Placebo Tests for Synthetic Controls

Ferman, Bruno and Pinto, Cristine

Sao Paulo School of Economics - FGV

1 April 2017

Placebo Tests for Synthetic Controls*

Bruno Ferman† Cristine Pinto‡

Sao Paulo School of Economics - FGV

First Draft: March, 2017

Please click here for the most recent version

Abstract

The synthetic control (SC) method has been recently proposed as an alternative to estimate treatment effects in comparative case studies. An important feature of the SC method is the inferential procedures based on placebo studies, suggested in Abadie et al. (2010). In this paper, we evaluate the statistical properties of these inferential techniques. We first show that the graphical analysis with placebos can be misleading, as placebo runs with lower expected squared prediction errors would still be considered in the analysis. Then we show that a test based on the the post/pre-intervention mean squared prediction error, as suggested in Abadie et al. (2010), ameliorates this problem. However, we show that such test can still have some size distortions, even if we consider a case in which the test statistic has the same marginal distribution for all placebo runs. Finally, we show that the fact that the SC weights are estimated can lead to important additional size distortions.

Keywords: synthetic control, difference-in-differences; linear factor model, inference, permutation test
JEL Codes: C12; C13; C21; C23

*Most of the results in this paper were originally included in an earlier version of Ferman and Pinto (2016b). We would like to thank Ivan Canay and Jin Hahn for comments and suggestions. We also thank Deivis Angeli for excellent research assistance.
†bruno.ferman@fgv.br
‡cristine.pinto@fgv.br

1 Introduction

In a series of influential papers, Abadie and Gardeazabal (2003), Abadie et al. (2010), and Abadie et al. (2015) proposed the Synthetic Control (SC) method as an alternative to estimate treatment effects in comparative case studies when there is only one treated unit. The main idea of the SC method is to use the pre-treatment periods to estimate weights such that a weighted average of the control units reconstructs the pre-treatment outcomes of the treated unit. Then they use these weights to compute the counterfactual of the treated unit in case it were not treated. According to Athey and Imbens (2016), *"the simplicity of the idea, and the obvious improvement over the standard methods, have made this a widely used method in the short period of time since its inception".* An important feature of the SC method is the inferential procedure based on placebo studies, suggested in Abadie et al. (2010).

In this paper, we consider the statistical properties of the inferential techniques proposed in Abadie et al. (2010). In the absence of random assignment, Abadie et al. (2010) and Abadie et al. (2015) interpret the p-value from their placebo tests as *"the probability of obtaining an estimate at least as large as the one obtained for the unit representing the case of interest when the intervention is reassigned at random in the data set"* (Abadie et al. (2015), page 500). While we agree this is a useful measure, it is important to evaluate the statistical properties of such tests. Analyzing the SC method in a linear factor model setting as the one considered in Ferman and Pinto (2016b), we derive the asymptotic distribution of the test statistics used in these placebo tests. Moreover, we evaluate whether such tests satisfy the conditions for the theory of randomization inference under an approximate symmetry assumption, developed in in Canay et al. (2014).

We first show that the graphical analysis proposed in Abadie et al. (2010) or a placebo test using the post-treatment mean squared prediction error (MSPE) as test statistic might lead to important size distortions, as the distribution of the post-treatment prediction errors for a given permutation might depend on, for example, the variance of the transitory shocks or the concentration of the SC weights. Such distortions can arise whether or not the SC estimator is asymptotically unbiased.[1] We also note that the strategy suggested in Abadie et al. (2010) of excluding placebos with a poor pre-treatment fit from the graphical analysis can be misleading as, under this strategy, placebos with a better pre-treatment fit relative to the treated unit would still be considered. Since placebos with a lower pre-treatment MSPE would tend to have a less volatile post-intervention prediction error, this may lead researchers to over-estate the significancy of their results. We recommend a slight modification in the graphical analysis to take this distortion into account.

[1] See Ferman and Pinto (2016b) for conditions under which the SC estimator is asymptotically unbiased.

Then we show that a placebo test using the ratio of post/pre-treatment MSPE as test statistic, also suggested in Abadie et al. (2010), can ameliorate this problem. If the SC estimator is asymptotically unbiased, then, under some conditions, the test statistics will have the same asymptotic (marginal) distribution for all permutations.[2] However, even under such conditions, we show that it is not possible to guarantee that the test is asymptotically valid, as the test statistics are generally not based on functions of the data that exhibit approximate symmetry, as would be required to apply the results on randomization tests under an approximate symmetry assumption from Canay et al. (2014). We provide examples in which we can have some size distortions even when the test statistics for all placebos have the same marginal distribution.

Finally, we show that the placebo test using the ratio of post/pre-treatment MSPE as test statistic can have important size distortions for at least three additional reasons due to the fact that the SC weights are estimated. First, if the SC estimator is asymptotically biased, then the expected value of the test statistic for the treated unit should be higher than for the control units, leading to over-rejection. Interestingly, our Monte Carlo (MC) simulations suggest that this over-rejection may appear even when the variance of the transitory shocks is small, in which case the pre-treatment fit should be good and the bias of the SC estimator should be small. This happens because, in this case, the variance of the SC estimator would be relatively small as well, so even a small bias could generate relevant size distortions. Second, we show that the post/pre ratio of MSPE may fail to properly correct the marginal distribution of the test statistics for a finite number of pre-treatment periods (T_0). This might happen because, with small T_0, the model might overfit the pre-treatment MSPE, so it might not provide a proper correction for the post-treatment MSPE. Finally, the fact that the SC method should only be used when there is a good pre-treatment fit while the placebos would be considered regardless of that can also lead to over-rejection. This happens because the test statistic of the treated unit would be conditional on a denominator close to zero, while the test statistic for the placebos would not.

A few recent papers analyzed in detail the placebo tests proposed in Abadie et al. (2010). Firpo and Possebom (2016) formalize the placebo test for the case where treatment is randomly assigned. In this case, the inference method suggested in Abadie et al. (2010) would provide valid inference for *unconditional* tests. Differently from Firpo and Possebom (2016), our paper considers the asymptotic properties of the placebo tests when we relax the hypothesis of random assignment. Also, even under random assignment,

[2] This will be the case if linear combinations of the transitory shocks and common factors are stationary, serially uncorrelated, and i.i.d. across units up to a scale parameter. We derive an alternative test statistic that guarantees the same asymptotic expected value and variance for all permutations under weaker conditions.

we can consider hypothesis testing conditional on the data on hand.[3] Hahn and Shi (2016) point out to the possibility of severe size distortions in placebo tests for the SC method in MC simulations. They focus, however, on a test statistic based on the post-intervention MSPE. We show that such severe size distortions can be strongly attenuated once we consider the post/pre ratio of MSPE as test the statistic, although the test may still present size distortions even if the test statistics for all placebos have the same marginal distribution. Finally, Ando and Sävje (2013) argue that the placebo test proposed by Abadie et al. (2010) is generally not valid and derive an alternative inference method. Differently from Ando and Sävje (2013), we consider the asymptotic properties of Abadie et al. (2010) placebo tests when the number of pre-intervention is large. Moreover, Ando and Sävje (2013) focus on the case in which the placebo test could have size distortions because the SC estimator would fail to reconstruct the factor loadings of the "treated" unit for some placebo runs, while we show that there might be size distortions even if we consider weights that satisfy this condition for all placebo runs.[4]

The remainder of this paper proceeds as follows. We present a brief review of the SC method in Section 2. In Section 3, we show that the placebo tests might have size distortions even when we consider an "infeasible" SC estimator that uses weights that correctly reconstruct the factor loadings of the treated unit. In Section 4, we consider additional sources of size distortions that are generated by the fact that the SC weights are estimated. We conclude in Section 5.

2 A Brief Review of The Synthetic Control Model

2.1 Setting

We consider the SC estimator in a linear factor models setting, as in Ferman and Pinto (2016b). Suppose we have a balanced panel of $J+1$ units indexed by i observed on $t = 1, ..., T$ periods. We want to estimate the treatment effect of a policy change that affected only unit $j = 1$ from period $T_0 + 1 \leq T$ to T. The potential outcomes are given by:

$$\begin{cases} y_{it}(0) = \delta_t + \lambda_t \mu_i + \epsilon_{it} \\ y_{it}(1) = \alpha_{it} + y_{it}(0) \end{cases} \tag{1}$$

[3] See Ferman and Pinto (2016a) for details on why conditional tests should be preferable when there are few treated units.

[4] Carvalho et al. (2015), Carvalho et al. (2016) and Powell (2016) consider extensions of the SC estimator, and derive large sample inferential techniques for hypothesis testing regarding the average effect across the post-treatment periods when both the number of pre- and post-treatment periods go to infinity. In this paper, we focus on the case in which the number of pre-treatment periods is large, but the number of post-treatment periods is finite.

where δ_t is an unknown common factor with constant factor loadings across units, λ_t is a $(1 \times F)$ vector of common factors, μ_i is a $(F \times 1)$ vector of unknown factor loadings, and the error terms ϵ_{it} are unobserved transitory shocks. We only observe $y_{it} = d_{it}y_{it}(1) + (1 - d_{it})y_{it}(0)$, where $d_{it} = 1$ if unit i is treated at time t. Note that the unobserved error $u_{it} = \lambda_t \mu_i + \epsilon_{it}$ might be correlated across units due to the presence of $\lambda_t \mu_i$. Since we hold the number of units $(J + 1)$ fixed and look at asymptotics when the number of pre-treatment periods goes to infinity, we treat the vector of unknown factor loads (μ_i) as fixed and the common factors (λ_t) as random variables. In order to simplify the exposition of our main results, we consider the model without observed covariates Z_i.

An important feature of our setting is that the SC estimator is only well defined if it actually happened that one unit received treatment in a given period. We define $D(1, T_0)$ as a dummy variable equal to 1 if unit 1 is treat after T_0 while all other units do not receive treatment.[5] Assumption 1 makes it clear that the sample a researcher observers when considering the SC estimator is always conditional on the fact that one unit was treated in a given period.

Assumption 1 (conditional sample) We observe a realization of $\{y_{1t}, ..., y_{J+1,t}\}$ for $t = 1, ..., T$ conditional on $D(1, T_0) = 1$.

We also impose that the treatment assignment is not informative about the first moment of the transitory shocks.

Assumption 2 (transitory shocks) $E[\epsilon_{jt}|D(1, T_0)] = E[\epsilon_{jt}] = 0$

Assumption 2 implies that, once we condition on the common factors λ_t, the transitory shocks are mean-independent from the treatment assignment. This assumption implies that $E[y_{jt}(0)|D(1, T_0), \lambda_t] = E[y_{jt}(0)|\lambda_t]$ and $E[y_{jt}(1)|D(1, T_0), \lambda_t] = E[y_{jt}(1)|\lambda_t]$. Note that this assumption excludes the possibility that treatment assignment is informative about the transitory shocks. However, we still allow for the possibility that the treatment assignment to unit 1 is correlated with the unobserved common factors. More specifically, we allow for $E[\lambda_t|D(1, T_0)] \neq E[\lambda_t]$.

We define Φ_1 as the set of weights such that a weighted average of the factor loadings of the control units reconstructs the factor loadings of the treated unit. That is:

$$\Phi_1 = \left\{ \mathbf{w}_1^* \in \mathbb{R}^J \mid \mu_1 = \sum_{j \neq 1} w_1^{j*} \mu_j, \ \sum_{j \neq 1} w_1^{j*} = 1, \text{ and } w_1^{j*} \geq 0 \right\}$$

[5] That is, one can think of $D(1, T_0)$ as a product between two indicator variables, one for the event that the treated unit is unit 1, and the other one that the treatment starts after T_0.

where w_1^j is the weight associated to unit j when we re-construct the factor loadings of unit 1. Note that it may be that $\Phi_1 = \varnothing$.

If we knew $\mathbf{w}_1^* \in \Phi_1$, then we could consider an *infeasible* SC estimator using these weights, $\hat{\alpha}_{1t}^* = y_{1t} - \sum_{j \neq 1} w_1^{j^*} y_{it}$. For a given $t > T_0$, we have that:

$$\hat{\alpha}_{1t}^* = y_{1t} - \sum_{j \neq 1} w_1^{j^*} y_{it} = \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j \neq 1} w_1^{j^*} \epsilon_{jt} \right) \tag{2}$$

Therefore, under Assumption 2, we have that $E[\hat{\alpha}_{1t}^* | D(1, T_0) = 1] = \alpha_{1t}$, which implies that this *infeasible* SC estimator is unbiased.

2.2 The SC estimator

The main idea of the SC method consists of estimating the SC weights $\widehat{\mathbf{w}}_1 = \{\hat{w}_1^j\}_{j \neq 1}$ using information on the pre-treatment period, so that we can construct the SC estimator $\hat{\alpha}_{1t} = y_{1t} - \sum_{j \neq 1} \hat{w}_1^j y_{jt}$ for $t > T_0$. Abadie et al. (2010) suggest a minimization problem to estimate these weights using the pre-intervention data. They define a set of K economic predictors where X_1 is a $(K \times 1)$ vector containing the economic predictors for the treated unit and X_0 is a $(K \times J)$ matrix of economic predictors for the control units.[6] The SC weights are estimated by minimizing $||X_1 - X_0 \mathbf{w}_1||_V$ subject to $\sum_{i=2}^{J+1} w_1^j = 1$ and $w_1^j \geq 0$, where V is a $(K \times K)$ positive semidefinite matrix. They discuss different possibilities for choosing the matrix V, including an iterative process where V is chosen such that the solution to the $||X_1 - X_0 \mathbf{w}_1||_V$ optimization problem minimizes the pre-intervention prediction error. In other words, let $\mathbf{Y}_1^P$ be a $(T_0 \times 1)$ vector of pre-intervention outcomes for the treated unit, while $\mathbf{Y}_0^P$ be a $(T_0 \times J)$ matrix of pre-intervention outcomes for the control units. Then the SC weights would be chosen as $\widehat{\mathbf{w}}_1(V^*)$ such that V^* minimizes $||\mathbf{Y}_1^P - \mathbf{Y}_0^P \widehat{\mathbf{w}}_1(V)||$.

Here we focus on the specification that uses all pre-treatment outcome lags as economic predictors. In this case, the optimization problem to derive the SC weights simplifies to:

$$\widehat{\mathbf{w}}_1 \quad = \quad \underset{\mathbf{w} \in W}{\operatorname{argmin}} \frac{1}{T_0} \sum_{t=1}^{T_0} \left[y_{1t} - \sum_{j \neq 1} w_1^j y_{jt} \right]^2$$

where $W = \{\mathbf{w}_1 \in \mathbb{R}^J | w_1^j \geq 0 \text{ and } \sum_{j \neq 1} w_1^j = 1\}$.

Ferman and Pinto (2016b) show that the SC weights will, in general, converge to weights that do not

[6] Economic predictors can be, for example, linear combinations of the pre-intervention values of the outcome variable or other covariates not affected by the treatment.

reconstruct the factor loadings of the treated unit. That is, in general, $\widehat{\mathbf{w}}_1 \to_p \bar{\mathbf{w}}_1 \notin \Phi_1$, even if $\Phi_1 \neq \varnothing$. In a setting in which the pre-treatment averages of the first and second moments of the common factors and transitory shocks converge, they show that SC estimator will converge to:

$$\hat{\alpha}_{1t} = y_{1t} - \sum_{j \neq 1} \hat{w}_1^j y_{it} \xrightarrow{d} \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j \neq 1} \bar{w}_1^j \epsilon_{jt} \right) + \lambda_t \left(\mu_1 - \sum_{j \neq 1} \bar{w}_1^j \mu_j \right) \tag{3}$$

where, in general, $\mu_1 \neq \sum_{j \neq 1} \bar{w}_1^j \mu_j$. This implies that the SC estimator will be asymptotically biased when the fact that treatment was assigned to unit 1 after time T_0 is informative about the unobserved common factors.

Ferman and Pinto (2016b) also show that, in a setting in which a subset of the common factors include a linear time trend or $I(1)$ processes, then the SC weights will converge to weights that reconstruct the factor loadings associated to the linear time trend of to the $I(1)$ processes. In this case, the SC estimator would be asymptotically unbiased even if treatment assignment is informative about these non-stationary common factors. However, the SC weights would still fail to reconstruct the factor loadings associated to the stationary common factors, so it will be asymptotically biased if treatment assignment is informative about the stationary common factors.

2.3 Inference: placebo tests

As argued in Abadie et al. (2010), large sample inferential techniques are not well suited to comparative case studies when the number of units in the comparison group is small.[7] They propose instead placebo tests where they apply the SC method to every potential control in the sample. First, they consider a graphical analysis where they compare the post-treatment prediction error of the SC estimator with the prediction error for each of SC placebo estimator. Then they consider whether the prediction error when one considers the actual treated unit is "unusually" large relative to the distribution of prediction errors for the units in the donor pool. Note that the graphical analysis suggested in Abadie et al. (2010) does not provide a clear decision rule on whether the null hypothesis should be rejected. Still, this analysis would implicitly reject the null when the post-intervention MSPE for the SC estimate is greater than the post-intervention MSPE for the placebo estimates. We consider, therefore, the post-intervention MSPE as the test statistic in order

[7] Carvalho et al. (2015), Carvalho et al. (2016) and Powell (2016) rely on large sample inferential techniques. Instead of testing the null hypothesis of no effect for all post-treatment periods, they test whether the average effect across time is equal to zero. If both the number of pre- and post-intervention periods is large, then they are able to derive the asymptotic distribution of the estimator. This method would not work if one wants to test the null of no effect for all post-treatment periods or if the number of post-intervention periods is finite.

to analyze potential distortions in such graphical analysis:

$$t_i^{\text{post}} = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \left[y_{it} - \sum_{j \neq i} \hat{w}_i^j y_{jt} \right]^2 \tag{4}$$

Then they also suggest a placebo test comparing the post/pre-treatment MSPE as test statistic.

$$t_i^{\text{ratio}} = \frac{\frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \left[y_{it} - \sum_{j \neq i} \hat{w}_i^j y_{jt} \right]^2}{\frac{1}{T_0} \sum_{t=1}^{T_0} \left[y_{it} - \sum_{j \neq i} \hat{w}_i^j y_{jt} \right]^2}$$

Abadie et al. (2010) and Abadie et al. (2015) recognize that the assumptions required in the classical randomization inference setting (in particular, random treatment assignment) are rather restrictive in the SC setting. Still, they argue that it is possible to interpret the p-values from their placebo tests as *"the probability of obtaining an estimate at least as large as the one obtained for the unit representing the case of interest when the intervention is reassigned at random in the data set"* (Abadie et al. (2015), page 500). While we agree that this interpretation of the placebo tests p-values is useful, it is important to consider the statistical properties of such tests.

3 Placebo tests with "infeasible" SC estimator

We start considering the properties of the placebo test using an *infeasible* SC estimator which uses weights that correctly reconstruct the factor loadings of the treated unit. This way we are able to disentangle the potential problems that arise due to the estimation of the SC weights as compared to problems that would arise even for an *infeasible* SC estimator.

3.1 Graphical analysis & post-MSPE

We consider first the graphical analysis suggested in Abadie et al. (2010). As mentioned in Section 2.3, the graphical analysis would suggest that the treatment effect is different from zero if t_1^{post} is "unusually" large relative to the distribution of $\{t_i^{\text{post}}\}_{i=1}^{J+1}$. Assuming that we know $\mathbf{w}_i^* \in \Phi_i$ for all $i = 1, ..., J+1$, then we

have that:[8]

$$t_i^{\text{post}} = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \left[\epsilon_{it} - \sum_{j \neq i} w_i^{j^*} \epsilon_{jt} \right]^2 \tag{5}$$

There are at least three reasons why this test statistic might not have the same (marginal) asymptotic distribution for all permutations. First, the transitory shock might be heteroskedastic. Ferman and Pinto (2016a) show that this would usually be true in the Differences-in-Differences setting if we have unit x time aggregate values and there is variation in the number of observations per unit. This would be the case, for example, if one uses the Current Population Survey (CPS). Note that, in this case, t_i^{post} would tend to attain higher values when the treated unit is small relative to the units in the donor pool. Second, even if the transitory shock is homoskedastic, the variance of $\epsilon_{it} - \sum_{j \neq i} w_i^{j^*} \epsilon_{jt}$ will depend on the weights $\{w_i^{j^*}\}_{j \neq i}$. If the weights for unit i are more concentrated around a few units in the donor pool, then the variance of t_i^{post} should be higher than if the weights were more evenly distributed. Finally, t_i^{post} would not have the same distribution as t_1^{post} if, for some i, $\Phi_i = \varnothing$. In this case, the distribution of t_i^{post} would also depend on the common factors λ_t. Hahn and Shi (2016) provide MC simulations showing that a permutation test using t_i^{post} as test statistic may severely over-reject under the null, even if one uses an infeasible SC estimator that relies on weights that correctly reconstruct the factor loadings of the treated unit.

Abadie et al. (2010) correctly noticed that the outcome variable may not be well reproduced for some units by a convex combination of the other units for the pre-intervention periods, and that the post-intervention MSPE for these units should be high as well. For this reason, they exclude placebos in which the pre-intervention MSPE is 20 times (or 5 times) larger than the pre-intervention MSPE for the treated unit. Note that, considering the *infeasible* SC estimator and using that $\mathbf{w}_i \in \Phi_i$ for all i, then the prediction error would be $\epsilon_{it} - \sum_{j \neq i} w_i^{j^*} \epsilon_{jt}$ whether time t is either pre- or post-intervention. Therefore, assuming that ϵ_{it} is stationary, then it would be likely that, in our setting, t_i^{post} under the null has the same asymptotic marginal distribution as t_1^{post} if the pre-intervention MSPE for unit i and unit 1 are similar. Note, however, that Abadie et al. (2010) procedure only excludes placebos with pre-intervention MSPE *higher* than the pre-intervention MSPE for the treated unit. Therefore, if there are many placebos with lower pre-intervention MSPE, then

[8] Ando and Sävje (2013) argue that in most applications it would not be reasonable to assume that this assumption is valid for all i. We believe that this condition might be reasonable in some applications. For example, this condition is satisfied if we have different groups of units where time trends are different across groups but parallel within groups, as considered in Ferman et al. (2016) and Ferman and Pinto (2016b). We analyze this case in detail in our MC simulations. In this case, the main idea of the SC estimator would be to select the control units that follow the same time trend as the treated unit. We consider below the implications in case assumption 1 is not valid for all i.

the test would over-reject the null since t_1^{post} would tend to attain larger values. In this case, Abadie et al. (2010) graphical analysis could be misleading, even if we consider weights that in that correctly reconstruct the factor loadings of the treated unit.

One possibility to ameliorate this problem is to re-scale the post- and pre-intervention prediction errors of the control units using the pre-intervention MSPE. More specifically, for placebo i, we can divide its prediction error by its the squared root of its pre-intervention MSPE, and multiply it by the squared root of unit 1 pre-intervention MSPE. As described in detail below in Section 3.2, under some conditions, this strategy would imply in prediction errors with the same variance for all placebos. Note that this strategy precludes the necessity of choosing arbitrary cut-offs for the exclusion of ill-fitting placebo runs.[9]

3.2 Post/pre-MSPE ratio

A second inference procedure suggested by Abadie et al. (2010) is a placebo test using the ratio of post/pre-intervention MSPE (t_i^{ratio}). According to them, *"the main advantage of looking at ratios is that it obviates choosing a cut-off for the exclusion of ill-fitting placebo runs"*.

Assuming again an *infeasible* SC estimator which uses weights that correctly reconstruct the factor loadings of the treated unit, we have that:

$$t_i^{\text{ratio}} = \frac{\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\left[\epsilon_{it} - \sum_{j\neq i} w_i^{j^*}\epsilon_{jt}\right]^2}{\frac{1}{T_0}\sum_{t=1}^{T_0}\left[\epsilon_{it} - \sum_{j\neq i} w_i^{j^*}\epsilon_{jt}\right]^2} \tag{6}$$

Note that, if we let $T_0 \to \infty$, then:

$$t_i^{\text{ratio}} \to_d \frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\left[\frac{\epsilon_{it} - \sum_{j\neq i} w_i^{j^*}\epsilon_{jt}}{\sqrt{var(\epsilon_{it} - \sum_{j\neq i} w_i^{j^*}\epsilon_{jt})}}\right]^2 \tag{7}$$

Therefore, t_i^{ratio} will have the same asymptotic (marginal) distribution for all i if $Q_{it} = \epsilon_{it} - \sum_{j\neq i} w_i^{j^*}\epsilon_{jt}$ is stationary, serially uncorrelated, and i.i.d. across i up to a scale parameter. If we assume that $\frac{E[Q_{it}^4]}{(E[Q_{it}^2])^2}$ is constant, and still maintain that errors are serially uncorrelated and stationary, then the test statistic has, asymptotically, the same expected value and variance for all placebos.

[9] If we constraint the SC unit to convex combinations of the control units, as in Abadie et al. (2010), then there is no guarantee that the pre-treatment prediction error will have mean zero. This will be particularly relevant for cases in which the SC weights are estimated or when $\Phi_j = \varnothing$ for some j. An alternative would be to use a demeaned SC estimator, as recommended in Ferman and Pinto (2016b). This is equivalent to relaxing the no-constant constraint, as presented in Doudchenko and Imbens (2016).

Note that if we relax the assumption that errors are serially uncorrelated, then there is no guarantee that the variance of t_i^{ratio} will be the same for all i. This happens because the denominator in 7 will not correctly re-scale the variance of the post-intervention MSPE, due to the serial correlation. Instead, we can construct an alternative test statistic $\tilde{t}_i$ that has asymptotically the same expected value and variance for all placebos. Define $S_i = \frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\left[\epsilon_{it} - \sum_{j\neq i} w_i^{j*}\epsilon_{jt})\right]^2$. We can use:

$$\tilde{t}_i = \frac{\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\left[y_{it} - \sum_{j\neq i} w_i^{j*} y_{jt}\right]^2 - \widehat{E[S_i]}}{\sqrt{\widehat{var(S_i)}}} \tag{8}$$

where $\widehat{E[S_i]}$ is an estimator for $E[S_i]$ and $\widehat{var(S_i)}$ is an estimator for $var[S_i]$. With large T_0, we can construct a new time serie $S_{it} = \frac{1}{T-T_0}\sum_{t'=t}^{t+T-T_0}\left[y_{it'} - \sum_{j\neq i} w_i^{j*} y_{jt'}\right]^2$ using the pre-treatment periods and calculate $\widehat{E[S_i]}$ and $\widehat{var(S_i)}$. In Appendix A.1, we provide conditions such that these are consistent estimators, and show that, in this case, the asymptotic distribution of $\tilde{t}_i$ has expected value equal to zero and variance equal to 1 for all placebos.[10]

The test statistics t_i^{ratio} and $\tilde{t}_i$ help prevent that test statistics for different placebos have wildly different asymptotic (marginal) distributions, which could generate severe size distortions. However, it is important to note that, even if the test statistics for all placebos have the same asymptotic (marginal) distributions, it is not possible to guarantee that the placebo test is asymptotically valid. Following Canay et al. (2014), such test would be asymptotically valid if the test statistics are based on a function of the data that exhibits approximate symmetry. In the SC setting, this will not generally be the case, because the SC estimator is a function of transitory shocks of the treated and control units, which induces correlation between test statistics in different permutations. With fixed J, this correlation will not vanish, even when $T_0 \to \infty$, as noticed in Powell (2016). We provide now examples in which the test statistics can have the same asymptotic distribution for all permutations, but we still can have size distortions.

Again, we assume that we know $\mathbf{w}_i \in \Phi_i$ and that we know the variance of $y_{it} - \sum_{j\neq i} w_i^j y_{jt}$. Consider first a model with two common factors, $\lambda_t = (\lambda_t^1, \lambda_t^2)$, where $\mu_i = (1,0)$ for $i = 1,2,3$ and $\mu_i = (0,1)$ for $i = 4, ..., 20$. Assume also that $\epsilon_{it} \overset{i.i.d.}{\sim} N(0,1)$ for all i and t. An (infeasible) SC estimator for the treatment effect at time t in this model for units $i = 1,2,3$ uses the average of the other 2 units that have the same factor loadings to construct the SC estimator, while for units $i = 4, ..., 20$ it uses the average of the the other

[10] Note that this test statistic can also be used with the feasible SC estimator. In this case, we also need to impose assumptions on the time series of the common factors λ_t.

16 units that have the same factor loadings. Now consider the vector $\left(\frac{\hat{\alpha}_1}{\sqrt{var(\hat{\alpha}_1)}}, \frac{\hat{\alpha}_2}{\sqrt{var(\hat{\alpha}_2)}}, ..., \frac{\hat{\alpha}_{20}}{\sqrt{var(\hat{\alpha}_{20})}}\right)'$, where $\hat{\alpha}_j$ is the SC estimator using unit j as treated. For all i, $\frac{\hat{\alpha}_i}{\sqrt{var(\hat{\alpha}_i)}} \sim N(0,1)$. However:

$$cov\left(\frac{\hat{\alpha}_i}{\sqrt{var(\hat{\alpha}_i)}}, \frac{\hat{\alpha}_k}{\sqrt{var(\hat{\alpha}_k)}}\right) = \begin{cases} -0.5 \text{ if } i \in \{1,2,3\} \text{ and } k \in \{1,2,3\}/\{i\} \\ 0 \text{ if } i \in \{1,2,3\} \text{ and } k \in \{4,...,20\} \\ -0.06 \text{ if } i \in \{4,...,20\} \text{ and } k \in \{4,...,20\}/\{i\} \end{cases} \tag{9}$$

Therefore, while all elements in this vector have the same marginal distribution, the conditional distributions are not the same for all placebos. This implies a mild under-rejection of 4.3% for a 5% test when we consider unit 1 as treated.[11] Intuitively, this happens because the high correlation between $\frac{\hat{\alpha}_1}{var(\hat{\alpha}_1)}$ and $\frac{\hat{\alpha}_2}{var(\hat{\alpha}_2)}$ implies that, when $\frac{\hat{\alpha}_1}{var(\hat{\alpha}_1)}$ is extreme, the realization of $\frac{\hat{\alpha}_2}{var(\hat{\alpha}_2)}$ is likely to be extreme as well. On the contrary, when a realization of $\frac{\hat{\alpha}_i}{var(\hat{\alpha}_i)}$ for $i > 3$ is extreme, it does not imply that the realizations of other $\frac{\hat{\alpha}_k}{var(\hat{\alpha}_k)}$ for $k \neq i$ are likely to be extreme as well. Of course, if we do have random assignment, then this placebo test would still have the correct size for *unconditional* tests, as the conditions for randomization inference would be satisfied (see Fisher (1935)). However, if the probability that a unit with the same characteristics as unit 1 is more likely to receive treatment, then we would have under-rejection.

We now show another example in which heteroskedasticity can also generate size distortions, even if the linear factor structure is symmetric. Assume now that we have 20 units in total. We have 5 common factors $\lambda_t = (\lambda_t^1, ..., \lambda_t^5)$, and $\mu_i = (1,0,0,0,0)$ for units $i = 1, ..., 4$, $\mu_i = (0,1,0,0,0)$ for units $i = 5, ..., 8$, and so on. Consider that $var(\epsilon_{1t}) = \sigma^2$ and $var(\epsilon_{it}) = 1$ for all $i > 1$. We calculate the infeasible SC estimator $\hat{\alpha}_i$ as the minimum variance estimator such that $\mathbf{w}_i \in \Phi_i$.[12] In this case, a higher σ^2 implies a lower correlation between $\hat{\alpha}_1$ and $\hat{\alpha}_i$ for $i \in \{2,3,4\}$. This happens because, when σ^2 is higher, then the SC estimator $\hat{\alpha}_i$ for $i \in \{2,3,4\}$ will assign lower weights for y_{1t}. If $\sigma^2 = 2$, then rejection rate is 5.3%, while rejection rate is 5.5% if $\sigma^2 = 5$. If $\sigma^2 < 1$, then we increase the correlation between $\hat{\alpha}_1$ and $\hat{\alpha}_i$ for $i \in \{2,3,4\}$. If $\sigma^2 = 0.5$, then rejection rate is 4.6%, while if $\sigma^2 = 0.1$, then rejection rate is 4%. Again, these results suggest that heteroskedasticity can generate size distortions in the permutation test even when the marginal distributions of the test statistics are the same for all permutations. However, based on our examples, size distortions are relatively mild even if we consider a highly heteroskedastic model.

[11] This rejection rate was calculated based on 10.000.000 MC simulations.

[12] In this case, if $i = 1$ or $i > 4$, then we construct the SC unit as the simple average of the other units that have the same factor loading as the treated unit. If $i \in \{2,3,4\}$, then we construct the SC unit assigning weight equal to $\frac{1}{2\sigma^2+1}$ for unit 1 and $\frac{\sigma^2}{2\sigma^2+1}$ for the other two units.

4 Placebo tests with estimated SC weights

The results from Section 3 show that, even considering an infeasible SC estimator that correctly reconstructs the factor loadings of the treated unit, the placebo tests may have (mild) size distortions. We now consider additional problems that may arise due to the fact that the SC weights are estimated. We consider three possibilities: (i) when the SC estimator is asymptotically biased; (ii) when the transitory shocks are heteroskedastic, and; (iii) when the SC analysis is conditional on a good pre-treatment fit.

For these three cases, we consider MC simulations of a linear factor model in which all units are divided into groups that follow different time trends. In our first DGP, we consider a model with stationary common factors:

$$Y_{j,t}^0 = \delta_t + \lambda_t^k + \epsilon_{j,t} \tag{10}$$

for some $k = 1, ..., K$. We consider the case in which $J+1 = 20$ and $K = 10$. Therefore, units 1 and 2 follow the trend λ_t^1, units 3 and 4 follow the trend λ_t^2, and so on. We consider that λ_t^k is normally distributed with variance equal to one, and we vary the serial correlation of λ_t and the variance of $\epsilon_{j,t} \sim N(0, 0.1)$.

In our second DGP, we modify the linear factor model such that a subset of the common factors is $I(1)$. In this case, we consider DGP which includes a non-stationary trend ϕ_t^r that follows a random walk:

$$Y_{j,t}^0 = \delta_t + \lambda_t^k + \phi_t^r + \epsilon_{jt} \tag{11}$$

for some $k = 1, ..., K$ and $r = 1, ..., R$. We consider in our simulations $K = 10$ and $R = 2$. Therefore, units $j = 2, ..., 10$ follow the same non-stationary path ϕ_t^1 as the treated unit, although only unit $j = 2$ also follows the same stationary path λ_t^1 as the treated unit.

In both models, we impose that there is no treatment effect, i.e., $Y_{j,t} = Y_{j,t}^0 = Y_{j,t}^1$ for each time period $t \in \{1, ..., T_0\}$. We fix the number of post-treatment periods $T - T_0 = 10$ and we vary the number of pre-intervention periods in the DGPs, $T_0 \in \{12, 32, 100, 400\}$.

4.1 SC estimator is asymptotically biased

Once we consider that the SC estimator relies on estimated weights, we have that:

$$
\begin{aligned}
t_i^{\text{ratio}} &= \frac{\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\left[y_{it}-\sum_{j\neq i}\hat{w}_i^j y_{jt}\right]^2}{\frac{1}{T_0}\sum_{t=1}^{T_0}\left[y_{it}-\sum_{j\neq i}\hat{w}_i^j y_{jt}\right]^2} \\
&\xrightarrow{d} \frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\left[\frac{\epsilon_{it}-\sum_{j\neq i}\bar{w}_i^j\epsilon_{jt}+\lambda_t(\mu_i-\sum_{j\neq i}\bar{w}_i^j\mu_j)}{\sqrt{var(\epsilon_{it}-\sum_{j\neq i}\bar{w}_i^j\epsilon_{jt}+\lambda_t(\mu_i-\sum_{j\neq i}\bar{w}_i^j\mu_j))}}\right]^2
\end{aligned}
\tag{12}
$$

where $\widehat{w}_i^j \xrightarrow{p} \bar{w}_i^j$.

Following Ferman and Pinto (2016b), it will generally be the case that $\bar{\mathbf{w}}_i \notin \Phi_i$, so $\mu_i \neq \sum_{j\neq i}\bar{w}_i^j\mu_j$. As a consequence, the SC estimator will be asymptotically biased if the fact that unit 1 was treated after T_0 is informative about the common factors. In this case, since the test statistic depends on the common factors even when $T_0 \to \infty$, the expected value of the test statistic of the treated unit will usually be higher than the expected value of the placebo test statistics, leading to over-rejection.

We explore the implications of the bias of the SC estimator for the placebo tests in the MC simulations described above. We set $var(\epsilon_{it}) = \sigma_\epsilon^2 \in \{0.1, 0.5, 1\}$ and we set λ_t as an AR(1) process with 0.5 serial correlation. Note that the permutation test would work in this case if we were able to use $\mathbf{w}_i^* \in \Phi_i$.[13]

In columns 1 to 3 of Table 1 we present rejection rates in a stationary model when we have that $E[\lambda_t^1|D(1,T_0)=1] = 1$ for $t > T_0$, while in columns 4 to 6 we present rejection rates when $E[\lambda_t^1|D(1,T_0)=1] = 2$ for $t > T_0$. As expected, the placebo test over-rejects the null, as the expected value of the test statistic is higher for the treated unit. Interestingly, we find the largest over-rejection when $\sigma_\epsilon^2 = 0.1$, in which case we found that the misallocation of weights (and, therefore, the asymptotic bias) should be relatively lower. This happens because, while the bias is lower in this case, the variance of the SC estimator is also lower. We present in columns 7 to 12 of Table 1 the same results for the non-stationary model. The placebo test still over-rejects the null, but not as much as in the stationary model. The reason is that the variance of the SC estimator is higher in the non-stationary model, due to the small discrepancy in the factor loadings of the treated and SC units associated with the non-stationary common factor for a fixed T_0. Overall, these results suggest that, when the SC estimator is biased, then the placebo test can over-reject the null even when the bias of the SC estimator is relatively small.

[13] In this case, the infeasible SC estimator is equal to $y_{it} - y_{i't}$, where i' is the pair that follows the same parallel trend as i. Therefore, for all i, the correlation between i and j will be equal to one if j is the pair of j, and zero otherwise.

4.2 Heteroskedasticity with finite and large T_0

We consider next whether heteroskedasticity can generate size distortions in this model. In this case, we consider the same model where $J+1=20$ units are divided into 10 groups of 2 units each, but we set the variance of the transitory shocks of the treated unit equal to 0.1, while the variance of the transitory shocks of the control units is equal to one. We present in column 1 of Table 2 rejection rates when transitory shocks and common factors are serially uncorrelated, using the test statistic proposed in Abadie et al. (2010) (t_i^{ratio}). With $T_0 = 1000$, rejection rate is around 5%. This was expected given that, with serially uncorrelated transitory shocks and common factors, t_i^{ratio} would have the same asymptotic marginal distribution for all placebos.[14] With finite T_0, however, our simulation results suggest that the size distortion can actually be relevant even if the common factors are serially uncorrelated. We over-reject the null when the treated unit has a lower variance. Note that, with a finite T_0, t_i^{ratio} is given by:

$$t_i^{\text{ratio}} = \frac{\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\left[\epsilon_{it} - \sum_{j\neq i}\widehat{w}_i^j\epsilon_{jt} + \lambda_t(\mu_i - \sum_{j\neq i}\widehat{w}_i^j\mu_j)\right]^2}{\frac{1}{T_0}\sum_{t=1}^{T_0}\left[\epsilon_{it} - \sum_{j\neq i}\widehat{w}_i^j\epsilon_{jt} + \lambda_t(\mu_i - \sum_{j\neq i}\widehat{w}_i^j\mu_j)\right]^2} \tag{13}$$

While both numerator and denominator of the test statistic depend on a linear combination of common and transitory shocks, the weights $\widehat{w}_i^j$ are chosen as to minimize the denominator. If T_0 is not large enough relative to J, we might "over-fit" the model. As a consequence, the denominator (in-sample prediction error) would not provide an adequate correction for the variance of the numerator (out-of-sample prediction error), so the marginal distribution of the test statistic would depend on the variance of the treated unit. One possible solution to this problem is to use pre-treatment periods not used in the estimation of the SC weights in the denominator. However, this implies not using all pre-treatment outcome lags as economic predictors exactly when T_0 is small. Also, the variance of the denominator should be large if one leaves out only a few pre-treatment lags, which would imply in a test with low power. Another possible solution might be to avoid over-fitting using a different method to estimate the SC weights that takes into account the fact that the number of parameters might be large relative to the number of pre-treatment periods. Doudchenko and Imbens (2016) consider the use of regularization methods such as best subset regression or LASSO to estimate the SC weights.

We present in column 3 of Table 2 rejection rates when common factors follow an AR(1) process with

[14] Differently from the infeasible SC estimator, the actual SC estimator will not assign 100% of the weight to the pair of the treated unit, even when $T_0 \to \infty$. Therefore, there is no guarantee the the placebo test is asymptotically valid even in this case. Still, our MC simulations suggest that asymptotic size distortions are negligible *for this particular DGP*.

serial correlation equal to 0.9. In this case, the test statistic t_i^{ratio} does not have the same asymptotic marginal distribution for all placebos. The problem is that the test statistic t_i^{ratio} does not properly take into account the serial correlation in the common factors. This implies an over-rejection even when T_0 is large. In this case, an alternative test statistic, $\tilde{t}_i$, that properly corrects the *marginal* distributions of the test statistics when $T_0 \to \infty$ provides rejection rates close to 5% when T_0 is large (column 4 of Table 2). With finite T_0, however, we have over-rejection when the treated unit has a lower variance, whether we use t_i^{ratio} or $\tilde{t}_i$, as in the case with serially uncorrelated common factors. The results using the non-stationary DGP are qualitatively similar (columns 5 to 8 of Table 2).

4.3 Conditional on a good pre-fit

Finally, note that Abadie et al. (2010) and Abadie et al. (2015) suggest that the SC estimator should not be used if the pre-treatment fit is poor. However, when they recommend the placebo test using the t_i^{ratio} test statistic, they suggest that all placebos should be considered. In other words, t_1^{ratio} is conditional on a good pre-treatment fit, while t_i^{ratio} for $i > 1$ is unconditional. This may lead to over-rejection because it will be more likely that the denominator of t_1^{ratio} should be close to zero relative to the denominator of t_i^{ratio}. We evaluate now whether this might generate size distortions. We consider an homoskedastic model in which the SC estimator is asymptotically unbiased (that is, treatment assignment is uncorrelated with common factors). Note that this model is consistent with random assignment of the treated unit. The only difference is that we will only consider simulations in which the pre-treatment fit for the actual SC estimator is good. As a measure of goodness of pre-treatment fit, we consider a pre-treatment normalized mean squared error index, as suggested in Ferman et al. (2016):

$$\widetilde{R}^2 = 1 - \frac{\frac{1}{T_0}\sum_{t=1}^{T_0}\left(y_{1t} - \widehat{y}_{1t}\right)^2}{\frac{1}{T_0}\sum_{t=1}^{T_0}\left(y_{1t} - \overline{y}_1\right)^2} \tag{14}$$

where $\overline{y}_1 = \frac{\sum_{t=1}^{T_0} y_{1t}}{T_0}$. Note that this measure is always lower than one, and it is close to one when the pre-treatment fit is good. If this measure is equal to one, then we have a perfect fit.[15]

We present in Table 3 rejection rates conditional on a good pre-treatment fit for the treated unit. We also present in this table the probability of having a good pre-treatment match. The results suggest that the test may over-reject when the probability of finding a good match is not high. As an extreme example, if we set a threshold for good fit as $\widetilde{R} > 0.9$ and look at the $(T_0, \sigma_\epsilon^2) = (20, 0.1)$ case, then we would have a

[15] Differently from the R^2 measure, this measure can be negative, which would suggest a poor pre-treatment fit.

probability of 13% of having a good pre-treatment fit, and we would have a rejection rate of 10.3% for a 5% test if we consider only SC estimators that provided a good pre-treatment fit. If the probability of having a good fit is close to one (which is usually the case in the non-stationary model), then over-rejection is very mild.

5 Conclusion

We consider the statistical properties of the placebo tests proposed in Abadie et al. (2010). We first show that the graphical analysis based on placebos may be misleading even if we consider an infeasible SC estimator. Then we show that, under some conditions, the placebo test that uses the ratio of the post/pre-intervention MSPE ameliorates the problem as, under some conditions, the test statistics for all placebo runs will have the same asymptotic *marginal* distribution. However, even under such conditions, we show that the test statistics may still have some size distortions. We provide examples in which we can have size distortions even when we consider an infeasible SC estimator that correctly reconstructs the factor loadings of the treated unit. While the size distortions we find in these examples are relatively small, further research is necessary to determine whether there might be examples in which size distortions could be more severe, or whether there is a bound to the size distortions we might have in the SC placebo test, when we consider this infeasible SC estimator. Finally, we show that, once we take into account that the SC weights are estimated, then we can have important size distortions. This will be the case when the SC estimator is asymptotically biased, when we have heteroskedasticity with a finite number of pre-treatment periods, and when we consider that the SC estimator should only be used when there is a close-to-perfect pre-treatment fit.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statiscal Association*, 2010, *105* (490), 493–505.

_ , _ , and _ , "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 2015, *59* (2), 495–510.

_ and Javier Gardeazabal, "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 2003, *93* (1), 113–132.

Ando, Michihito and Fredrik Sävje, "Hypothesis Testing with the Synthetic Control Method," 2013. Working Paper.

Athey, Susan and Guido Imbens, "The State of Applied Econometrics-Causality and Policy Evaluation," *arXiv preprint arXiv:1607.00699*, 2016.

Canay, Ivan A., Joseph P. Romano, and Azeem M. Shaikh, "Randomization Tests under an Approximate Symmetry Assumption?," 2014.

Carvalho, Carlos V., Ricardo Mansini, and Marcelo C. Medeiros, "ArCo: An Artificial Counterfactual Approach for Aggregate Data," February 2015. Working Paper.

_ , _ , and _ , "The Perils of Counterfactual Analysis with Integrated Processes," December 2016. Working Paper.

Doudchenko, Nikolay and Guido Imbens, "Balancing, regression, difference-in-differences and synthetic control methods: A synthesis," 2016.

Ferman, Bruno and Cristine Pinto, "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity," 2016. Working Paper.

_ and _ , "Revisiting the Synthetic Control Estimator," 2016. Working Paper, Available at https://mpra.ub.uni-muenchen.de/77930/1/MPRA_paper_77930.pdf.

_ , _ , and Vitor Possebom, "Cherry Picking with Synthetic Controls," 2016. Working Paper.

Firpo, Sergio and Vitor Possebom, "Synthetic Control Estimator: A Generalized Inference Procedure and Confidence Sets," April 2016.

Fisher, R.A., *The design of experiments. 1935*, Edinburgh: Oliver and Boyd, 1935.

Hahn, Jinyong and Ruoyao Shi, "Synthetic Control and Inference," 2016.

Powell, David, "Synthetic Control Estimation Beyond Case Studies: Does the Minimum Wage Reduce Employment?," 2016. RAND Corporation, WR-1142.

Table 1: **Permutation test with asymptotically biased estimator**

	Stationary model					
	$E[\lambda_t^1 \mid D(1,T_0)=1, t>T_0]=1$			$E[\lambda_t^1 \mid D(1,T_0)=1, t>T_0]=2$		
	$\sigma_\epsilon^2=0.1$	$\sigma_\epsilon^2=0.5$	$\sigma_\epsilon^2=1$	$\sigma_\epsilon^2=0.1$	$\sigma_\epsilon^2=0.5$	$\sigma_\epsilon^2=1$
	(1)	(2)	(3)	(4)	(5)	(6)
$T_0=5$	0.068	0.064	0.062	0.128	0.107	0.099
	[0.001]	[0.001]	[0.001]	[0.002]	[0.002]	[0.001]
$T_0=20$	0.126	0.092	0.082	0.321	0.220	0.182
	[0.002]	[0.001]	[0.001]	[0.002]	[0.002]	[0.002]
$T_0=50$	0.157	0.115	0.100	0.392	0.297	0.243
	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]
$T_0=100$	0.174	0.127	0.109	0.416	0.324	0.270
	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]

	Non-stationary model					
	$E[\lambda_t^1 \mid D(1,T_0)=1, t>T_0]=1$			$E[\lambda_t^1 \mid D(1,T_0)=1, t>T_0]=2$		
	$\sigma_\epsilon^2=0.1$	$\sigma_\epsilon^2=0.5$	$\sigma_\epsilon^2=1$	$\sigma_\epsilon^2=0.1$	$\sigma_\epsilon^2=0.5$	$\sigma_\epsilon^2=1$
	(7)	(8)	(9)	(10)	(11)	(12)
$T_0=5$	0.057	0.056	0.055	0.080	0.072	0.069
	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
$T_0=20$	0.073	0.066	0.063	0.121	0.104	0.096
	[0.001]	[0.001]	[0.001]	[0.002]	[0.002]	[0.001]
$T_0=50$	0.082	0.072	0.068	0.136	0.120	0.110
	[0.001]	[0.001]	[0.001]	[0.002]	[0.002]	[0.002]
$T_0=100$	0.090	0.080	0.075	0.142	0.127	0.118
	[0.001]	[0.001]	[0.001]	[0.002]	[0.002]	[0.002]

Notes: this table presents MC simulations results on a permutation test where the SC estimator is asymptotically biased. Columns 1 to 6 present results for a stationary model, while columns 7 to 12 present results for a model with both non-stationary and stationary common factors. Standard errors in brackets.

Table 2: **Permutation Test with Heteroskeadsticity**

	Stationary model				Non-stationary model			
	without serial correlation		with 0.9 serial correlation		without serial correlation		with 0.9 serial correlation	
	t_i^{ratio}	$\tilde{t}_i$	t_i^{ratio}	$\tilde{t}_i$	t_i^{ratio}	$\tilde{t}_i$	t_i^{ratio}	$\tilde{t}_i$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$T_0 = 5$	0.137	-	0.240	-	0.116	-	0.186	-
	[0.002]	-	[0.003]	-	[0.002]	-	[0.003]	-
$T_0 = 20$	0.089	0.082	0.178	0.158	0.076	0.071	0.136	0.117
	[0.002]	[0.002]	[0.003]	[0.003]	[0.002]	[0.002]	[0.002]	[0.002]
$T_0 = 50$	0.071	0.070	0.129	0.118	0.061	0.058	0.108	0.091
	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]
$T_0 = 100$	0.062	0.063	0.104	0.092	0.057	0.053	0.093	0.080
	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]
$T_0 = 1000$	0.050	0.050	0.071	0.054	0.048	0.048	0.072	0.053
	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]

Notes: this table presents rejection rates when the variance of the transitory shocks for the treated unit is 0.1 while the variance of the transitory shocks for the control unit is 1. Columns 1 and 2 consider the stationary model when the common factor is serially uncorrelated using, respectively, the test statistic suggested in Abadie et al. (2010) and the one suggested in equation 8. Columns 3 and 4 present results when the serial correlation of the common factor is 0.9. Columns 5 to 8 present results for the non-stationary model. It is not possible to calculate $\tilde{t}_i$ with $T_0 = 5$. Standard errors in brackets.

Table 3: **Conditional Permutation Test**

	Stationary model			Non-stationary model		
	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.5$	$\sigma^2_\epsilon = 1$	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.5$	$\sigma^2_\epsilon = 1$
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel i: conditional on $\widetilde{R}^2 > 0.8$					
$T_0 = 5$	0.068	0.099	0.105	0.059	0.081	0.091
	[0.000]	[0.001]	[0.001]	[0.000]	[0.001]	[0.001]
	(0.732)	(0.504)	(0.474)	(0.848)	(0.614)	(0.546)
$T_0 = 20$	0.059	0.222	0.395	0.050	0.064	0.081
	[0.000]	[0.006]	[0.022]	[0.000]	[0.001]	[0.001]
	(0.644)	(0.013)	(0.001)	(0.982)	(0.556)	(0.296)
$T_0 = 50$	0.056	0.240	-	0.050	0.052	0.056
	[0.000]	[0.085]	-	[0.000]	[0.000]	[0.000]
	(0.703)	(0.000)	(0.000)	(1.000)	(0.832)	(0.552)
$T_0 = 100$	0.054	-	-	0.050	0.050	0.052
	[0.000]	-	-	[0.000]	[0.000]	[0.000]
	(0.767)	(0.000)	(0.000)	(1.000)	(0.972)	(0.819)
	Panel ii: conditional on $\widetilde{R}^2 > 0.9$					
$T_0 = 5$	0.101	0.159	0.168	0.074	0.122	0.141
	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
	(0.489)	(0.313)	(0.296)	(0.668)	(0.405)	(0.352)
$T_0 = 20$	0.103	0.302	0.000	0.054	0.086	0.122
	[0.001]	[0.063]	[0.000]	[0.000]	[0.001]	[0.002]
	(0.130)	(0.000)	(0.000)	(0.841)	(0.194)	(0.062)
$T_0 = 50$	0.109	-	-	0.050	0.057	0.063
	[0.003]	-	-	[0.000]	[0.001]	[0.001]
	(0.035)	(0.000)	(0.000)	(0.985)	(0.445)	(0.189)
$T_0 = 100$	0.105	-	-	0.050	0.052	0.054
	[0.007]	-	-	[0.000]	[0.000]	[0.001]
	(0.005)	(0.000)	(0.000)	(1.000)	(0.738)	(0.439)

Notes: this table presents rejection rates conditional on the a good pre-treatment fit for the treated unit. Columns 1 to 3 present results for the stationary model, while columns 4 to 6 present results for the non-stationary model. Panel i defines good pre-treatment fit as a $\widetilde{R}^2 > 0.8$ for the regression of the pre-treatment outcomes of the treated unit on the pre-treatment outcomes of the SC unit. Panel ii defines good pre-treatment fit as $\widetilde{R}^2 > 0.9$. In parenthesis, we present the probability of having a good match. Standard errors in brackets.

A Supplemental Appendix: Placebo Tests for Synthetic Controls

A.1 Permutation test

We now prove that that the test statistic $\tilde{t}_i$ has, asymptotically, the same expected value and variance for all permutations. We have that:

$$\widetilde{t}_i = \frac{\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\left[y_{it} - \sum_{j\neq i}\widehat{w}_i^j y_{jt}\right]^2 - \widehat{E\left[S_i\right]}}{\sqrt{\widehat{Var\left[S_i\right]}}}$$

where $S_i = \frac{1}{T-T_0}\sum_{t=T_0}^{T}\left[\epsilon_{it} - \sum_{j\neq i}\overline{w}_i^j\epsilon_{jt} + \lambda_t\left(\mu_i - \sum_{j\neq i}\overline{w}_i^j\mu_j\right)\right]^2$. We use $T-T_0$ blocks of a combination of pre-treatment variables defined by $\widehat{P_{ik}} = \frac{1}{T-T_0}\sum_{s=k}^{k+T-T_0-1}\left(y_{is} - \sum_{j\neq i}\widehat{w}_i^j y_{js}\right)^2$ for $k = 1, ..., 2T_0 - T$. In this case, the expectation of S_i is estimated by:

$$\widehat{E\left[S_i\right]} = \frac{1}{2T_0 - T}\sum_{k=1}^{2T_0-T}\left[\widehat{P_{ik}}\right]$$

and the estimator of the variance is:

$$\widehat{Var\left[S_i\right]} = \frac{1}{2T_0 - T}\sum_{k=1}^{2T_0-T}\left[\widehat{P_{ik}} - \widehat{E\left[S_i\right]}\right]^2$$

We need to impose the following assumptions. Consider the sequence $\{P_{ik}\}_{k=1}^{2T_0-T}$. We assume that:

1. P_{ik} is a covariance-stationary sequence.
2. P_{ik} is α-mixing with size $-\frac{r}{r-1}$, $r > 4$.
3. $E\left[|P_{ik}|^{r+\delta}\right] < \Delta < 0$ for some $\delta > 0$ at all s.
4. $\frac{1}{T_0}\sum_{s=1}^{T_0} P_{ik}^2 \rightarrow_p E\left[P_{ik}^2\right]$

Lemma 1 *Under assumptions 1-5, and assuming that $\widehat{w}_i^j \rightarrow_p \overline{w}_i^j$, then we have that the expected value of the asymptotic distribution of $\tilde{t}_i$ is equal to zero and the asymptotic variance is equal to 1.*

Proof. *Using the result that* $\widehat{w}_i^j \to_p \overline{w}_i^j$,

$$\frac{1}{2T_0 - T} \sum_{k=1}^{2T_0 - T} \left[\widehat{P_{ik}}\right] = \frac{1}{2T_0 - T} \sum_{k=1}^{2T_0 - T} \left[P_{ik}\right] + o_p\left(1\right)$$

Under assumptions 1-5, and using Corollary 3.48 in White(1999),

$$\frac{1}{2T_0 - T} \sum_{k=1}^{2T_0 - T} \left[\widehat{P_{ik}}\right] \to_p E\left[P_{ik}\right] = E\left[\frac{1}{T - T_0} \sum_{t=T_0}^{T} \left(y_{is} - \sum_{j \neq i} \overline{w}_i^j y_{js}\right)^2\right]$$

Under assumption 2 in the main text,

$$E\left[\frac{1}{T - T_0} \sum_{t=T_0}^{T} \left(y_{is} - \sum_{j \neq i} \overline{w}_i^j y_{js}\right)^2\right] = E\left[\left(y_{is} - \sum_{j \neq i} \overline{w}_i^j y_{js}\right)^2\right]$$

Using the model for y_{is} *and under the condition that* $\sum_{j=2}^{J+1} \overline{w}_i^j = 1$,

$$E\left[\left(y_{is} - \sum_{j \neq i} \overline{w}_i^j y_{js}\right)^2\right] = E\left[\left(\epsilon_{is} - \sum_{j \neq i} \overline{w}_i^j \epsilon_{js} + \lambda_s \left(\mu_{is} - \sum_{j \neq i} \overline{w}_i^j \mu_j\right)\right)^2\right]$$

At the end,

$$\widehat{E\left[S\right]} \to_p E\left[S\right]$$

Using a proof analogous to the lemma above, we can show that $\widehat{Var\left[S\right]} \to_p Var\left[S\right]$.

Therefore:

$$\widetilde{t}_i \to_d \frac{\frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \left(\varepsilon_{it} - \sum_{j \neq i} \overline{w}_i^j \varepsilon_{jt} + \lambda_t \left(\mu_i - \sum_{j \neq i} \overline{w}_i^j \mu_j\right)\right)^2 - E\left[\frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \left(\varepsilon_{it} - \sum_{j \neq i} \overline{w}_i^j \varepsilon_{jt} + \lambda_t \left(\mu_i - \sum_{j \neq i} \overline{w}_i^j \mu_j\right)\right)^2\right]}{\sqrt{Var\left[\frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \left(\varepsilon_{it} - \sum_{j \neq i} \overline{w}_i^j \varepsilon_{jt} + \lambda_t \left(\mu_i - \sum_{j \neq i} \overline{w}_i^j \mu_j\right)\right)^2\right]}}$$

■