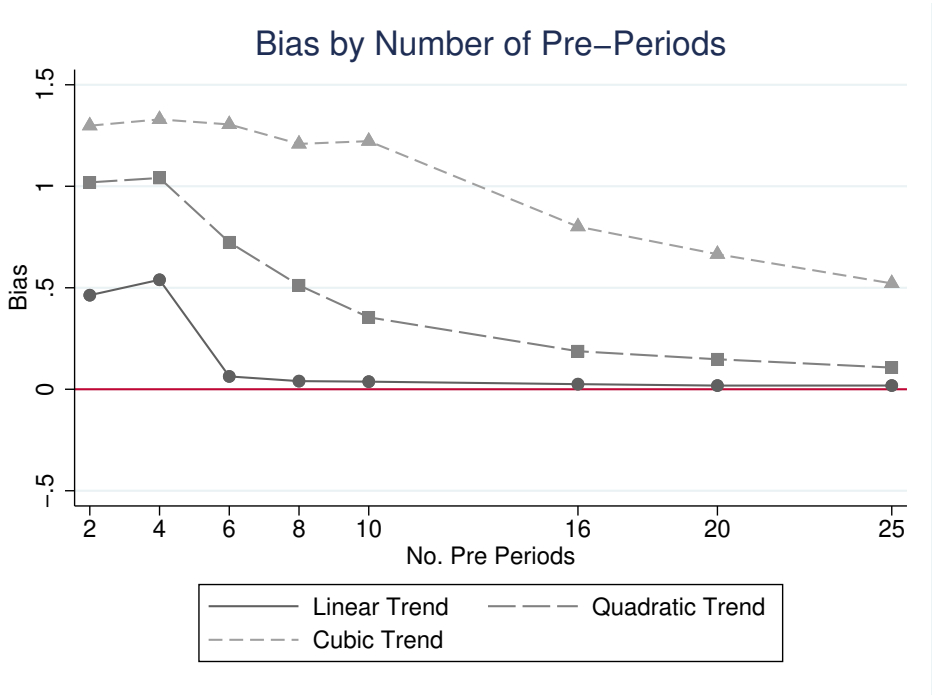


Figure 8: Bias by Pre-Period Length

This figures shows mean bias estimates across number of control units (G) and trend (D). We specify synthetic controls to match on all pre-period values of Y. The DGP includes different time trends, varying numbers of pre-periods, and 30 donor groups. Results are averaged over 3000 Monte Carlo replications.

5 Bias Reduction Strategy

The bias that we identify in our Monte Carlo simulations is the result of a type of "over-fitting" of the weights to match the model errors (noise). The bias is stable conditional on the particular

location of the treated unit in the joint distribution of unobserved parameters. Given assumptions about the underlying DGP, researchers can estimate this location and perform a simulation-based approach to estimating the bias.

Using a parametric bootstrap, we simulate the magnitude of bias that results from matching on noise given the observable features of the data, including the relative location of the treatment group parameters in the joint distribution of donor unit parameters. We then subtract this estimated bias from the main estimate. This approach involves: estimating the underlying structural model; simulating realizations of the synthetic control estimate when the treatment group is chosen to be similar to the one in the original data relative to the realized sample; and then averaging the treatment effect estimate across many simulation runs to estimate the bias in the original analysis[12].

In this section we lay out the procedure for the proposed correction and demonstrate an empirically feasible implementation that reduces bias relative to matching on all pre-period values of Y.

5.1 Procedure for Proposed Bias Correction

The correction procedure is as follows:

1. We assume the the DGP comes from the polynomial family of functions. This requires choosing or estimating the polynomial order (D). To estimate $\hat{D}$, we estimate a series of random-intercepts and random-slopes models on the underlying panel from polynomial degree 0 to 4. We then use the Bayesian information criteria (BIC) as a model selection procedure, choosing the polynomial order with lowest BIC[13].

2. After $\hat{D}$ is chosen, estimate the mean and standard deviation of each unobserved factor loading μ_d (we assume that each is generated from a normal distribution, though other distributions are theoretically viable) using a GMM mixed-effects model.[14] The same model is used to estimate $\hat{\sigma}$, the standard deviation of the error term.

3. Generate artificial data for the control units by randomly drawing from the assumed DGP, using the estimated hyper-parameter values and setting G, T_0 and T as in the sample.

[12] Frölich (2004), Huber et al. (2013), and Busso et al. (2014), establish the potential role of empirical Monte Carlo simulations in helping inform estimator choice for researchers. A major take away from these papers is that the assumed DGP needs to match the applied data environment in order for the simulations to be useful. Recently Advani et al. (2019) evaluate Monte Calro approaches for choosing among alternative potential estimators. They conclude that this method can perform poorly at identifying the bias of estimators. Our setting differs from that studied in Advani et al. (2019) because our bias (driven by selecting on noise) is a type of "mechanical bias", rather than being driven by covariate imbalance or other finite-sample violations of a selection on observables assumption. As our simulations below show, the bias-correction approach we use is able to measure and correct for this bias.

[13] A researcher might well choose other model selection techniques.

[14] We use Stata's xtmixed command with the outcome as the dependent variable and D-degree orthogonal polynomial transformations of the period variable as independent variables.

4. Create a simulated treatment unit. Using the actual (non-simulated) data, estimate each unit's μ parameter values with a linear regression conditional on the chosen $\hat{D}$:

$$Y_{it} = \sum_{d=0}^{\hat{D}} \lambda_{d,t} * \mu_{d,i} + \eta_{it} \tag{11}$$

   Then assign the treated unit the values of μ with the corresponding rank from the simulated parameters generated in step #2. Then generate treated unit outcomes according to EQ. 11, drawing $\eta_{it} \sim N(0, \hat{\sigma}^2)$.

5. Run synthetic control estimation using the same matching specification as used for the real data (in our simulations below we match on all pre-treatment Ys, but again other specifications are possible), and record the estimated treatment effect.

6. Repeat steps 3-5 many times and compute the average treatment effect[15]. This is an estimate of the bias.

7. Subtract the estimated bias from the treatment effect estimated from the actual data, resulting in a bias-corrected estimate.

We show below that this approach can reduce or eliminate the bias from matching on noise. This approach is limited in that the researcher assumes both the functional form of the DGP and the distribution type (normal, uniform, etc.) of the unobserved parameters. This can be seen by examining Appendix Figure A1. If we assume, for instance, a uniform distribution of μ_0 but the actual distribution is normal, then the difference between the gray triangles (uniform) and red circles (normal) would be the degree of inaccuracy in the correction.

5.2 Implementation of Bias Correction

The simulations in this section continue to use the parameter settings from previous simulations, and we simulate polynomial trend functions of degrees D=1 to 3. For each degree polynomial we estimate a specification that matches on all pre-period values of the outcome variable, and one that applies the simulation-based correction. We then graph, separately for each D, the mean difference between the treatment group and the synthetic control group (the bias) across periods. For each

[15] For each "nested" simulation described in these steps, as with all other analysis in this paper, we exclude data realizations where the treated unit has any pre-treatment outcome observations that are the largest or smallest of all units. In rare cases within our simulations, when $\hat{D}$ is mis-specified the treated unit can receive extreme estimated parameter values, so that nearly every simulated data realization is rejected because treated unit outcomes are very large or small. We therefore use the following rule: if out of the first 20 nested simulations, four or fewer are successful (i.e. the treated unit never has the largest or smallest outcome in a pre-treatment period), we terminate the correction procedure and use the original baseline estimate as our "corrected" estimate.

D, we run 3000 simulations, each with 100 "nested" simulation runs used to estimate bias with the Monte Carlo bias correction procedure.

We consider empirical estimation of D when the true DGP is known to come from the family of polynomial trend functions from degree 0 to 4. We run a series of Monte Carlo simulations where the correction procedure is based on using $\hat{D}$ and moments of the empirical distribution of outcomes instead of (the empirically unobservable) D itself. We also provide results on estimate precision.

Figure 9 shows how accurately the BIC estimate identifies the polynomial order of the DGP ($\hat{D}$). The histogram shows the distribution of $\hat{D}$ under three true values of D (linear, quadratic, and cubic trends) and the parameter settings used in Figure 10. In each case, the BIC chooses the correct degree of heterogeneity more than half the time, increasing from about 50% in the linear case, to almost 100% in the case of a cubic polynomial. In every case, BIC chooses $\hat{D} \geq D$.

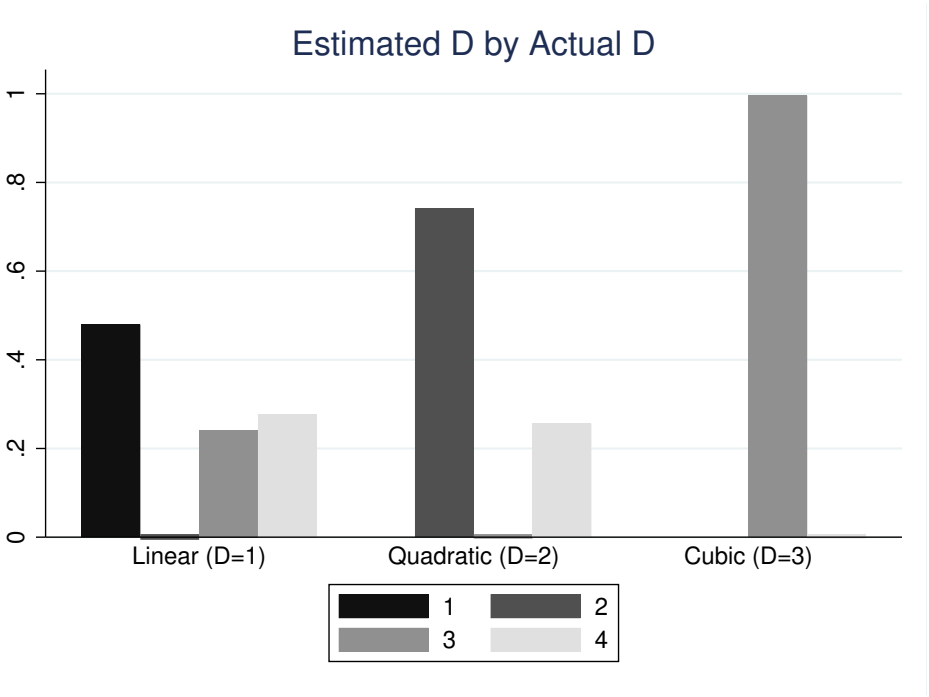


Figure 9: Estimates of D Given Candidate Family of Polynomial Functions

The bars indicate the proportion of the time $\hat{D}$ is selected given the true D (x-axis). $\hat{D}$ is selected via first running a series of regression models with D-degree polynomial trend parameters. One regression is run for each D (ranging between 1 and 4) and the model with the lowest BIC is selected. Results are over 3,000 Monte Carlo replications of baseline DGP across each (true) dimension of D.

5.3 Performance of Bias Correction

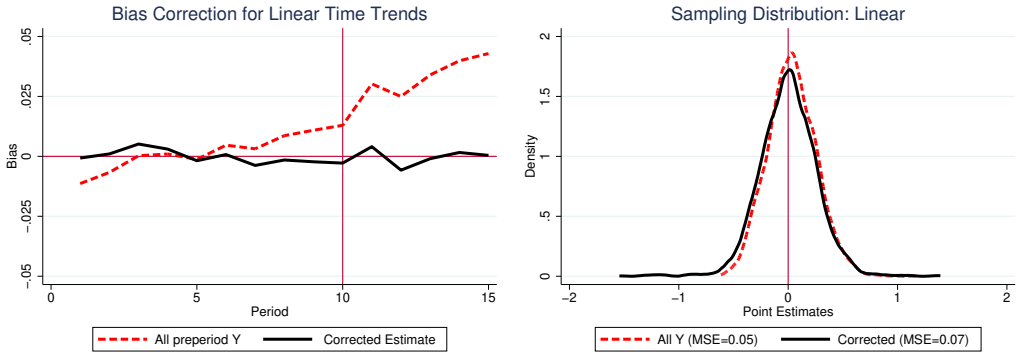
We conduct a new series of simulation exercises to demonstrate the effectiveness of our bias reduction strategy. We use the parameter settings from previous simulations, and we simulate polynomial trend

functions of degrees D=1 to 3. For each degree polynomial we estimate a specification that matches on all pre-period values of the outcome variable, and one that applies the Monte Carlo correction discussed above. We then graph, separately for each D, the mean difference between the treatment group and the synthetic control group (the bias) across periods. For each D, we run 3000 simulations, each with 100 "nested" simulation runs of the correction used to estimate bias within each overall simulation run.

Figure 10 shows the results of this exercise. Our Monte Carlos bias correction strategy substantially reduces bias relative to the baseline of matching on all pre-period outcome variables across DGPs of different polynomial orders. Further, the correction is nearly unbiased through true DGPs of degree 2.

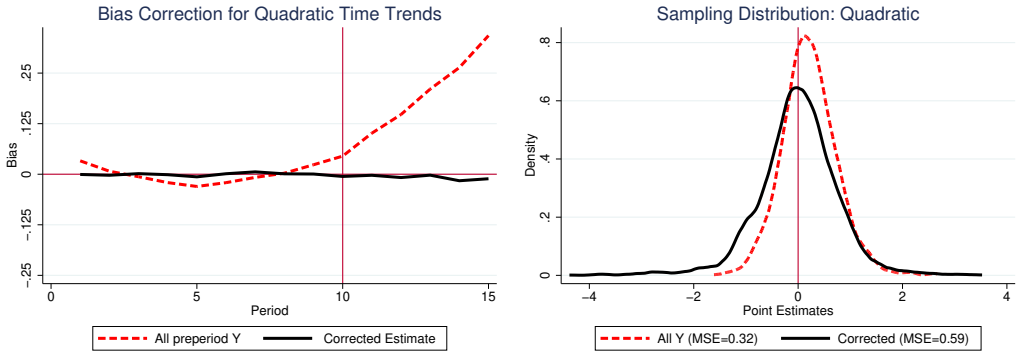
A number of lessons discussed previously continue to hold. For the specification that matches on all pre-period outcome values (the red short-dashed line), increasing D (moving down panels of Figure 10) leads to increased bias. The pre-period fit is on average quite good. The bias increases over time following a functional form determined by the curvature in the underlying trend, due to the mismatch between $\mu_{i=treated}$ and μ_S.

We additionally include density plots of the sampling distributions of the estimates and their MSE in the right side of Figure 10. Here we graph, for each D, a kernel density estimate of the sampling distribution of $\hat{\delta}$ from each specification. This shows that while the correction procedures do indeed reduce bias, they also increase sampling variance. None of the corrected estimates have an MSE below that achieved from matching on all pre-period Y values at any dimensionality D. While the corrections do continue to reduce bias, the cost in terms of variance is sufficiently high that their MSE is higher than the more biased estimates from matching on all pre-period outcomes.



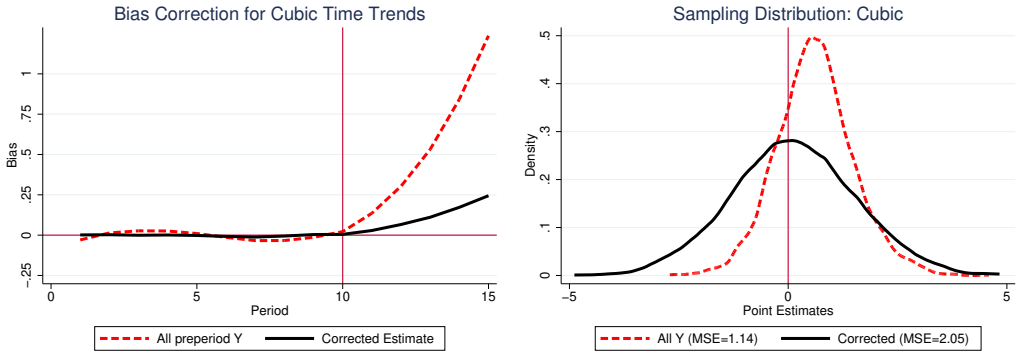
(a) D=1

(b) Linear



(c) D=2

(d) Quadratic



(e) D=3

(f) Cubic

Figure 10: Performance of Bias Correction Procedures when DGP is unknown (Feasible)

This figure graphs bias and corrected estimates over 3,000 Monte Carlo simulations of the baseline DGP, while varying the trends (D). We take the sum of the intercept and slope coefficients for each group and then select the treatment group to be the one at the 70th percentile of that distribution. Rows denote trend (linear, quadratic, cubic). The left column graphs mean bias by trend (D) across periods. The right column provides estimates of the density and reports the true mean squared error.

6 Empirical Application: Free Trade and Economic Growth

One important question for empirical researchers is the extent to which the bias we describe in this paper should concern practitioners. That is, can it be sufficiently large to meaningfully alter substantive conclusions about the real world? In this section, we show that there can be scope for meaningful "matching on noise" bias in the context of a real world empirical example.

6.1 Empirical Example: Trade Liberalization and GDP per capita

Billmeier and Nannicini (2013) (henceforth BN) use a series of synthetic control models to analyze the effect of national-level free trade agreements on economic growth. They first identify every country that experienced a significant trade liberalization episode between 1963 and 2005, and perform synthetic control estimates for each liberalizing country for which basic data requirements were met[16]. For each episode of trade liberalization, they performed a separate synthetic control estimation, using other countries in the same region that did not liberalize trade during the sample window as the donor units.

We use replication data from BN to evaluate possible matching on noise bias using our proposed bootstrap correction procedures. To fit the estimation framework in our paper, we differ from BN in that we match on all available pre-treatment outcome values rather than the set of variables they use. Further, we use ln(GDP/capita) rather than levels of GDP/capita to maintain comparability of estimated bias magnitudes across countries. Thus we do not consider this work an empirical replication nor our corrected point estimates as directly informing the literature on trade liberalization. The exercise is instead meant to investigate the scope for bias using real-world data in a meaningful context.

Our application focuses on the BN estimates that use only nearby countries as synthetic control donor units, and thus we drop all countries in the Middle-East and North Africa regions, which use alternative donor pools. We also limit our sample to countries with 5 or more periods of data both before and after the free trade agreement and good quality fits for our bootstrap procedure as described in section 5.2. This leaves us with 17 countries from the original 30 countries in the BN sample.

For each country in the remaining replication sample, we estimate two synthetic controls models, using only data from the 5 periods before and after the trade liberalization. We estimate a baseline synthetic control model matching only on all pre-period Y values and our feasible Monte Carlo correction.

[16] BN use World Bank GDP per capita data as an outcome.

6.2 Empirical Results

The results are summarized in Figure 11. The top panel shows the Monte Carlo correction bias estimates. The Monte Carlo produces smooth estimates across time, given the correction works off averaging across polynomial time-trend simulations. Estimates largely fan out from zero, following the cubic trend mismatch when choosing a simulated-treatment group at the appropriate part of the unobserved variable distribution. The magnitudes of the estimated bias in the post period are meaningfully large, ranging from +/- 25% of GDP by three years after liberalization.

The bottom panel shows the difference between the baseline estimate and the feasible correction estimate (that is, the estimated bias). The BIC criteria chose D=3 (a cubic model for trends), the largest value of D among the choice set. The scale of the y-axis is log points of GDP per capita. The x-axis shows the baseline synthetic control estimate of the impact of trade liberalization after 5 years, in log points of GDP per capita, and the y-axis shows the corrected estimates. A 45 degree line is plotted to give context for the magnitude of the corrections. The panel graphs corrected estimates across the baseline estimate of matching on all pre-period Y.

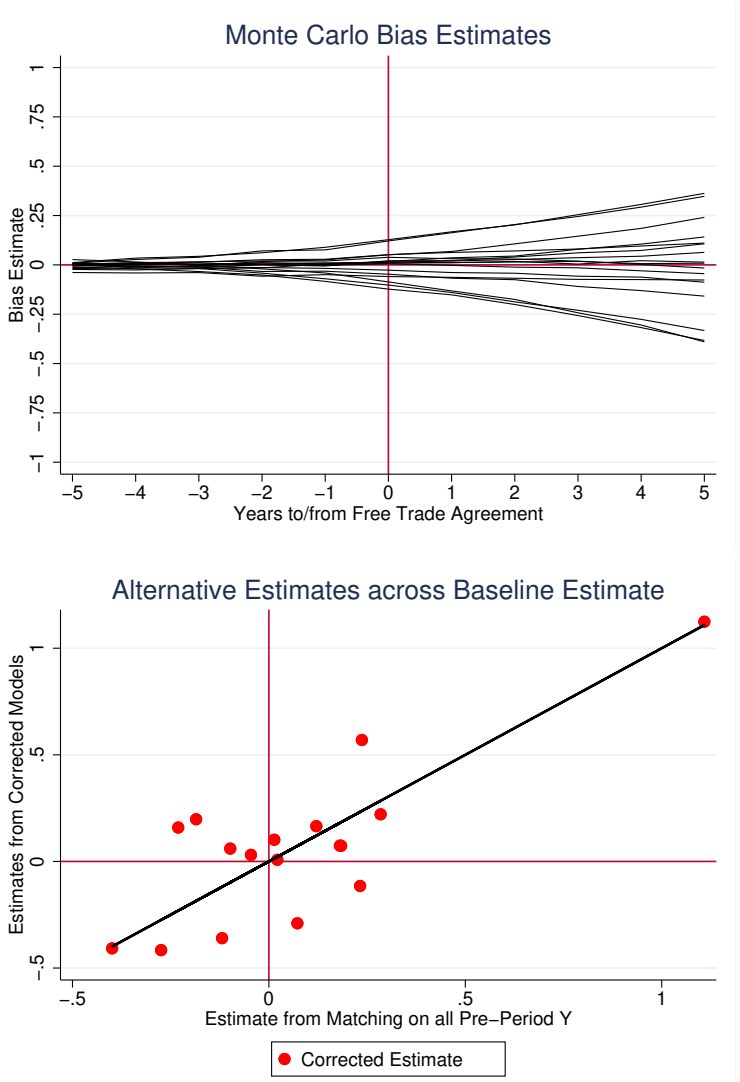


Figure 11: Bias Estimates in Free Trade Application

The top panel graphs mean bias estimates from the feasible bootstrap correction from 17 synthetic controls estimates of the effects of Free Trade on (ln) $GDP per capita, following Billmeier and Nannicini (2013). The bottom panel provides a scatter plot of corrected estimates across baseline estimates (matching on all Pre-Period Y).

In the overall replication sample of 17 countries, the median absolute value of bias estimates is 0.11 log points, approximately 11% of GDP per capita. However, this relatively moderate bias conceals potentially large bias in several studies: the 75th percentile of absolute estimated bias is 0.33 log points. Fully 35% of Monte Carlo corrected estimates are of the opposite sign from the baseline estimates.

This disagreement on bias estimates within-country can further be seen in Appendix Figure A5. Here we graph country-specific estimates of both synthetic control models across year for 9 of the 17 countries in the estimation sample. The figures are ordered from the lowest estimated baseline treatment effect in the upper left to the largest in the bottom right. The baseline estimate is in dashed red, while the solid black lines represents the feasible Monte Carlo estimates. Of the examples in Appendix Figure A5, two of the nine countries have estimates that disagree on sign, and three additional countries have estimated magnitudes that disagree considerably.

Taken together, these exercises demonstrate that the bias that results from "matching on noise" can be meaningfully large in a real world empirical application.

7 An Alternative to Synthetic Control: Interactive Fixed-Effects

The factor structure justifying synthetic control estimation in Abadie et al. (2010) suggests an alternative estimation approach[17]. The alternative is to directly estimate the individual factor loadings (μ_i) for each control group, using the interactive fixed effects approach of Bai (2009). If control units have differential polynomial trends, rather than weighting those groups to approximate the trends of the treated group: we can directly estimate those trends. The interactive fixed effects approach of Bai (2009) allows us to do this by providing a way of identifying the factor loadings by separately estimating the factors (λ_t), and factor loadings (μ_i) simultaneously.

We employ this technique, estimating the interactive fixed effect in each draw of our Monte Carlo simulation. Our setup is otherwise identical to our evaluation of the synthetic control estimates above, with 10 pre-treatment periods, 30 donor groups and normally distributed draws of μ[18]. This approach allows us to horse race the two methods and compare (in the context of our simulations) the bias of synthetic control models to that of interactive fixed effects.

We estimate the following interactive fixed effects model:

[17] We thank the Associate Editor for this suggestion.

[18] Code to estimate interactive fixed effects was provided by Matthieu Gomez, Erik Loualiche, and Dave Kleinschmidt online at: https://github.com/FixedEffects/InteractiveFixedEffectModels.jl

$$Y_{it} = \sum_{t=10}^{15} \gamma_t Treated_i + \sum_{r=1}^{R} \lambda_{r,t}\mu_{r,i} + \lambda_{0,t} + \mu_{0,i} + \epsilon_{it} \tag{12}$$

Where $\sum_{t=10}^{15} \gamma_t Treated_i$ is a vector of treated status by time dummies for each post treatment time period, and γ_t captures the treatment effect in time t. $\mu_{0,i}$ and $\lambda_{0,t}$ represent standard unit and time fixed effects and are always included in addition to the interactive fixed effects. The interactive fixed effects themselves are captured by: $\sum_{r=1}^{R} \lambda_{r,t}\mu_{r,i}$ which reflect R distinct factors, where R needs to be chosen in the process of estimation. In practice we will vary R to be from zero (a standard two way fixed effects model) to $R = 2$ (allowing for two additional latent factors). Note that $\lambda_{r,t}\mu_{r,i}$ allows for a different unit-time specific effect in each period. If the model is specified correctly, γ_t will capture the treatment effect for each time period post treatment.

Figure 12 presents bias and dispersion estimates for interactive two-way fixed-effects models instead of the synthetic control models. The left column graphs bias and the right-column graphs the corresponding sampling distribution of post-period bias across units. Rows indicate the degree of the true underlying polynomial time trends, D: 1) linear; 2) quadratic; and 3) cubic. For each degree of DGP, we estimate three models: a standard two-way fixed-effects model with the treatment effect enforced to be zero in each pre-period, and interactive two-way fixed-effect models with 1 and 2 latent factors (R) specified. We run all three estimations 3000 times for each treated unit rank, and record the average bias.

When there are linear trends ($D = 1$), the standard two-way fixed effects model has bias that increases over time in the post period. This is expected because the true differential trends across units are not controlled for. Interactive fixed effects, allowing the common factor to have a single rank, eliminates the bias completely. When there are quadratic trends ($D = 2$, first column second row), allowing a rank of $R = 2$ for the common factor reduces the bias to zero (with a small decrease in the bias if $R = 1$). Strikingly, as long as the model is not under-specified, there appears to be no bias from matching on noise as we found in our main Monte Carlo simulations for synthetic control methods. Focusing on the right column, the trade off for including too many latent parameters in the model is apparently a relatively small increase in variance of the underlying estimates. Further, the sign and magnitude of the bias from interactive fixed-effects models are similar to those of the synthetic controls estimates, but only when the model under-specifies the number of unobserved latent parameters. While the primary focus of this paper is not on interactive fixed effects, we believe these results suggest this method has promise as a potential competitor or supplement to synthetic controls. Future work could further explore the relative benefits and drawbacks of the two methods in more detail.

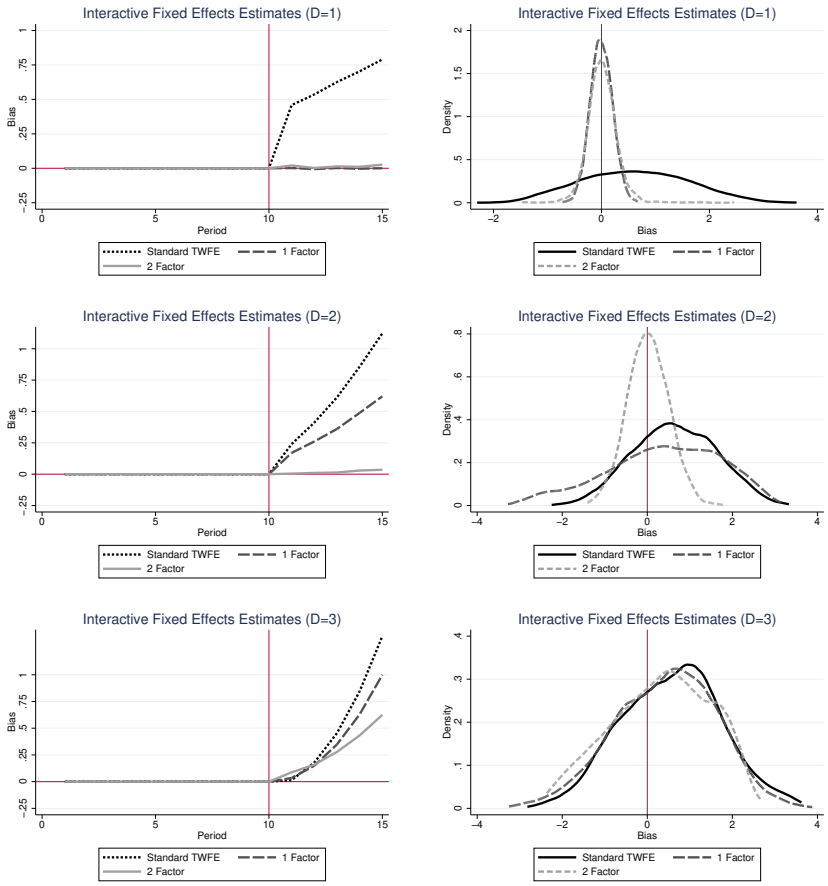


Figure 12: Interactive FE Results

The figure graphs mean bias estimates for interactive fixed-effect models over 3,000 Monte Carlo simulations of the baseline DGP while varying the trends (D). We take the sum of the intercept and slope coefficients for each group and then select the treatment group to be the one at the 70th percentile of that distribution. Rows denote trend (linear, quadratic, cubic). The left column graphs mean bias by trend (D) across periods. The right column provides estimates of the density and reports the true mean squared error, truncated at the 99th percentile. The short-dashed line represents standard two way fixed-effects (TWFE) estimates, while the long-dash and solid line estimates come from interactive fixed-effect models with 1 and 2 unobserved factors respectively, and also saturated with group and time fixed effects, and treatment by time dummies for each post period.

8 Conclusion

Using a range of Monte Carlo simulations and a real world empirical example, we document bias in synthetic controls estimates stemming from matching on idiosyncratic noise in the outcome variable. Within the context of our assumed DGP, the magnitude of the bias is a function of the variance of model errors of the outcome variable, the complexity of the underlying DGP, and features of the data sample including number of potential donor units and length of pre-period. In a real world application, we demonstrate that the bias can be substantial.

We then present a simulation-based correction that directly estimates the bias in a feasibily implementable manner. The correction can lead to bias improvements under a range of settings for our polynomial-trend family of DGPs. However, the correction generates higher MSE in the particular environments we simulate. Further, it requires knowledge of the parametric form of the unobserved factors (in our setting, polynomial trends), which are not typically available to a researcher.

Finally, we present interactive fixed effects as an alternative to synthetic controls. While fully exploring the relative benefits and costs of interactive fixed effects compared to synthetic controls is beyond the scope of the paper, our simulations suggest that interactive fixed effects do not produce the same bias from matching on noise present in synthetic controls.

In sum, our paper sheds light on a poorly understood problem with an increasingly popular program evaluation method. Our proposed correction procedures point towards future research possibilities and refinements. Finally, we hope researchers employing synthetic control methods to evaluate real world programs will employ similarly-motivated models and specifications in their work to convince themselves and readers that their matches are on meaningful features of the world, and not on noise.

References

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the 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.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the basque country. *American economic review 93*(1), 113–132.

Acemoglu, D., S. Johnson, A. Kermani, J. Kwak, and T. Mitton (2016). The value of connections in turbulent times: Evidence from the united states. *Journal of Financial Economics 121*(2), 368–391.

Advani, A., T. Kitagawa, and T. Słoczyński (2019). Mostly harmless simulations? using monte carlo studies for estimator selection. *Journal of Applied Econometrics 34*(6), 893–910.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021). Synthetic difference-in-differences. *American Economic Review 111*(12), 4088–4118.

Athey, S. and G. W. Imbens (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives 31*(2), 3–32.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

Ben-Michael, E., A. Feller, and J. Rothstein (2021). The augmented synthetic control method. *Journal of the American Statistical Association 116*(536), 1789–1803.

Billmeier, A. and T. Nannicini (2013). Assessing economic liberalization episodes: A synthetic control approach. *Review of Economics and Statistics 95*(3), 983–1001.

Bohn, S., M. Lofstrom, and S. Raphael (2014). Did the 2007 legal arizona workers act reduce the state's unauthorized immigrant population? *Review of Economics and Statistics 96*(2), 258–269.

Busso, M., J. DiNardo, and J. McCrary (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics 96*(5), 885–897.

Cavallo, E., S. Galiani, I. Noy, and J. Pantano (2013). Catastrophic natural disasters and economic growth. *Review of Economics and Statistics 95*(5), 1549–1561.

Chen, A., A. Graves, M. Resnick, and R. Michael (2018). Does spending more get more? health care delivery and fiscal implications from a medicare fee bump. *Journal of Policy Analysis and Management 37*(4), 706–731.

Courtemanche, C. J. and D. Zapata (2014). Does universal coverage improve health? the massachusetts experience. *Journal of Policy Analysis and Management 33*(1), 36–69.

Cunningham, S. and M. Shah (2017, 12). Decriminalizing Indoor Prostitution: Implications for Sexual Violence and Public Health. *The Review of Economic Studies 85*(3), 1683–1715.

Donohue, J. J., A. Aneja, and K. D. Weber (2019). Right-to-carry laws and violent crime: A comprehensive assessment using panel data and a state-level synthetic control analysis. *Journal of Empirical Legal Studies 16*(2), 198–247.

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

Eren, O. and I. S. Ozbeklik (2016). What do right-to-work laws do? evidence from a synthetic control method analysis. *Journal of Policy Analysis and Management 35*(1), 173–194.

Ferman, B. and C. Pinto (2016). Revisiting the synthetic control estimator.

Ferman, B. and C. Pinto (2017). Placebo tests for synthetic controls. Technical report, University Library of Munich, Germany.

Ferman, B. and C. Pinto (2021). Synthetic controls with imperfect pretreatment fit. *Quantitative Economics 12*(4), 1197–1221.

Ferman, B., C. Pinto, and V. Possebom (2020). Cherry picking with synthetic controls. *Journal of Policy Analysis and Management 39*(2), 510–532.

Fitzpatrick, M. D. (2008). Starting school at four: The effect of universal pre-kindergarten on children's academic achievement. *The BE Journal of Economic Analysis & Policy 8*(1).

Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics 86*(1), 77–90.

Gautier, P. A., A. Siegmann, and A. Van Vuuren (2009). Terrorism and attitudes towards minorities: The effect of the theo van gogh murder on house prices in amsterdam. *Journal of Urban Economics 65*(2), 113–126.

Gurantz, O. (2020). What does free community college buy? early impacts from the oregon promise. *Journal of Policy Analysis and Management 39*(1), 11–35.

Hinrichs, P. (2014). Affirmative action bans and college graduation rates. *Economics of Education Review 42*, 43–52.

Hu, L., R. Kaestner, B. Mazumder, S. Miller, and A. Wong (2018). The effect of the affordable care act medicaid expansions on financial wellbeing. *Journal of public economics 163*, 99–112.

Huber, M., M. Lechner, and C. Wunsch (2013). The performance of estimators based on the propensity score. *Journal of Econometrics 175*(1), 1–21.

Kaul, A., S. Klößner, G. Pfeifer, and M. Schieler (2022). Standard synthetic control methods: The case of using all preintervention outcomes together with covariates. *Journal of Business & Economic Statistics 40*(3), 1362–1376.

Kiesel, K. and S. B. Villas-Boas (2013). Can information costs affect consumer choice? nutritional labels in a supermarket experiment. *International Journal of Industrial Organization 31*(2), 153–163.

Klasik, D. (2013). The act of enrollment: The college enrollment effects of state-required college entrance exam testing. *Educational researcher 42*(3), 151–160.

Kreif, N., R. Grieve, D. Hangartner, A. J. Turner, S. Nikolova, and M. Sutton (2016). Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health economics 25*(12), 1514–1528.

Lee, W.-S. (2010). Comparative case studies of the effects of inflation targeting in emerging economies. *Oxford Economic Papers 63*(2), 375–397.

Nannicini, T. and R. Ricciuti (2010). Autocratic transitions and growth.

Peri, G. and V. Yasenov (2019). The labor market effects of a refugee wave synthetic control method meets the mariel boatlift. *Journal of Human Resources 54*(2), 267–309.

Pinotti, P. (2015). The economic costs of organised crime: Evidence from southern italy. *The Economic Journal 125*(586), F203–F232.

Smith, B. (2013). *Cross-Country Determinants of Growth: A Microeconometric Approach*. University of California, Davis.

Smith, B. (2015). The resource curse exorcised: Evidence from a panel of countries. *Journal of Development Economics 116*, 57–73.

Sun, J., F. Wang, H. Yin, and B. Zhang (2019). Money talks: The environmental impact of china's green credit policy. *Journal of Policy Analysis and Management 38*(3), 653–680.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis 25*(1), 57–76.

Online appendix for "Matching on Noise: Finite Sample Bias in the Synthetic Control Estimator"

A Calculating the Bias in the Illustrative model

In this section we derive the bias in the estimated treatment effect for the model of Section 3.1 of the paper. We consider a scenario with only three units, in which the treated unit lies in between the two donor units. The only unobserved factor loading is the fixed effect for each unit. $\mu_0 = 0$ is the location of the fixed effect for the lower control unit, $\mu_1 = 1$ is the location of the fixed effect for the upper control unit, and $\mu_{treated}$ is the location of the fixed effect for the treated unit. $\mu_{treated}$ is a parameter that can vary, and the bias in the estimated treatment effect will depend on this parameter. There are two time periods. The first period is used to estimate SC weights, and the second period is used to estimate the treatment effect. The realized outcomes $Y_{i,t}$ are the unit fixed effects, plus iid errors ϵ. For the treated unit in time 2, the realized outcome also includes the treatment effect δ.

A.1 Bias in the treatment effect estimate

We consider the factor loadings $(\mu_0, \mu_1, \mu_{treated})$ as fixed, and we condition on these when we compute the bias. We consider the model errors (ϵ's) to be random. The bias is the expected error in the estimated treatment effect, with expectation taken over the distribution of the ϵ's

The SC weights on the donor units come from fitting the first period realization. There is a weight w_0 for unit 0 and a weight w_1 for unit 1. The ideal "target" weights depend on the true locations of the factor loadings:

$$w_0^{tgt} = \frac{\mu_1 - \mu_{treated}}{\mu_1 - \mu_0}$$

$$w_1^{tgt} = \frac{\mu_{treated} - \mu_0}{\mu_1 - \mu_0}$$

The ideal SC weights will recover the true factor loading for the treated unit:

$$w_0^{tgt} \cdot \mu_0 + w_1^{tgt} \cdot \mu_1 = \mu_{treated}$$

The estimated SC weights depend on the realized locations in period 1. In the formulas below the model errors ϵ refer to first period errors only.

$$\widehat{w_0} = \frac{\mu_1 - \mu_{treated} + \epsilon_1 - \epsilon_{treated}}{\mu_1 - \mu_0 + \epsilon_1 - \epsilon_0}$$

$$\widehat{w_1} = \frac{\mu_{treated} - \mu_0 + \epsilon_{treated} - \epsilon_0}{\mu_1 - \mu_0 + \epsilon_1 - \epsilon_0}$$

The Period 2 estimated treatment effect is:

$$\begin{aligned}\widehat{\delta} &= Y_{treated,t=2} - Y_{SC,t=2} \\ &= \delta + [\mu_{treated} - \widehat{w_0}\mu_0 - \widehat{w_1}\mu_1] + [\epsilon_{treated,t=2} - \widehat{w_0}\epsilon_{0,t=2} - \widehat{w_1}\epsilon_{1,t=2}]\end{aligned}$$

The second period model errors (ϵ) are independent of everything and have mean 0, so the second term in brackets will also have expectation 0. So the bias in the estimated treatment effect is:

$$\begin{aligned}Bias &= E[\mu_{treated} - \widehat{w_0} \cdot \mu_0 - \widehat{w_1} \cdot \mu_1] \\ &= E[(w_0^{tgt} - \widehat{w_0}) \cdot \mu_0 + (w_1^{tgt} - \widehat{w_1}) \cdot \mu_1]\end{aligned}$$

The bias in the estimated treatment effect comes from the bias in the estimated SC weights.

A.2 The bias formula

The estimated weights ($\widehat{w_0}, \widehat{w_1}$) depend on the realizations on the first period's ϵ. We can substitute in, and then work to simplify:

$$\begin{aligned}(w_0^{tgt} - \widehat{w_0}) \cdot \mu_0 &= (\frac{\mu_1 - \mu_{treated}}{\mu_1 - \mu_0} - \frac{\mu_1 - \mu_{treated} + \epsilon_1 - \epsilon_{treated}}{\mu_1 - \mu_0 + \epsilon_1 - \epsilon_0}) \cdot \mu_0 \\ (w_1^{tgt} - \widehat{w_1}) \cdot \mu_1 &= (\frac{\mu_{treated} - \mu_0}{\mu_1 - \mu_0} - \frac{\mu_{treated} - \mu_0 + \epsilon_{treated} - \epsilon_0}{\mu_1 - \mu_0 + \epsilon_1 - \epsilon_0}) \cdot \mu_1\end{aligned}$$

Note: The $\epsilon_{treated}$ term is only in the numerator, is independent of everything else, and has mean 0. So we can pull this part out of the equations and set equal to 0 in expectation. This means that the existence of $\epsilon_{treated}$ does not have an influence on the bias.

If we simplify further by plugging in the values $\mu_0 = 0$ and $\mu_1 = 1$, we then get:

$$Bias = E(\mu_{treated} - \frac{\mu_{treated} - \epsilon_0}{1 + \epsilon_1 - \epsilon_0})$$

With the expectation taken over the distribution of the first period model errors, ϵ_1 and ϵ_0. This can also be written as:

$$Bias = E(\frac{\mu_{treated} \cdot (\epsilon_1 - \epsilon_0) + \epsilon_0)}{1 + \epsilon_1 - \epsilon_0})$$

A.3 Calculation of bias as a function of $\mu_{treated}$ and σ_ϵ

Figure 2 in Section 3.1 shows how the bias depends on $\mu_{treated}$ and on σ_ϵ. To calculate this, we use the fact that $\epsilon_{treated}$ plays no role in the bias, and set $\epsilon_{treated} = 0$. We then need to integrate over the distribution of $\epsilon_{i=0,t=1}$ and $\epsilon_{i=1,t=1}$. To do so, we need to determine a distribution for the ϵ. We use a truncated normal distribution. The truncation points are chosen to guarantee that $(\mu_0 + \epsilon_0) \leq (\mu_{treated}) \leq (\mu_1 + \epsilon_1)$. This ensures that the realized observed outcomes Y preserve the same ranking as the underlying fixed effects: $Y_{i=0,t=1} \leq Y_{i=treated,t=1} \leq Y_{i=1,t=1}$. To implement this we first define a Normal distribution with standard deviation σ_{norm}, and then truncate the distribution at $\pm min(\mu_{treated}, 1 - \mu_{treated})$. Figure 2, discussed in Section 3.1, shows a calculation of the bias using numerical integration, for different values of of where the treated unit falls relative to the donor units ($\mu_{treated}$). The x-axis in Figure 2 shows the standard deviation of the resulting truncated variable, and the y-axis shows the bias in the estimated treatment effect.

A.4 SCAE for the illustrative example

In Section 3.2.2 of the paper, we introduce the Synthetic Control Average Error (SCAE), which is the weighted sum of the model errors in the pre-treatment periods, using the synthetic control weights. In this subsection we calculate the SCAE for the illustrative example of Section 3.1 in the paper. This calculation shows that for the illustrative example, the expected value of the SCAE produces the same formula as our definition of the bias of the synthetic control estimator.

The SCAE under the assumptions of the illustrative example is

$$\begin{aligned} SCAE &= \widehat{w_0} \cdot \epsilon_0 + \widehat{w_1} \cdot \epsilon_1 \\ &= \frac{(\mu_1 - \mu_{treated} + \epsilon_1 - \epsilon_{treated}) \cdot \epsilon_0}{\mu_1 - \mu_0 + \epsilon_1 - \epsilon_0} + \frac{(\mu_{treated} - \mu_0 + \epsilon_{treated} - \epsilon_0) \cdot \epsilon_1}{\mu_1 - \mu_0 + \epsilon_1 - \epsilon_0} \end{aligned}$$

Taking expectations over ϵ, the terms for $\epsilon_{treated}$ drop out. Combining remaining terms we have:

$$E[SCAE] = E[\frac{\mu_1\epsilon_0 - \mu_{treated}\epsilon_0 + \mu_{treated}\epsilon_1 - \mu_0\epsilon_1}{\mu_1 - \mu_0 + \epsilon_1 - \epsilon_0}]$$

For the illustrative model of section 3.1, $\mu_0 = 0$ and $\mu_1 = 1$. So the expected SCAE simplifies to:

$$E[SCAE] = E[\frac{\mu_{treated}(\epsilon_1 - \epsilon_0) + \epsilon_0}{1 + \epsilon_1 - \epsilon_0}]$$

This is the same as the formula for the bias of the SC estimator for this illustrative example.

B Alternative Distributions of μ_0

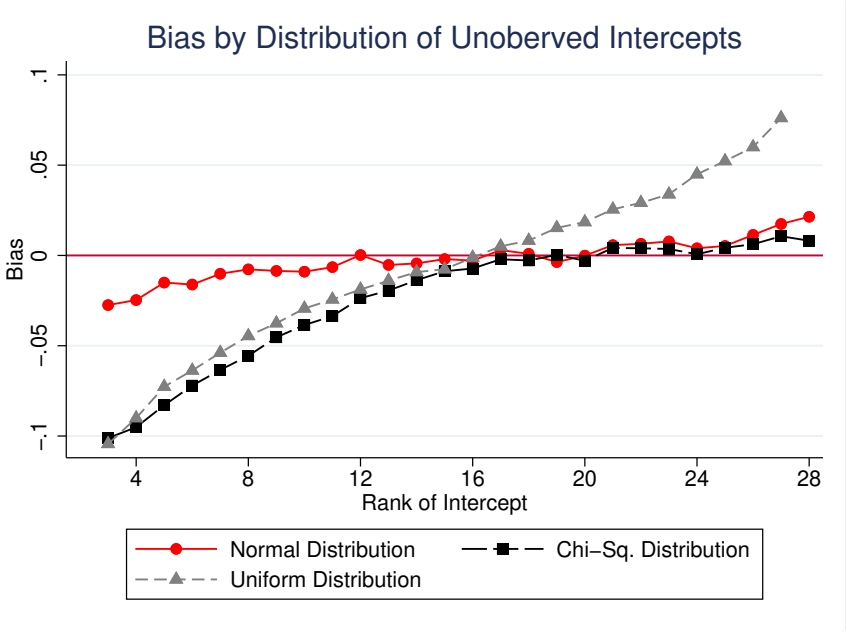


Figure A1: Alternative Distributions of μ_0

The figure graphs mean bias by rank of treated group intercept, drawing the intercepts from alternative distributions. The normal distribution line recreates the line in Figure 4a. Estimates come from 3,000 Monte Carlo repetitions using models matching on all pre-period values of Y and varying distributions of μ.

C Variability of Estimates

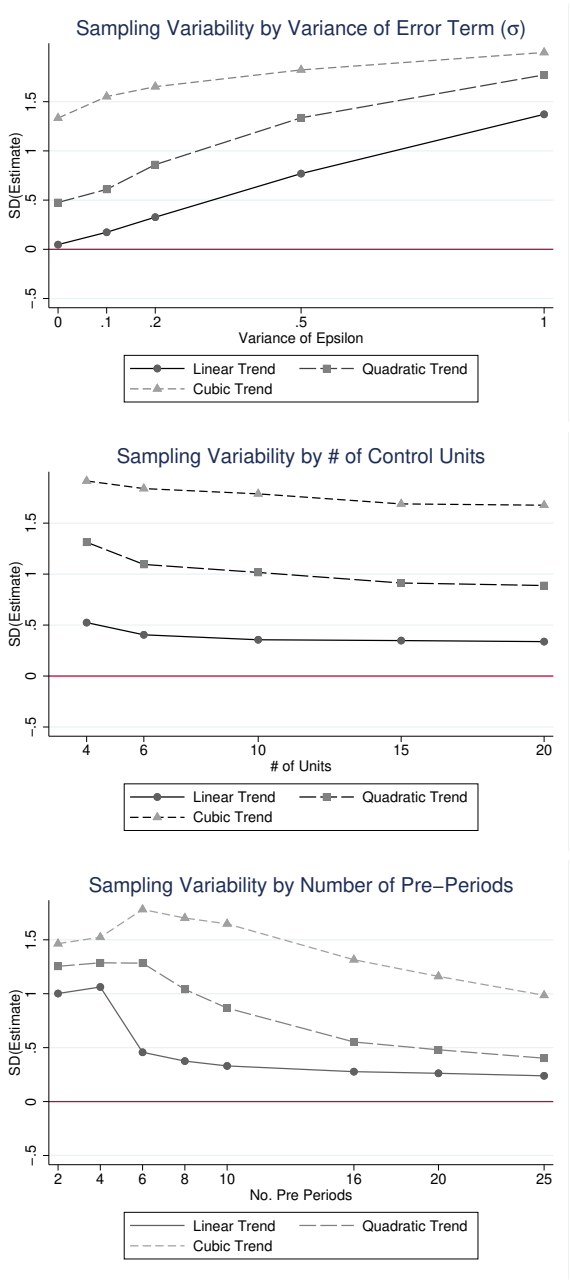


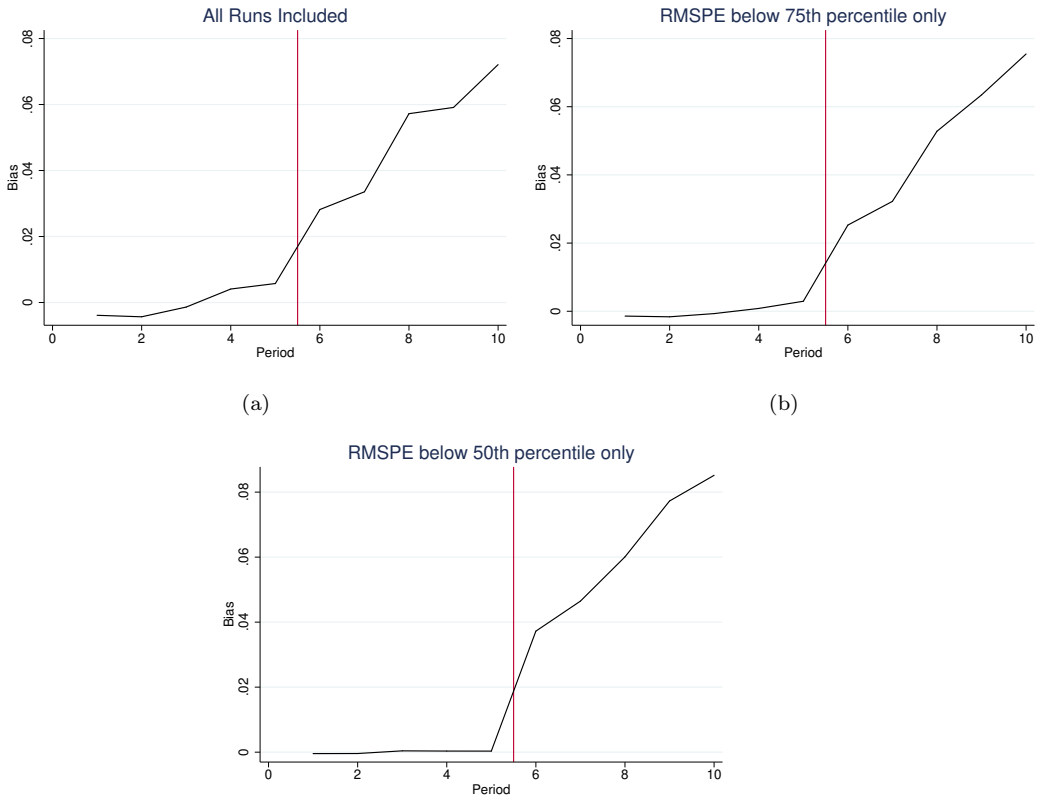
Figure A2: Sampling Variability Across Data Features

These figures above mimic Figures 5, 6 and 7 above, graphing the standard deviation of the synthetic control estimates, as opposed to the the mean, varying the parameters of interest. Estimates come from 3,000 Monte Carlo repetitions using models matching on all pre-period values of Y.

D Pre-Treatment Trend in Bias

Figure 3 shows a slight pre-treatment trend in the bias for higher values of error term variance. As mentioned in the main text, this occurs because in some cases no perfect pre-treatment match is available. These are less interesting cases, since they will be apparent to the researcher in the form of bad pre-treatment matches. Our paper focuses on the case where there is a good pre-treatment match. Figure A3 demonstrates that the bias is not primarily driven by lack of available pre-treatment matches. Figure A3a recreates the case in Figure 3, where error variance is equal to 0.2 and the treated unit trend rank is 22nd out of 31. There is a slight pre-trend in the bias. But if we drop Monte Carlo runs where the root mean squared pre-treatment error (RMSPE, a measure of the pre-treatment fit) is below the 75th percentile across all Monte Carlo runs and recreate the same graph, the trend is diminished. If we further limit the RMSPE to cases below the 50th percentile, where all cases are effectively a perfect match, the bias is nearly exactly zero throughout the pre-treatment period. But in all cases there is still a bias in the post period due to the spurious matching on error terms discussed in this paper.

We considered requiring a maximum RMSPE for a given Monte Carlo run for all analysis performed in this paper (recall that for all analysis we do require that the treated unit outcome is never the highest or lowest of all units in any pre-treatment period, making a match impossible). However, given the multitude of empirical settings we analyze, particularly when varying the dimensionality of the DGP, a consistent RMSPE standard that both enforces an excellent match and does not exclude the vast majority of runs proved impractical. Therefore we highlight the issue here and acknowledge that some of the bias we show in our analysis may be driven by mismatches in the pre-period (though also note that in most analyses we use 30 control units and conservatively set the treated unit rank at the 70th percentile, so good matches are theoretically feasible), but a bias exists even in the case of perfect matching.



(a) (b)

(c)

Figure A3: Bias by period for various matching-quality thresholds

This figure shows mean bias by period for 3 different rules for pre-period match quality of the synthetic controls estimate. The top left panel shows bias results when all Monte Carlo simulations are included. The top right panel limits results to those where the pre-period RMSE is below the 75th percentile of realized RMSE across simulation runs. The bottom panel includes only the lowest 50% pre-period RMSE realized across simulation runs. Results are averaged over 3,000 Monte Carlo replications of the baseline DGP with linear trends.

E Holding Factor Loadings (μ_i) Fixed in Simulations

In our paper we think of the matching on noise bias as being conditional on fixed locations of μ_i, and depending on the relative position of $\mu_{i=treated}$ compared to the donor units. However, in our Monte Carlo simulations we draw random values for μ_i and choose as our treatment group the unit with the appropriate rank of $\sum_{d=0}^{D} \mu_i$. That is, we fix the relative rank of $\mu_{i=treated}$, by randomly generating the values of the mu_is and then always choosing $\mu_{i=treated}$ at its predetermined rank.

This approach allows for easy implementation in our simulations, and lets us show how bias arises when conditioning on the ranking of the treated group among the controls. In this appendix we show that we obtain similar results when we fix the factor loadings μ_i across simulations.

Fixing the factor loadings μ_i requires us to choose what values to fix them at. We do this as follows. First, we take a univariate standard normal distribution and divide it into five segments with equal probability, with four "cut points" dividing the segments (these cut points are at approximately $\{\pm 0.253, \pm 0.842\}$). We assign μ_1 to take the value of the conditional mean within each segment. We next divide the standard normal into six segments (these cut points are at approximately $0, \{\pm 0.532, \pm 1.400\}$), and assign μ_0 to take the value of the conditional mean within each segment. The 30 donor units each take one of the values from the $(5 \times 6 = 30)$ combinations. We select the values of the factor loadings of the treated unit to yield the 70th percentile of the sum of μ_0 and μ_1 from a bivariate standard normal with zero correlation. This corresponds to the 70th percentile of a N(0,2) distribution, so we assign the treated factor loadings to be $mu_0 = mu_1 = 0.3708$ (since the 70th percentile of a N(0,2) is $0.3708 + 0.3708 = 0.7416$).

Using these fixed values of μ, we then repeat our Monte Carlo analysis, allowing for random draws of only ϵ across Monte Carlo replications. We present the results in Appendix Figure A4, which is analogous to Figure 3 in the main text. The pattern of the bias is similar to the baseline results in Figure 3. There is no bias when σ_ϵ is zero; and the bias is monotonically increasing with the variance of the error term. The key difference with Figure 3 is that the magnitude of the bias is slightly smaller for each given level of σ_ϵ in each pre-period: with the largest bias for $\sigma_\epsilon = 0.5$ being slightly above .3 in period 10 rather than slightly below 0.4.

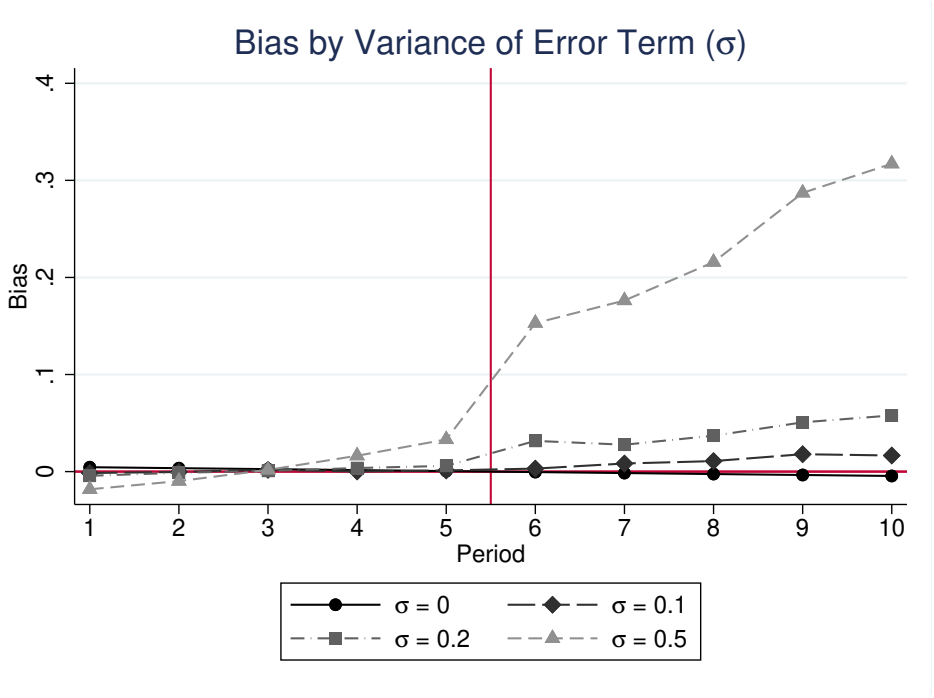


Figure A4: Mean bias estimates by period across Var(ϵ) (σ).

This figure shows the average bias in our Monte Carlo simulations across different pre-periods; fixing the values of the control mu_i (factor loadings for the intercept and trend). We set the treatment group such that the sum of the intercept and slope coefficients is chosen to be the one at the 70th percentile of that distribution (in this case $mu_0 = mu_1 = 0.3708$ since the 70th percentile of a N(0,2) is $0.3708 + 0.3708 = 7416$). We specify the synthetic control estimator to match on all pre-period values of Y. The DGP includes a linear time trend, 10 pre-periods, and 30 donor groups. Results are averaged over 3000 Monte Carlo replications. See text for more details

The similarity between Figure A4 and Figure 3 reassures us that our results are driven by the randomness of the ϵ. They are not an artifact of drawing different values of mu_i's across simulations. We see this as evidence that our approach of fixing the rank of $\mu_{i=treated}$ while drawing μ_i across simulations is conceptually similar.

F Example Corrections for Free Trade Application

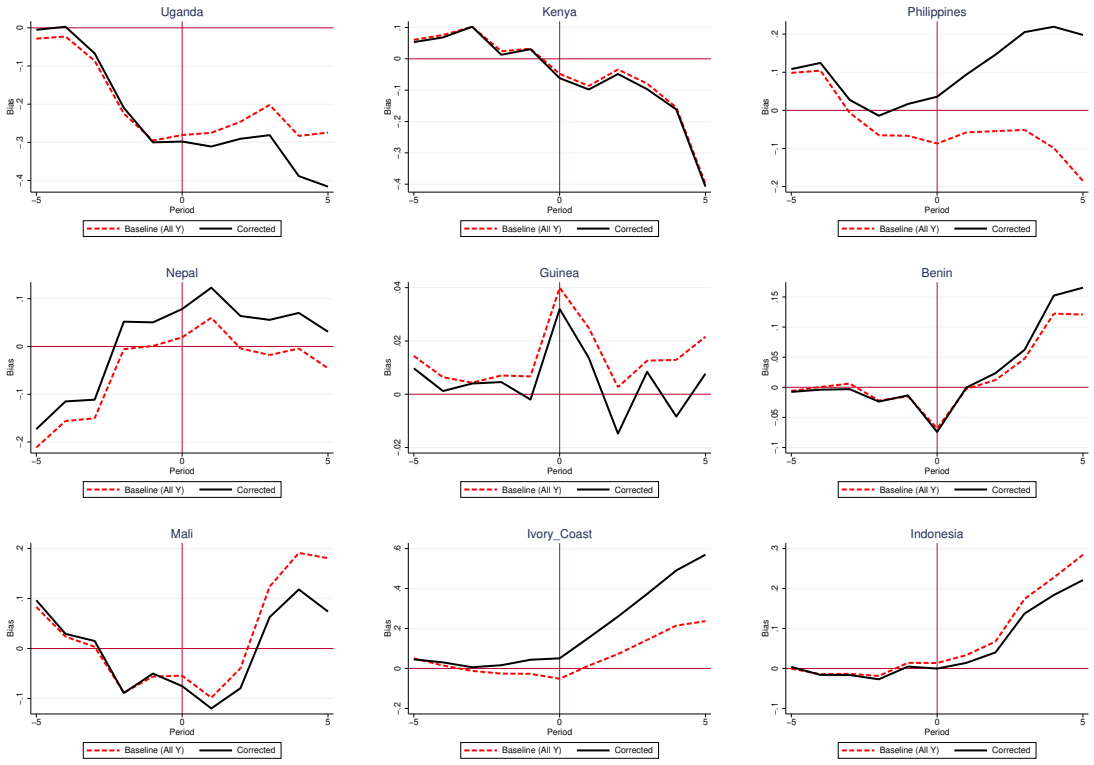


Figure A5: Free Trade Demo by Country

The figure above provides country-level estimates of the effects of Free Trade on (ln) GDP/cap (following Billmeier and Nannicini (2013)). Baseline estimates in the dashed red line match on all pre-period Y values; the solid black line represents mean bias estimates from our feasible bootstrap correction. We display 9 of the 17 estimates, sorted (beginning top left) from lowest to highest estimated average post-period treatment effect based on the baseline specification. Results for all countries are available upon request.