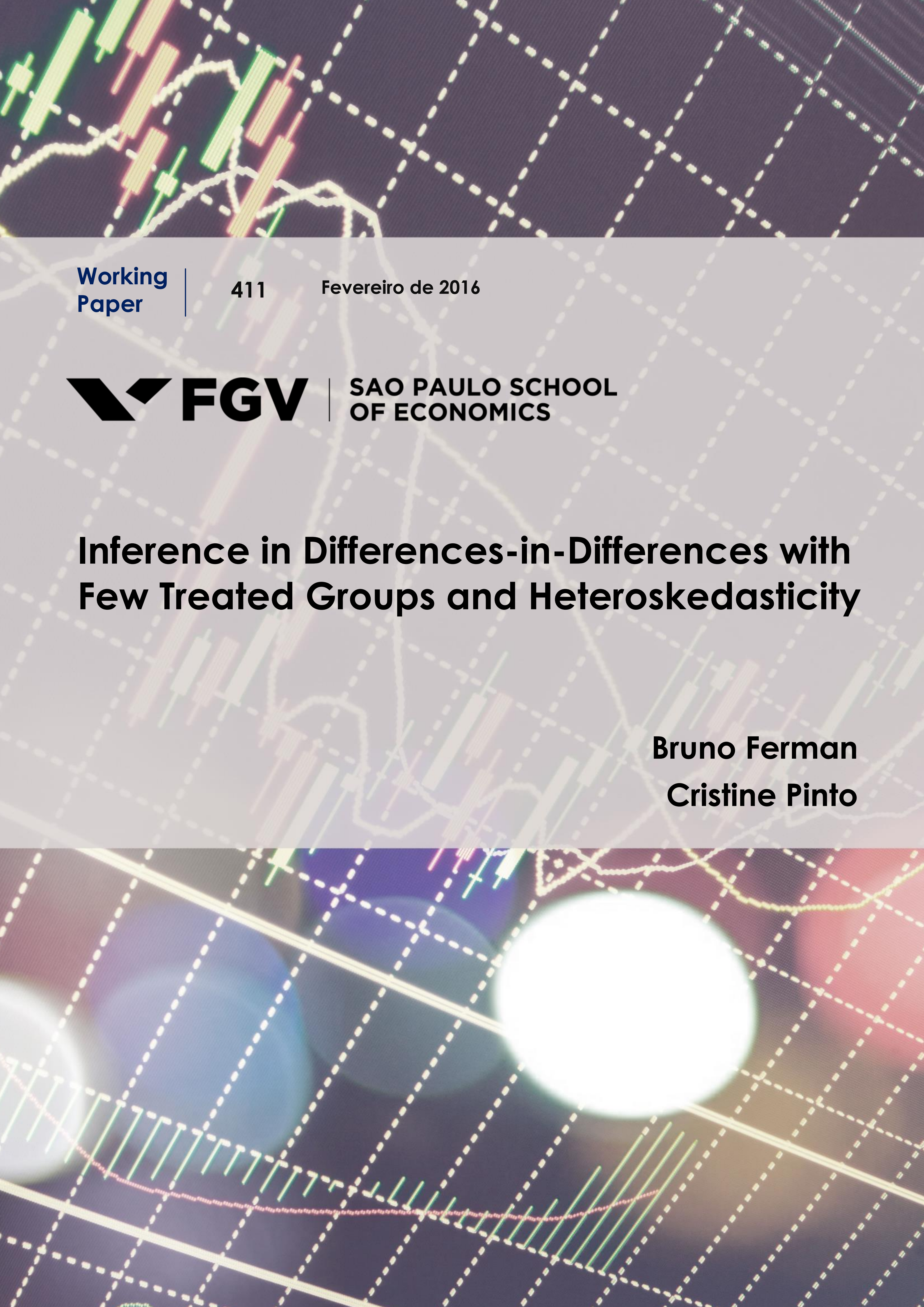


Working Paper | 411 Fevereiro de 2016

FGV | SAO PAULO SCHOOL OF ECONOMICS

Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity

Bruno Ferman

Cristine Pinto

Escola de Economia de São Paulo da Fundação Getulio Vargas FGV-EESP
www.eesp.fgv.br

Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity[*]

Bruno Ferman[†] Cristine Pinto[‡]

Sao Paulo School of Economics - FGV

First Draft: October, 2015

This Draft: May, 2016

Abstract

Differences-in-Differences (DID) is one of the most widely used identification strategies in applied economics. However, how to draw inferences in DID models when there are few treated groups remains an open question. We show that the usual inference methods used in DID models might not perform well when there are few treated groups and errors are heteroskedastic. In particular, we show that when there is variation in the number of observations per group, inference methods designed to work when there are few treated groups tend to (under-) over-reject the null hypothesis when the treated groups are (large) small relative to the control groups. This happens because larger groups tend to have lower variance, generating heteroskedasticity in the group x time aggregate DID model. We provide evidence from Monte Carlo simulations and from placebo DID regressions with the American Community Survey (ACS) and the Current Population Survey (CPS) datasets to show that this problem is relevant even in datasets with large numbers of observations per group. We then derive an alternative inference method that provides accurate hypothesis testing in situations where there are few treated groups (or even just one) and many control groups in the presence of heteroskedasticity. Our method assumes that we can model the heteroskedasticity of a linear combination of the errors. We show that this assumption can be satisfied without imposing strong assumptions on the errors in common DID applications. With many pre-treatment periods, we show that this assumption can be relaxed. Instead, we provide an alternative inference method that relies on strict stationarity and ergodicity of the time series. Finally, we consider two recent alternatives to DID when there are many pre-treatment periods. We extend our inference methods to linear factor models when there are few treated groups. We also derive conditions under which a permutation test for the synthetic control estimator proposed by Abadie et al. (2010) is robust to heteroskedasticity and propose a modification on the test statistic that provided a better heteroskedasticity correction in our simulations.

Keywords: differences-in-differences; inference; heteroskedasticity; clustering; few clusters; bootstrap; randomization inference; synthetic control; linear factor model

JEL Codes: C12; C21; C33

[*] We would like to thank Josh Angrist, Aureo de Paula, Marcelo Fernandes, Sergio Firpo, Bernardo Guimaraes, Michael Leung, Lance Lochner, Marcelo Moreira, Marcelo Medeiros, Whitney Newey, Vladimir Ponczek, Andre Portela, Vitor Possebom, Rodrigo Soares, Chris Taber, Gabriel Ulyssea and seminar participants at Sao Paulo School of Economics - FGV, PUC-Rio, Insper, Latin American Workshop in Econometrics, EPGE - FGV, USP, MIT Econometrics lunch, and Boston University for comments and suggestions.

[†] bruno.ferman@fgv.br

[‡] cristine.pinto@fgv.br

1 Introduction

Differences-in-Differences (DID) is one of the most widely used identification strategies in applied economics. However, inference in DID models is complicated by the fact that residuals might exhibit intra-group and serial correlations. Not taking these problems into account can lead to severe underestimation of the DID standard errors, as highlighted in Bertrand et al. (2004). Still, there is as yet no unified approach to deal with this problem. As stated in Angrist and Pischke (2009), *"... there are a number of ways to do this [deal with the serial correlation problem], not all equally effective in all situations. It seems fair to say that the question of how best to approach the serial correlation problem is currently under study, and a consensus has not yet emerged."*

One of the most common solutions to this problem is to use the cluster-robust variance estimator (CRVE) at the group level.[1] By clustering at the group level, we allow for unrestricted correlation in the within-group errors. More specifically, we allow not only for correlation in the errors of observations in the same group x time, but also for correlation in errors of observations in the same group at different time periods. One important advantage of CRVE is that it also allows for unrestricted heteroskedasticity. The variance of the DID estimator can be divided into two components: one related to the variance of the treated groups and another one related to the variance of the control groups. The CRVE takes heteroskedasticity into account by essentially estimating the variance separately for the treated and for the control groups. Bertrand et al. (2004) show that CRVE and pairs-bootstrap at the group level work well when the number of groups is large.[2] Even when there are only a small number of groups, it might still be possible to obtain tests with correct size even with unrestricted heteroskedasticity (Cameron et al. (2008), Brewer et al. (2013), Imbens and Kolesar (2012), Bell and McCaffrey (2002), Canay et al. (2014), and Ibragimov and Mller (2013)). However, these inference methods will eventually fail when the proportion of treated groups goes to zero or one, even if there are many groups in total (MacKinnon and Webb (2015b)). The problem is that, with a small number of treated groups, it is hard to estimate the variance component related to the treated groups based only on the residuals of the treated group. In the polar case where there is only one treated group, the CRVE estimate of this component of the variance would be identical to zero.[3]

[1] The CRVE was developed by Liang and Zeger (1986). We can think of this method as a generalization of the heterocedasticity-robust variance matrix due to White (1980). In typical applications the label "group" stands for states, counties or countries. More generally, we refer to group as the unit level that is treated. We will assume throughout that errors of individuals within a group can be correlated while errors of individuals in different groups are uncorrelated.

[2] Wooldridge (2003) provides an overview of cluster-sample methods in linear models. The author shows that when the number of groups increases and the groups sizes are fixed, the theory is well developed.

[3] Another alternative presented by Bertrand et al. (2004) is to collapse the pre- and post-information. This approach would take care of the auto-correlation problem. However, in order to allow for heteroskedasticity, one would have to use robust standard errors, in which case this method would also fail when there are few treated groups.

An alternative when there are few treated groups is to use information from the control groups in order to estimate the component of the variance related to the treated groups. Donald and Lang (2007), henceforth DL, deal with the case when the number of treated and control groups is small. They use small sample inference procedures on the group x time DID aggregate model. While working with aggregate data corrects for group x time common shocks, their model requires that errors are normal, homoskedastic and serially uncorrelated.[4] Conley and Taber (2011), henceforth CT, provide an interesting inference method to take both intra-group and serial correlations into account when the number of treated groups is small, but the number of control groups is large. Their method uses information on the residuals of the control groups to estimate the distribution of the DID estimator under the null. Cluster residual bootstrap provides another alternative when there are few treated clusters (Cameron et al. (2008)). In cluster residual bootstrap, we hold the treatment variable constant throughout the pseudo-samples, while resampling the residuals, so that we guarantee that every pseudo-sample will have the same number of treated groups. A crucial assumption for all these methods is that the errors (or a linear combination of the errors) are homoskedastic, so that we can use information on the residuals of the control group to assess the variance of the treated group. However, this homoskedasticity assumption might be very restrictive in DID applications. In particular, errors in the group x time DID aggregate model should be inherently heteroskedastic when there is variation in the numbers of observations used to calculate each group x time average.

In this paper, we first show that usual inference methods used in DID models might not perform well when the number of treated groups is small. Methods that allow for unrestricted heteroskedasticity do not work because they estimate the component of the variance related to the treated groups based on few observations. Also, alternative methods that use information from the control groups will not work properly in the presence of heteroskedasticity. In the particular case in which the number of observations per group varies, these methods tend to (under-) over-reject the null hypothesis when the number of observations in the treated groups is (large) small relative to the number of observations in the control groups. The problem is that variation in the number of observations per group invalidates the homoskedasticity assumption, because larger groups tend to have lower variance. The intuition of this result was already articulated in Assuncao and Ferman (2015) in an application of CT.[5] We formalize this idea and derive conditions under which

[4] Pengyuan et al. (2013) relax the normality assumption for the 2-periods case using a flexible skew-t distribution family to model group time effects. However, their method still requires homoskedasticity.

[5] Assuncao and Ferman (2015) exclude the comparison of placebo estimates when the placebo treated group is much smaller than the original treated group. As stated in Assuncao and Ferman (2015), "*One important caveat with this method [Conley and Taber (2011)] is that the number of observations in each treatment group × year cell in the placebo regressions will not be the same as in the original regression. This is particularly important when the number of observations in the treatment group is small relative to the control group. In this case, increasing the number of observations in the treatment group would*

this problem would be more or less relevant. We then provide evidence from Monte Carlo simulations and simulations with real datasets to show that this problem can be relevant even in datasets with very large numbers of observations per group. This occurs because, as the intra-group correlation approaches zero, increasing the number of observations per group has little impact on the heteroskedasticity. Therefore, a large number of individual observations per group should not be a reasonable justification for the assumption that group x time averages have homoskedastic residuals.

We then derive an alternative method for inference when there are only few treated groups (or even just one) and errors are heteroskedastic. The main assumption is that we know how the heteroskedasticity is generated. Under this assumption, we can re-scale the residuals of the control groups using the (estimated) heteroskedasticity structure in a way that allows us to use this information to estimate the distribution of the error for the treated groups. While our method is more general, this assumption would be satisfied in the particular example in which the heteroskedasticity is generated by variation in the number of observations per group. Also, our method only requires estimation of the heteroskedasticity structure of a linear combination of the errors, which implies that we do not have to impose strong assumptions on the structure of the serial correlation. Therefore, our method is more robust than econometric corrections that place a specific parametric form on the time-series process either to estimate the standard errors or to run a FGLS.[6] Also, our method works even when there are few pre- and/or post-treatment periods. We show that a cluster residual bootstrap with this heteroskedasticity correction provides valid hypothesis testing asymptotically when the number of control groups goes to infinity. Our Monte Carlo simulations and simulations with real datasets suggest that our method provides hypothesis testing with correct sizes when there are around 25 groups in total (1 treated and 24 controls). It is important to note that there is no heteroskedasticity-robust inference method in DID when there is only one treated group. Therefore, although our method is not robust to any form of unknown heteroskedasticity, it provides an important improvement relative to existing methods that rely on homoskedasticity.

With many pre-treatment and a fixed number of post-treatment periods, we can relax the assumption

reduce the variance of the estimator even if we hold the number of observations constant. If this correction is not used, then a placebo estimator using a state with few observations as the treatment group would have a much higher variance than our actual estimator, while a placebo estimator using a large state as the treatment group would tend to underestimate this variance."

[6] Bertrand et al. (2004) show that parametric corrections do not perform well because the coefficient on the auto-correlation parameter is downward biased and because the time-series process might not be correctly specified. Hansen (2007) proposes a bias correction for the auto-correlation estimators. Hausman and Kuersteiner (2008) use a second order expansion to provide a FGLS t-test that takes into consideration the fact that the covariance matrix of the errors has been estimated. Brewer et al. (2013) show that FGLS with Hansen (2007) bias correction combined with robust inference can produce tests with correct sizes even with few groups. However, their approach relies on using the FGLS residuals into the CRVE formula. Therefore, their method would not be appropriate when the proportion of treated groups goes to zero or one, which is the case analyzed in this paper. CT present in their online Appendix an alternative method when there is heteroskedasticity generated by population sizes. Their method, however, requires the full specification of the variance/covariance matrix.

that the structure of the heteroskedasticity is known. If we assume instead that, for each group, errors are strictly stationary and ergodic, then it is possible to apply Andrews (2003) end-of-sample instability test to a transformation of the DID model. This approach works even when there is only one treated and one control group. We also consider two estimation methods that have been recently proposed as alternatives to DID when there are many pre-treatment periods. We consider first the use of linear factor models for estimation of regional policies treatment effects, as suggested by Gobillon and Magnac (2013). We show that our inference methods can be extended to linear factor models when there are only few treated groups. We also consider inference using synthetic control methods (Abadie et al. (2010)). We derive conditions under which a permutation test for the synthetic control estimator proposed by Abadie et al. (2010) is robust to heteroskedasticity. We also propose an alternative test statistic for the permutation test that provided more reliable inference in our simulations.

Our inference method is related to the Randomization Inference (RI) approach proposed by Fisher (1935). In this approach, one uses a permutation test that calculates the test statistic under all possible combinations of treatment assignment, and rejects the null if the observed realization in the actual experiment is extreme enough. The RI approach assumes that treatment assignment is the only stochastic element of the model. In this case, RI provides exact hypothesis testing regardless of the characteristics of the residuals (Lehmann and Romano (2008)). However, a permutation test would not provide valid inference if the assignment mechanism is unknown.[7] Moreover, even under uniform assignment probabilities, a permutation test would only remain valid for *unconditional* tests (that is, before we know which groups were treated). We argue that a permutation test would not provide a satisfactory solution in our setting of few treated groups with heteroskedasticity. As suggested in Mcullagh (1992), unconditional tests can be "*...inappropriate and possibly misleading for hypotheses concerning the data at hand*", while "*...conditional probability calculation often provides an answer to which no right-thinking person could object*".[8] In our setting, once one knows that the treated groups are (large) small relative to the control groups, then one should know that a permutation test that does not take this information into account would (under-) over-reject the null when the null is true.[9] Therefore, such test would not have the correct size conditional on the data at hand. One alternative solution

[7] This would be the case if, for example, larger states are more likely to switch policies. Rosenbaum (2002) proposes a method to estimate the assignment mechanism under selection on observable. However, with few treated groups and many control groups, it would not be possible to reliably estimate this assignment mechanism. Note that it is possible that the DID identification assumptions are valid even when the assignment mechanism is not uniform.

[8] Many authors have recognized the need to make hypothesis testing conditional to the particular data at hand, including Fisher (1934), Pitman (1938), Cox (1958), Cox (1980), Fraser (1968), Cox and Hinkley (1979), Bradley Efron (1978), Barndorff-Nielsen (1980), Barndorff-Nielsen (1983), Barndorff-Nielsen (1984), Hinkley (1980), Mcullagh (1984), Casella and Goutis (1995), and Yates (1984)

[9] This is essentially what happens with CT inference method in this setting. In fact, CT provides an alternative way to implement their method that is "heuristically" motivated by the literature on permutation or randomization inference.

would be to follow Canay et al. (2014). In their setting, they show that it would be possible to incorporate this information if one had functions of the data that have the same limiting distribution under the null hypothesis in all permutations. Canay et al. (2014) and MacKinnon and Webb (2015a) consider RI tests using tests statistics that are asymptotically symmetric. However, their inference methods do not perform well with very few treated groups. In contrast, our method provides a valid correction for heteroskedasticity even when there is only one treated group.

The remainder of this paper proceeds as follows. In Section 2 we present our base model. We briefly explain the necessary assumptions in the existing inference methods, and explain why heteroskedasticity usually invalidates inference methods designed to deal with the case of few treated groups. Then we derive an alternative inference method that corrects for heteroskedasticity even when there is only one treated group. In Section 3 we consider the case with many pre-treatment periods. We first derive an alternative application of our method that relies on a different set of assumptions. Then we extend our inference method to linear factor models with few treated groups. In Section 4 we contrast our inference method with the RI approach, while in Section 5 we derive conditions under which a permutation test for the synthetic control estimator is robust to heteroskedasticity. We perform Monte Carlo simulations to examine the performance of existing inference methods and to compare that to the performance of our method with heteroskedasticity correction in Section 6, while we compare the different inference methods by simulating placebo laws in real datasets in Section 7. We conclude in Section 8.

2 Base Model

2.1 A Review of Existing Methods

We consider a group x time DID aggregate model:[10]

$$Y_{jt} = \alpha d_{jt} + \theta_j + \gamma_t + \eta_{jt} \tag{1}$$

where Y_{jt} represents the outcome of group j at time t; d_{jt} is the policy variable, so α is the main parameter of interest; θ_j is a time-invariant fixed effect for group j, while γ_t is a time fixed-effect; η_{jt} is a group x time error term that might be correlated over time, but uncorrelated across groups. Depending on the application,

[10] The group x time DID aggregate model takes any individual level within group x time cell correlation in the errors into account (DL and Moulton (1986)). However, there might still be correlation of individuals in the same group at different periods in the aggregate model, as suggested by Bertrand et al. (2004).

"groups" might stand for states, counties, countries, and so on. We assume that d_{jt} is nonstochastic.

There are N_1 treated groups and N_0 control groups. Let us assume that d_{jt} changes to 1 for all treated groups starting after date t^*. In this case, the DID estimator will be given by:

$$\begin{aligned}
\hat{\alpha} &= \frac{1}{N_1}\sum_{j=1}^{N_1}\left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} Y_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} Y_{jt}\right] - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} Y_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} Y_{jt}\right] \\
&= \alpha + \frac{1}{N_1}\sum_{j=1}^{N_1}\left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} \eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} \eta_{jt}\right] - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} \eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} \eta_{jt}\right] \\
&= \alpha + \frac{1}{N_1}\sum_{j=1}^{N_1} W_j - \frac{1}{N_0}\sum_{j=N_1+1}^{N} W_j \qquad (2)
\end{aligned}$$

where $W_j = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\eta_{jt}$.

It is clear from equation 2 that consistency of $\hat{\alpha}$ will depend on both $N_1 \rightarrow \infty$ and $N_0 \rightarrow \infty$. As shown in CT, if the number of treated groups (N_1) and the number of periods (T) are fixed, then the DID estimator is unbiased. However, this estimator is not consistent, since the first term, $\frac{1}{N_1}\sum_{j=1}^{N_1} W_j$, would not converge to zero when $N_0 \rightarrow \infty$.

The variance of the DID estimator, under the assumption that η_{jt} are independent across j, is given by:

$$var(\hat{\alpha}) = \left[\frac{1}{N_1}\right]^2 \sum_{j=1}^{N_1} var(W_j) + \left[\frac{1}{N_0}\right]^2 \sum_{j=N_1+1}^{N} var(W_j) \qquad (3)$$

Note that the variance of the DID estimator is the sum of two components: the variance of the treated groups pre/post comparison and the variance of the control groups pre/post comparison. We allow for any kind of correlation between η_{jt} and $\eta_{jt'}$, which is captured in the linear combination of the errors W_j.

When there are many treated and control groups, Bertrand et al. (2004) suggest that CRVE at the group level works well, as this method allows for unrestricted intra-group and serial correlation in the residuals η_{jt}. One important point is that this method is not only cluster-robust, but also heteroskedasticity-robust. The CRVE has a very intuitive formula in the DID framework:[11]

$$\widehat{var(\hat{\alpha})}_{\text{Cluster}} = \left[\frac{1}{N_1}\right]^2 \sum_{j=1}^{N_1} \widehat{W}_j^2 + \left[\frac{1}{N_0}\right]^2 \sum_{j=N_1+1}^{N} \widehat{W}_j^2 \qquad (4)$$

where $\widehat{W}_j = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\hat{\eta}_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\hat{\eta}_{jt}$.

[11] Up to a degrees-of-freedom correction.

With CRVE, we calculate each component of the variance of the DID estimator separately. In other words, we use the residuals of the treated groups to calculate the component related to the treated groups, and the residuals of the control groups to calculate the component related to the control groups. This way, CRVE allows for unrestricted heteroskedasticity. While CRVE is very appealing when there are many treated and many control groups, equation 4 makes it clear why it becomes unappealing when there are few treated groups. In the extreme case when $N_1 = 1$, we will have $\widehat{W}_1^2 = 0$ *by construction.* Therefore, the variance of the DID estimator would be severely underestimated (MacKinnon and Webb (2015b)). The same problem applies to other clustered standard errors corrections such as BRL (Bell and McCaffrey (2002)). It is also problematic to implement heteroskedasticity-robust cluster bootstrap methods such as pairs-bootstrap and wild cluster bootstrap when there are few treated groups. In pairs-bootstrap, there is a high probability that the bootstrap sample will not include a treated unit. Wild cluster bootstrap generates variation in the residuals of each j by randomizing whether its residual will be $\hat{\eta}_{jt}$ or $-\hat{\eta}_{jt}$. However, in the extreme case with only one treated, this leads to $\widehat{W}_1 = 0$. Therefore, the wild cluster bootstrap would not generate variation in the treated group. Another alternative presented by Bertrand et al. (2004) is to collapse the pre- and post-information. This approach would take care of the auto-correlation problem. However, in order to allow for heteroskedasticity, one would have to use heteroskedasticity-robust standard errors. In this case, this method would also fail when there are few treated groups.

It is clear, then, that the inference problem in DID models with few treated groups revolves around how to estimate the component of the DID estimator variance related to the treated group using the residuals $\hat{\eta}_{jt}$. Alternative methods use information on the residuals of the control groups in order to estimate the component of the variance related to the treated groups. These methods, however, rely on specific assumptions regarding the error terms. DL assume that the group x time errors are normal, homoskedastic, and serially uncorrelated. Under these assumptions, the variance of $\hat{\alpha}$ becomes:

$$var(\hat{\alpha}) = \frac{1}{NT}\frac{\sigma_\eta^2}{p(1-p)} \tag{5}$$

where $var(\eta_{jt}) = \sigma_\eta^2$ and p is the proportion of treated groups. Therefore, under these assumptions, one could recover the variance of $\hat{\alpha}$ by estimating σ_η^2 using the $T \times N$ estimated residuals $\hat{\eta}_{jt}$. As suggested by DL, if $T \times N$ is small, then one should compare the test statistic $t = \hat{\alpha}/\sqrt{\widehat{var(\hat{\alpha})}}$ to the student-t distribution instead of calculating the critical values based on the normal distribution. The assumption that errors are serially uncorrelated, however, might be unappealing in DID applications (Bertrand et al. (2004)).

CT provide an interesting alternative inference method that allows for unrestricted auto-correlation in the error terms. Their method uses the residuals of the control groups to estimate the distribution of the DID estimator under the null. The key difference relative to DL is that CT look at a linear combination of the residuals that takes into account any form of serial correlation instead of using the group x time level residuals. In the simpler case with only one treated group, $\hat{\alpha}-\alpha$ would converge to W_1 when $N_0 \rightarrow \infty$. In this case, they use $\{\widehat{W}_j\}_{j=2}^{N_0+1}$ (a linear combination of the control group residuals) to construct the distribution of W_1. While CT relax the assumption of no auto-correlation, it requires that errors are i.i.d. across groups, so that $\{\widehat{W}_j\}_{j=2}^{N_0+1}$ approximates the distribution of W_1 when $N_0 \rightarrow \infty$.

Finally, cluster residual bootstrap methods resample the residuals while holding the regressors constant throughout the pseudo-samples. The residuals are resampled at the group level, so that the correlation structure is preserved. It is possible that a treated group receives the residuals of a control group. While this helps when there are only few treated groups, a crucial assumption is that errors are homoskedastic.

2.2 Leading Example: Variation in Group Sizes

As seen in Section 2.1, CRVE in DID models with few treated groups severely underestimates the variance of $\hat{\alpha}$. Alternative methods such as DL, CT and cluster residual bootstrap require strong distributional assumptions on the errors. In particular, they all require some kind of homoskedasticity in the aggregate group x time model. In this section, we show that group x time DID aggregate models will be inherently heteroskedastic when there is variation in the number of observations per group and derive the implications of this heteroskedasticity for these inference methods. However, it is important to point out that this is not the only example that might generate heteroskedasticity in the group x time aggregate DID model.

We start with a simple individual-level DID model:

$$Y_{ijt} = \alpha d_{jt} + \theta_j + \gamma_t + \nu_{jt} + \epsilon_{ijt} \tag{6}$$

where Y_{ijt} represents the outcome of individual i in group j at time t; ν_{jt} is a group x time error term (possibly correlated over time), and ϵ_{ijt} is an individual-level error term. The main feature that defines a "group" in this setting is the assumption that errors $(\nu_{jt} + \epsilon_{ijt})$ of two individuals in the same group might be correlated, while errors of individuals in different groups are uncorrelated. For ease of exposition, we assume that ϵ_{ijt} are all uncorrelated, while allowing for unrestricted auto-correlation in ν_{jt}. However, our correction will require weaker assumptions on the error structure, as will be presented in Section 2.3.

When we aggregate by group x time, our model becomes the same as the one in equation 1:

$$Y_{jt} = \alpha d_{jt} + \theta_j + \gamma_t + \eta_{jt} \tag{7}$$

The important point is that errors in the group x time aggregate model (η_{jt}) are heteroskedastic across j, unless $M(j,t)$ is constant across j. More specifically:

$$\eta_{jt} = \nu_{jt} + \frac{1}{M(j,t)} \sum_{i=1}^{M(j,t)} \epsilon_{ijt} \tag{8}$$

where $M(j,t)$ is the number of observations in group t at time t. Therefore, assuming for simplicity that $M(j,t) = M_j$ is constant across j and T is fixed, we have that the variance of W_j conditional on M_j is given by:

$$\begin{aligned} var(W_j|M_j) &= var\left(\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} \eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\eta_{jt}\right) \\ &= var\left(\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} \nu_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\nu_{jt} + \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\left[\frac{1}{M_j}\sum_{i=1}^{M_j}\epsilon_{ijt}\right] - \frac{1}{t^*}\sum_{t=1}^{t^*}\left[\frac{1}{M_j}\sum_{i=1}^{M_j}\epsilon_{ijt}\right]\right) = \\ &= A + \frac{B}{M_j} \end{aligned} \tag{9}$$

for constants A and B, regardless of the auto-correlation of ν_{jt}.[12]

We are assuming so far that we have a panel of repeated cross-sections, so that ϵ_{ijt} are not correlated over time. If we had a panel and allow for the individual-level residuals to be auto-correlated, then we would have another term that would depend on the ϵ_{ijt} auto-correlation parameter divided by the number of observations, so we would still end up with the same formula, $var(W_j|M_j) = A + \frac{B}{M_j}$. This formula might also remain valid even in situations where, for example, the correlation between two observations in the same school is stronger than the correlation between two observations in the same state but in different schools.[13] Finally, we do not need to assume that the individual-level model is homoskedastic to have the formula $var(W_j|M_j) = A + \frac{B}{M_j}$.

[12] When the number of observations per group is not constant over time, the formula will be: $var(W_j) = \tilde{A} + \tilde{B}\left[\left(\frac{1}{T-t^*}\right)^2 \sum_{t=t^*+1}^{T} \frac{1}{M(j,t)} + \left(\frac{1}{t^*}\right)^2 \sum_{t=1}^{t^*} \frac{1}{M(j,t)}\right]$, for constants $\tilde{A}$ and $\tilde{B}$.

[13] This would be the case, for example, in a model $Y_{ikjt} = \alpha d_{jt} + \theta_j + \gamma_t + \nu_{jt} + \omega_{kjt} + \epsilon_{ikjt}$ for student i in school k, state j and time t. We allow for a common school-level shock ω_{kjt} in addition to the state-level shock ν_{jt}. If the number of schools grow at the same rate as the number of observations, then this model would also generate $var(W_j|M_j) = A + \frac{B}{M_j}$.

This heteroskedasticity in the error terms of the aggregate model implies that, when the number of observations in the treated groups are (large) small relative to the number of observations in the control groups, we would (over-) underestimate the component of the variance related to the treated group when we estimate it using information from the control groups. This implies that inference methods that do not take that into account would tend to (under-) over-reject the null hypothesis when the number of observations of the treated groups is (large) small.

Note that, if $A > 0$, this would not be a problem when $M(j,t) \to \infty$. In this case, $var(W_j) \to A$ for all j. In other words, when the number of observations in each group x cell is large, then the correlated part of the error would dominate. In this case, if we assume that the group x time error ν_{jt} is i.i.d., then $\frac{var(W_j)}{var(W_j')} \to 1$, which implies that the residuals of the control groups would be a good approximation for the distribution of the treated groups error even when the number of observations in each group is different. This is one of the main rationales used in DL to justify the homoskedasticity assumption in the aggregate model.

However, an interesting case occurs when $A = 0$. In this case, even though $var(W_j) \to 0$ for all j when $M_j \to \infty$, the ratios $\frac{var(W_j)}{var(W_{j'})}$ remain constant (unless $\frac{M_j}{M_{j'}} \to 1$), which implies that the aggregate model remains heteroskedastic even asymptotically. Therefore, CT, DL and cluster residual bootstrap would tend to (under-) over-reject the null hypothesis when the number of observations of the treated groups are (large) small relative to the number of observations of the control groups even when there is a large number of individual observations. While $A \approx 0$ implies that the group x time common shock is not strong, note that this does not mean that standard inference using OLS regression on the individual-level data would be reliable. Even in the absence of group x time common shocks, OLS standard errors might be underestimated if there is important serial correlation in the individual level error, ϵ_{ijt}.

2.3 Inference with Heteroskedasticity Correction

As discussed in Section 2.1, the main challenge in estimating the variance of $\hat{\alpha}$ when there are few treated groups is how to estimate the component related to the treated groups. The CRVE estimates this component of the variance without using information from the control groups. While this approach has the appealing property of allowing for unrestricted heteroskedasticity, it is unfeasible when the number of treated groups is small. On the other extreme, other methods surpass the problem of few treated groups by using information from the control groups. The problem with these approaches is that they require homoskeadsticity.

In this section, we derive an inference method that uses information from the control groups to estimate the variance of the treated groups while still allowing for heteroskedasticity. Intuitively, our approach assumes that we know how the heteroskedasticity is generated, which is the case when, for example, heteroskedasticity is generated by variation in the number of observations per group. Under this assumption, we can re-scale the residuals of the control groups using the (estimated) structure of the heteroskedasticity in a way that allows us to use this information to estimate the distribution of the error for the treated groups. Our method only requires information on the heteroskedasticity structure for a linear combination of the errors, which implies that we do not have to impose strong assumptions on the structure of the serial correlation of the errors. While we motivate our methods based on heteroskedasticity generated by variation in the number of groups, it is important to note that our method is more general.

More formally, we assume we have a total of N groups where the first N_1 groups are treated. For simplicity, we consider first the case where d_{jt} changes to 1 for all treated groups starting after date t^*. Let X_j be a vector of observable covariates that not necessarily enter in model 1 and d_j be an indicator variable equal to 1 if group j is treated. We will define our assumptions directly on the linear combination of the errors $W_j = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\eta_{jt}$. The main assumptions for our method are:

1. $\{W_j, X_j\}$ is i.i.d. across $j \in \{1, ..., N_1\}$, i.i.d. across $j \in \{N_1+1, ..., N\}$ and independently distributed across $j \in \{1, ..., N\}$.

2. $W_j|X_j, d_j \overset{d}{=} W_j|\tilde{X}_j$, where $\tilde{X}_j$ is a subset of X_j.

3. $W_j|\tilde{X}_j$ has the same distribution across $\tilde{X}_j$ up to a scale parameter. That is, $\frac{W_j}{\sqrt{var(W_j|\tilde{X}_j)}}|\tilde{X}_j$ does not depend on X_j.

4. $E[W_j|X_j, d_j] = E[W_j|X_j] = 0$

Note that assumption 1 does not require that treatment is randomly assignment. Instead, we might consider a case where, for example, larger states are more likely to switch policies than smaller states. Assumption 2 implies that, conditional on a subset of observable covariates, the distribution of W_j will be the same independently of the treatment status. This is crucial for our method, as it guarantees that we can extrapolate information from the control groups residuals to estimate the distribution of the treated groups errors. This assumption would not be required with large N_1 and N_0 for inference with heteroskedasticity-robust methods. In this case, it would be possible to allow for different distributions conditional on treatment status since there would be enough observations to estimate the variance component related to the treated groups using only information from the treated groups. In our setting, this would not be feasible since we assume that the number of treated groups is fixed and small. Assumption 3 implies that the distribution of $W_j|X_j$ only depends on X_j through the variance parameter. This assumption reduces the dimensionality of the problem. It would be possible to relax this assumption and estimate the conditional distribution of $W_j|\tilde{X}_j$ non-parametrically. However, this would require very large number of control groups, N_0. Finally, condition 4 is the standard identification assumption for DID.

Our method is an extension of the cluster residual bootstrap with H_0 imposed where we correct the residuals for heteroskedasticity. In cluster residual bootstrap with H_0 imposed, we estimate the DID regression imposing that $\alpha = 0$, generating the residuals $\{\widehat{W}_j^R\}_{i=1}^N$. If the errors are homoskedastic, then, under the null, $\widehat{W}_j^R$ would have the same distribution across j, which implies that we can resample with replacement $\mathcal{B}$ times from $\{\widehat{W}_j^R\}_{i=1}^N$, generating $\{\widehat{W}_{j,b}^R\}_{i=1}^N$. Then we can calculate our bootstrap estimates as $\hat{\alpha}_b = \frac{1}{N_1}\sum_{j=1}^{N_1}\widehat{W}_{j,b}^R - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\widehat{W}_{j,b}^R$. Note that, in our setting, we do not need to work with the group x time residuals $\hat{\eta}_{jt}$ to construct our bootstrap estimates. Instead, we can work with a linear combination of the residuals that takes into account any form of auto-correlation in the residuals. This is one of the key insights of CT.

As explained in Section 2.1, the problem with cluster residual bootstrap is that it requires the residuals to be homoskedastic. In Theorem 1 in Appendix A.1, we show that, if we know the variance of each random variable W_j conditional on X_j, then we can re-scale the residuals $\widehat{W}_{j,b}^R$ and use a cluster residual bootstrap on the re-scaled residuals even if the model is heteroskedastic. First, we normalize each observed $\widehat{W}_{j'}$ by $\widehat{W}_{j'}^{norm} = \widehat{W}_{j'}^R \frac{1}{\sqrt{var(W_{j'}|X_{j'})}}$. Then we generate a bootstrap sample with the re-scaled residuals $\widetilde{W}_{j,b} = \widehat{W}_{j,b}^{norm}\sqrt{var(W_j|X_j)}$. As a result, this procedure generates bootstrap estimators $\hat{\alpha}_b = \frac{1}{N_1}\sum_{j=1}^{N_1}\widetilde{W}_{j,b} - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\widetilde{W}_{j,b}$ with the same distribution as the DID estimator.[14] The main assumption we need is that $\{W_j\}_{j=1}^N$, which is a linear combination of the error terms η_{jt}, are independent across j and have the same distribution up to the variance parameter.[15] It is important to note that we only need the variance of a linear combination of the errors. This point is crucial for our method, because we do not need to know the serial correlation structure of the errors η_{jt}.

The main problem, however, is that $var(W_j|X_j)$ is generally unknown, so it needs to be estimated. In Theorem 2 in Appendix A.1, we show that this heteroskedasticity correction works asymptotically when

[14] As we assume a setting in which the number of treated groups is fixed and small, we have to consider for inference the distribution of $\hat{\alpha}$ conditional on $\{X_j\}_{j=1}^N$. Note that CT would be valid as unconditional inference if we assume that $\{W_j, X_j\}$ is i.i.d. across $j \in \{1, ..., N\}$. However, CT would not provide a reasonable solution conditional on the data at hand. As we show in our example in Section 2.2, CT would provide a biased test conditional on the information about group sizes. If $\{W_j, X_j\}$ is not identically distributed (as is allowed in assumption 1), then it would be impossible to consistently estimate $var(W_j|d_j = 1)$ because we would only have a finite number of treated observations (unless we have more information about the distribution of $X_j|d_j$). Therefore, it would not be possible to conduct unconditional inference. In Section 4, we discuss in more detail the case for conditional inference in the RI framework.

[15] Note that this assumption is weaker than assuming that the sequences $\{\eta_{j1}, ..., \eta_{jT}\}$ are independent and have the same distribution up to a variance parameter across j.

$N_0 \to \infty$ if we have a consistent estimator for $var(W_j|X_j)$. That is, we can use $var\widehat{(W_j|X_j)}$ to generate $\widetilde{\widehat{W}}_{j,b} = \widehat{W}^R_{j,b}\sqrt{\frac{var\widehat{(W_j|X_j)}}{var\widehat{(W_{j,b}|X_{j,b})}}}$. In theory, one could estimate the conditional variance function non-parametrically. In practice, however, a non-parametric estimator would likely require a large number of control groups.

In our leading example where heteroskedasticity is generated by variation in group sizes, we show in Section 2.2 that we can derive a parsimonious function for the conditional variance without having to impose a strong structure on the error terms. More specifically, in this example, the conditional variance function would be given by $var(W_j|X_j, d_j) = var(W_j|M_j) = A + \frac{B}{M_j}$, for constants A and B, where X_j is the set of observable variables which includes M_j. We show in Theorem 3 in Appendix A.1 that we can get a consistent estimator for $var(W_j|M_j)$ by regressing $\hat{W}_j^2$ on $\frac{1}{M_j}$ and a constant.[16] Note that we do not need individual-level data to apply this method, provided that we have information on the number of observations that were used to calculate the group x time averages.

Finally, a problem with cluster bootstrap methods when there are few clusters is that there will be few possible combinations of bootstrap samples (Cameron et al. (2008), Webb (2014), and MacKinnon and Webb (2015a)). To ameliorate this problem, we apply the idea of wild cluster bootstrap to our method. Therefore, for each j, we sample either $\widetilde{\widehat{W}}_{j,b}$ with probability 0.5 or $-\widetilde{\widehat{W}}_{j,b}$ with probability 0.5. This procedure provides a smoother bootstrap distribution.

Summarizing, our bootstrap procedure, for this specific case, consists of:

1. Calculate the DID estimate:

$$\hat{\alpha} = \frac{1}{N_1}\sum_{j=1}^{N_1}\left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} Y_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} Y_{jt}\right] - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} Y_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} Y_{jt}\right]$$

2. Estimate the DID model with H_0 imposed ($Y_{jt} = \alpha_0 d_{jt} + \theta_j + \gamma_t + \eta_{jt}$), and obtain $\{\widehat{W}_j^R\}_{i=1}^N$. Usually the null will be $\alpha_0 = 0$.

3. Estimate $var(W_j|M_j)$ by regressing $\left(\widehat{W}_j^R\right)^2$ on a constant and $\frac{1}{M_j}$.

4. Use $var\widehat{(W_j|M_j)}$ to obtain the normalized residuals $\widehat{W}_{j'}^{norm} = \widehat{W}_{j'}^R \frac{1}{\sqrt{var\widehat{(W_{j'}|M_{j'})}}}$

5. Do $\mathcal{B}$ iterations of this step. On the b^{th} iteration:

[16] When the number of observations per group is not constant over time, we regress $\widehat{W}_j^2$ on $\left[\left(\frac{1}{T-t^*}\right)^2 \sum_{t=t^*+1}^{T} \frac{1}{M(j,t)} + \left(\frac{1}{t^*}\right)^2 \sum_{t=1}^{t^*} \frac{1}{M(j,t)}\right]$ and a constant.

(a) Resample with replacement N times from $\{\widehat{W}_j^{norm}\}_{i=1}^N$ to obtain $\left\{\widetilde{\widehat{W}}_{j,b}\right\}_{i=1}^N$, where $\widetilde{\widehat{W}}_{j,b} = \widehat{W}_{j,b}^{norm}\sqrt{\widehat{var(W_j|M_j)}}$ with probability 0.5 and $-\widehat{W}_{j,b}^{norm}\sqrt{\widehat{var(W_j|M_j)}}$ with probability 0.5.

(b) Calculate $\hat{\alpha}_b = \frac{1}{N_1}\sum_{j=1}^{N_1}\widetilde{\widehat{W}}_{j,b} - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\widetilde{\widehat{W}}_{j,b}$.

6. Reject H_0 at level a if and only if $\hat{\alpha} < \hat{\alpha}_b[a/2]$ or $\hat{\alpha} > \hat{\alpha}_b[1-a/2]$, where $\hat{\alpha}_b[q]$ denotes the q^{th} quantile of $\hat{\alpha}_1, ..., \hat{\alpha}_{\mathcal{B}}$.

The method described above works when all the treated groups start treatment in the same period t^*. Consider a general case where there are N_0 control groups and N_k treated groups that start treatment after period $t_k^\star$, with $k = 1, ..., K$. We show in Appendix A.2 that, for large N_0, the DID estimator is a weighted average of K DID estimators, each one using one $k > 0$ as treated groups and $k = 0$ as control groups. The weights are given by $\frac{N_k(T-t_k^\star)t_k^\star}{\sum_{k=1}^K N_k(T-t_k^\star)t_k^\star}$. Therefore, the weights are increasing in the number of treated groups that start treatment after $t_k^\star$ (N_k) and are higher when $t_k^\star$ divides the total number of periods in half. Define $\widehat{W}_j^{R,k} = \frac{1}{T-t_k^*}\sum_{t=t_k^*+1}^T \hat{\eta}_{jt}^R - \frac{1}{t_k^*}\sum_{t=1}^{t_k^*}\hat{\eta}_{jt}^R$. We generalize our method to this case by estimating K functions $\widehat{var(W_j^k|M_j)}$ by regressing $(\widehat{W}_j^{R,k})^2$ on a constant and $\frac{1}{M_j}$. Each function $\widehat{var(W_j^k|M_j)}$ provides the proper rescale for the residuals of the DID regression using k as the treated groups. Then we calculate $\hat{\alpha}_b$ as a weighted average of these K DID estimators.

We also show in Appendix A.3 that our method applies to DID models with covariates. With group x time level covariates, we estimate the OLS DID regressions in steps 1 and 2 of the bootstrap procedure with covariates. The other steps remain the same. If we have individual-level data, then we run the individual-level OLS regression with covariates in step 2 and then aggregate the residuals of this regression at the group x time level. The other steps in the bootstrap procedure remain the same.

It is important to note that our inference procedure is closely related to Generalized Least Squares (GLS) in that we assume a structure for the heteroskedasticity. There are, however, two important differences. First, we only assume a structure for the heteroskedasticity of a linear combination of the errors, W_j, instead of assuming a structure on the full variance/covariance matrix of the group x time level errors, η_{jt}. This is a crucial component of our method since working with W_j circumvents the problem of specifying the serial correlation structure of the errors. Second, since we are considering the case with a fixed number of treated groups, we cannot rely on asymptotic approximations. More specifically, the asymptotic distribution of $\hat{\alpha}$ when the number of control groups grows would depend on the distribution of the group x time aggregate shocks, ν_{jt}, which will not necessarily be normally distributed. Our bootstrap method allows us to calculate critical values without relying on asymptotic normality.

Finally, it might be possible to use the estimated structure of the heteroskedasticity to derive a more efficient estimator. Note that one of the main advantages of our method is that we do not need to specify the full variance/covariance structure of η_{jt}, which implies that we cannot construct a "full" GLS estimator. Still, we can use the estimated heteroskedasticity structure of W_j to improve efficiency. Note that $\hat{\alpha} = \frac{1}{N_1}\sum_{j=1}^{N_1} \bigtriangledown Y_j - \frac{1}{N_0}\sum_{j=N_1+1}^{N} \bigtriangledown Y_j$, where $\bigtriangledown Y_j = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T} Y_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} Y_{jt}$.[17] Given the estimated structure of the variance of W_j, we could construct a more efficient estimator by changing the weights used to calculate the treated and the control averages:

$$\hat{\alpha}' = \sum_{j=1}^{N_1} \omega_j \bigtriangledown Y_j - \sum_{j=N_1+1}^{N} \omega_j \bigtriangledown Y_j \tag{10}$$

where $\omega_j = \frac{\frac{1}{\sqrt{var\widehat{(W_j|X_j)}}}}{\sum_{j=1}^{N_1}\sqrt{\frac{1}{var\widehat{(W_j|X_j)}}}}$ if j is treated and $\omega_j = \frac{\frac{1}{\sqrt{var\widehat{(W_j|X_j)}}}}{\sum_{j=N_1+1}^{N}\sqrt{\frac{1}{var\widehat{(W_j|X_j)}}}}$ if j is control. The only difference in our inference procedure is that one should calculate the bootstrap estimates in step 5.b using this formula and then compare the estimate $\hat{\alpha}'$ to the bootstrap distribution. Note that with only one treated group and many control groups, there should not be much difference relative to the original DID estimator. In this case, both estimators would converge in distribution to $\alpha + W_1$. However, with more than one treated group this adjustment can be more relevant.

3 Heteroskedasticity Correction with Large t^*

3.1 DID with Large t^*

As explained in Section 2.3, an important assumption of our method is that we need to know how the heteroskedasticity was generated. We now show that this assumption can be relaxed if we have a large number of pre-treatment periods and a small number of post-treatment periods. Under strict stationarity and ergodicity of the time series, we can apply Andrews (2003) end-of-sample instability test to a transformation of the DID model. The main idea is that with large $t^\star$ and small $T - t^\star$ the DID estimator would converge in distribution to a linear combination of the post-treatment errors. Therefore, under strict stationarity and ergodicity, we can use blocks of the pre-treatment periods to estimate the distribution of $\hat{\alpha}$. This is essentially the idea of CT method but exploiting the time instead of the cross-section variation.

[17] If we consider the individual-level DID estimator, then the averages for the treated and control groups would be weighted by the sample size of each group.

If we collapse the cross-section variation using the transformation $\tilde{Y}_t = \frac{1}{N_1}\sum_{j=1}^{N_1} Y_{jt} - \frac{1}{N_0}\sum_{j=N_1+1}^{N} Y_{jt}$, then:

$$\tilde{Y}_t = \begin{cases} \tilde{\theta} + \tilde{\eta}_t, \text{ for } t = 1, ..., t^* \\ \alpha + \tilde{\theta} + \tilde{\eta}_t, \text{ for } t = t^* + 1, ..., T \end{cases} \tag{11}$$

where $\tilde{\theta} = \frac{1}{N_1}\sum_{j=1}^{N_1} \theta_j - \frac{1}{N_0}\sum_{j=N_1+1}^{N} \theta_t$ and $\tilde{\eta}_t = \frac{1}{N_1}\sum_{j=1}^{N_1} \eta_{jt} - \frac{1}{N_0}\sum_{j=N_1+1}^{N} \eta_{jt}$.

Therefore, this is a particular case of Andrews (2003) end-of-sample instability test in a model that includes only a constant.[18] We want to test whether the average of $\tilde{Y}_t$ is different after the treatment. With group-level covariates, we can estimate the OLS DID model and then work construct $\tilde{Y}_t$ using $Y_{jt} - X'_{jt}\hat{\beta}$. Since $\hat{\beta}$ is consistent, this approach will work under strict stationarity and ergodicity of η_{jt}. The same approach works if we have individual-level covariates.[19]

This approach is interesting because we do not need to assume the structure of the heteroskedasticity. Also, this approach works even if we have as few as one treated and one control group. However, this approach is unfeasible if there are few pre-treatment periods. Moreover, the stationarity assumption might be violated if, for example, there is variation in the number of observations per group across time. For example, if we divide the US states in the CPS by quartiles of number of observations for each year from 1979 to 2014, then 35 out of the 51 states belonged to 3 or 4 different quartiles depending on the survey year. In this scenario, our method using the function $\widehat{var(W_j|M_j)}$ would still provide a valid alternative, provided that we have a large number of control groups and we know how the heteroskedasticity was generated.

3.2 Linear Factor Model

We now consider inference in linear factor model, an estimation method that has been studied in the panel data setting in Bai (2009) and analyzed in detail for estimation of treatment effects of regional policies as a generalization of DID in Gobillon and Magnac (2013). We show that the inference methods we propose can be expanded to linear factor models with few treated groups.

[18] Note that the DID estimator would be given by $\hat{\alpha} = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T} \tilde{Y}_t - \frac{1}{t^*}\sum_{t=1}^{t^*} \tilde{Y}_t$.

[19] With group-level covariates, we consider a model $Y_{jt} = \alpha d_{jt} + X'_{jt}\beta + \theta_j + \gamma_t + \eta_{jt}$. With individual-level covariates, we consider a model $Y_{ijt} = \alpha d_{jt} + X'_{ijt}\beta + \theta_j + \gamma_t + \nu_{jt} + \epsilon_{ijt}$. In this case, we have to impose the strict stationarity and ergodicity assumptions on $\eta_{jt} = \nu_{jt} + \frac{1}{M(j,t)}\sum_{i=1}^{M(j,t)} \epsilon_{ijt}$.

Gobillon and Magnac (2013) consider a model in which the potential outcome in the absence of treatment is given by:

$$Y_{jt}(0) = x_{jt}\beta + f_i'\lambda_i + \eta_{jt}^{LFM} \tag{12}$$

where x_{jt} are covariates, λ_i is a $L \times 1$ vector of individual effects or *factor loadings*, and f_t is a $L \times 1$ vector of time effects or *factors*. The treatment effect is given by α_{jt}, so that:

$$Y_{jt}(1) = Y_{jt}(0) + \alpha_{jt} \tag{13}$$

This model allows for more flexibility relative to the usual DID model. As shown in Gobillon and Magnac (2013), we can go back to the usual DID model by setting the restrictions $\lambda_i = (\theta_i, 1)'$ and $f_t = (1, \gamma_t)'$. They assume that we know the number of factors in the true DGP and that the factors are sufficiently strong so that the consistency condition for factores and factor loadings is satisfied.

As suggested in Gobillon and Magnac (2013), it is possible to estimate this model in two steps. In the first step, we estimate the linear factor model in equation 12 using the sample composed of non treated observations over the whole period and of treated observations in the pre-treatment ($t \leq t^*$). If t^* and N_0 tend to ∞, then we get consistent estimators for β, f_t and λ_t. In the second step, we estimate the counterfactual term imputing the estimated β, f_t and λ_t. More specifically, we have that the average treatment on the treated effect in period t is given by:

$$E[Y_{jt}(1) - Y_{jt}(0)| \text{ treated }] = E[\alpha_{jt}| \text{ treated }] = E[Y_{jt} - x_{jt}\beta - \lambda_i' f_t| \text{ treated }] \tag{14}$$

Therefore, we can use the empirical counterpart $\hat{\alpha}_t = \frac{1}{N_1}\sum_{j=1}^{N_1}\left[Y_{jt} - x_{jt}\hat{\beta} - \hat{\lambda}_i'\hat{f}_t\right]$ to estimate $E[\alpha_{jt}| \text{ treated }]$. If we let N_0 and t^* go to ∞ while N_1 is fixed, then:

$$\begin{aligned} \hat{\alpha}_t &= \frac{1}{N_1}\sum_{j=1}^{N_1}\left[Y_{jt} - x_{jt}\hat{\beta} - \hat{\lambda}_i'\hat{f}_t\right] \rightarrow \frac{1}{N_1}\sum_{j=1}^{N_1}[Y_{jt} - x_{jt}\beta - \lambda_i' f_t] = \\ &= E[\alpha_{jt}| \text{ treated }] + \frac{1}{N_1}\sum_{j=1}^{N_1}\eta_{jt}^{LFM} \end{aligned} \tag{15}$$

If we want to estimate the average treatment on the treated as defined in Gobillon and Magnac (2013), we just need to use $\hat{\alpha} = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\hat{\alpha}_t$. As N_0 and t^* go to ∞ while N_1 and $T - t^*$ are fixed, $\hat{\alpha} - \alpha$ will converge to $\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\left[\frac{1}{N_1}\sum_{j=1}^{N_1}\eta_{jt}^{LFM}\right]$. In other words, with fixed N_1 and fixed $T-t^*$ the error of the linear factor model estimator will be dominated by the error of the treated groups.

This result is a natural extension of CT.[20] The key point is that factors and factor loads are consistently estimated, so we can use the residuals from the linear factor model $\hat{\eta}_{jt}^{LFP}$ to estimate the distribution of η_{jt}^{LFM}. Since we have both $t^* \to \infty$ and $N_0 \to \infty$, we have two alternatives in this case. We can exploit the cross-section variation using the estimated residuals from the control groups, $\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\hat{\eta}_{jt}^{LFM}$ for $j > N_1$, to approximate the distribution of the errors of the treated groups, $\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\eta_{jt}^{LFM}$ for $j \leq N_1$. Under homoskedasticity across j, this is essentially CT method applied to linear factor models with few treated groups. If errors are heteroskedastic, then we can use our method, provided that we know how the heteroskedasticity was generated. Alternatively, we can exploit the time series variation as shown in Section 3 provided that η_{jt}^{LFM} is strictly stationary and ergodic.

4 Randomization Inference and Permutation Tests

We assume in our model that treatment assignment is nonstochastic, while the stochastic elements in the model come from η_{jt}, ν_{jt}, and ϵ_{ijt}. This departs crucially from Randomization Inference (RI), which considers that the only stochastic component of the model is the treatment assignment (Fisher (1935)). In RI, one calculates the test statistic under all possible combinations of treatment assignment, and rejects the null if the observed realization in the actual experiment is extreme enough. This idea is closely related to CT. In fact, CT propose an alternative way to implement their method which is *heuristically* motivated by the literature on permutation tests and RI. As stated in Lehmann and Romano (2008), RI provides exact test statistics based solely on the null of no treatment effects and the fact that treatment was randomly assigned, not depending on any assumption regarding the characteristics of outcome, covariates and residuals. Young (2015) argues that many published papers in Economics that use standard inference methods in randomized experiments produce invalid testing, and proposes the use of RI methods. While we agree that RI provides a powerful inference method in randomized experiments, we believe this approach would not provide satisfactory inference method in our setting.

First, permutation tests should take into account the treatment assignment mechanism. In standard permutation tests where one considers all possible combinations of treatment/control status, it is implicitly assumed that all units had uniform assignment probability. This assumption guarantees that the random-

[20] Note that we get an equivalent formula in the DID model if we let N_0 and t^* go to ∞ while N_1 and $T-t^*$ are fixed.

ization distribution would provide the exact distribution of the test statistic (for an *unconditional* test). However, this assumption might be unreasonable in many DID applications. For example, one might believe that larger states switch policies more often than smaller states. Note that the DID identification assumption can still be valid in this scenario, while the assumption of uniform assignment probability is violated. Therefore, a standard permutation test that considers all possible combinations of treatment/control status would not provide valid inference. In this scenario, permutations tests that take into account the assignment mechanism would be valid. However, the assignment mechanisms will usually be unknown to the econometrician in DID applications. Moreover, with only one or just a few treated groups there is little hope in reliably estimating the assignment mechanism from the data. Therefore, with unknown assignment mechanisms, the assumptions for the RI approach would not be satisfied, while our method would still provide valid inference.

We argue now that, even under uniform assignment probability, RI would usually provide an unsatisfactory solution to the inference problem in our setting. The key point is that, while a standard permutation test would provide a valid for an unconditional test (that is, before we know the information on the size of the treated groups), it would not provide a reasonable inference solution for a researcher given the data he has at hand. The case for conditional inference traces back Fisher (1934), Pitman (1938), and Cox (1958). As Mcullagh (1992) argues, unconditional tests can be "*...inappropriate and possibly misleading for hypotheses concerning the data at hand*", while "*...conditional probability calculation often provides an answer to which no right-thinking person could object*". In our example described in Section 2.2, if one has information on group sizes, then we argue it would be more reasonable to consider a test conditional on this information. The reason is that once one knows that the treated groups are (large) small relative to the control groups, then one knows that a permutation test that ignores this information would (under-) over-reject the null when the null is true, even if the assignment probabilities were uniform. In this case, this test would no longer have the correct size given the data at hand. One way of incorporating this information in a permutation test would be to apply the permutation conditional on the information of group sizes. However, if there are few control groups with the same size as the treated groups, then one would not have many possible permutations. In the extreme case where there is one treated group and no control group of the same size, this conditional permutation test would generate a p-value interval of $[0, 1]$.

Canay et al. (2014) provides alternative randomization tests that can incorporate information on groups sizes. Differently from the traditional RI theory (Fisher (1935)), Canay et al. (2014) assume that the limiting distribution of a function of the data exhibits approximate symmetric under the null, and not exact symmetry as in the traditional randomization tests. In their setting, they can work with asymptotic results in nature. Their approach might be useful in our setting because, in this alternative framework, one could do a permutation test using a test statistic that does not depend asymptotically on the size of the groups.

As an application of their method, Canay et al. (2014) suggest an inference method for DID with a finite number of treated groups (N_1) and many control groups (N_0).The main idea of their method is to consider N_1 DID estimators where in each one they use one of the treated groups and all control groups to construct a DID estimator for α (say, $\hat{\alpha}_j$, for $j = 1, ..., N_1$). With the sequence $(\hat{\alpha}_1, ..., \hat{\alpha}_{N_1})$ they construct a test statistic given by $T = \frac{|\bar{\hat{\alpha}} - \alpha_0|}{s_{\hat{\alpha}}/\sqrt{N_1}}$, where $\bar{\hat{\alpha}} = \frac{1}{N_1}\sum_{j=1}^{N_1} \hat{\alpha}_j$ and $s_{\hat{\alpha}} = \frac{1}{N_1 - 1}\sum_{j=1}^{N_1}(\hat{\alpha}_j - \bar{\hat{\alpha}})^2$. Then they consider sign changes given by $(g_1\hat{\alpha}_1, ..., g_{N_1}\hat{\alpha}_{N_1})$ where $g_j \in \{-1, 1\}$ and recalculate the test statistic. Let $T^{(1)} \leq T^{(2)} \leq ... \leq T^{(N_1)}$ be the ordered values of the test statistic considering all sign change transformation, and let $T^{(k)}$ be the $N_1(1-a)^{th}$ largest value of the test statistics, where a is the significance level. The test will reject with probability 1 if $T > T^{(k)}$ and with probability zero if $T < T^{(k)}$. If $T = T^{(k)}$, we would have to randomize whether we reject the null, so that asymptotically the test has the correct size. This randomization when $T = T^{(k)}$ translates into an important drawback for their method when there are very few treated groups. For example, if $N_1 = 2$ then there would be a 50% probability (under the null) that rejection will be decided by randomization.[21] While the test would still be unbiased, it would have poor power, as they point out in remark S.2.5 (Canay et al. (2014)).[22] This happens because they rely on variation among the treated units. Since we also exploit variation from the control groups, our method does not suffer from this problem even when we have only one treated group.

Finally, in a recent paper MacKinnon and Webb (2015a) suggest a permutation test on a t-statistic, which is constructed using CRVE. Following Canay et al. (2014) idea, their method works when the numbers of treated and control groups are large enough, as asymptotically the t-statistic would have the same distribution under the null for all permutations. However, their method does not work well with very few treated groups. In particular, their method collapses to CT when there is only one treated group. The reason is that CRVE would assign an estimated variance for the treated group equal to zero, so there would not be much variation in the *estimated* variance of the placebo estimators. The key point is that we go back to the original problem of estimating the variance of the treated groups using CRVE with few treated groups. In contrast, our method provides a valid correction for heteroskedasticity even when there is only one treated group.

[21] In the extreme case with $N_1 = 1$ it would not be possible to calculate their test statistic, since $s_{\hat{\alpha}}$ would be equal to zero. If we consider other alternatives for the test statistic such as $|\bar{\hat{\alpha}} - \alpha_0|$, then all sign transformations would generate exactly the same value for the test statistics. In this case, the test would collapse to randomly deciding whether to reject the null with probability 5%. Therefore, although the test would reject with correct size under the null, the power would also be equal to 5% for any alternative hypothesis.

[22] Canay et al. (2014) also consider a setting with both N_1 and N_0 finite. This alternative would also have poor power with very few treated groups.

5 Synthetic Controls Method

For large t^*, we also consider inference using synthetic control methods (Abadie et al. (2010)). The Synthetic Control estimator was proposed by Abadie and Gardeazabal (2003) and Abadie et al. (2010) to deal with situations where there is only one treated group and when the number of pre-treatment periods is large. This method extends the traditional DID framework by using a data-driven procedure to construct a suitable comparison group. The main idea is to use the pre-treatment period to construct a counterfactual for the treated group given by $\hat{Y}_{1t}^N = \sum_{j=2}^{N_0+1} w_j^* Y_{jt}$, where the weights w_j^* are estimated so that the differences between actual and estimated pre-treatment outcomes (Y_{1t} and $\hat{Y}_{1t}^N$) and covariates (Z_1 and $\hat{Z}_1$) are minimized.[23] In the Synthetic Control approach, one needs to decide which variables to include to estimate the weights w_j^*. Particularly important for our application, one can either include the Y_{jt} for all pre-treatment t, or leave some of the pre-treatment Y_{jt} out.

The inference method suggested in Abadie et al. (2010) is a permutation test where one estimates placebo regressions using each of the control units as a placebo treatment. In essence, this is the same as what CT do in the DID framework. However, one important difference relative to permutation tests on the treatment parameter is that Abadie et al. (2010) suggest that one should look at the ratio of post-/pre-treatment Mean Squared Predicted Error (MSPE). One of their motivations to look at this ratio is to obviate the necessity of excluding placebo runs that did not provide a good fit prior to the treatment. For example, if the outcome variable of one placebo group is always lower than the outcome variables of the other groups, then the estimated counterfactual outcome for this group would always be atypically higher than the actual outcome, both before and after the treatment. Therefore, when one divides by the pre-treatment MSPE, this corrects for the fact that the Synthetic Control estimators for this placebo group will always be large. We now evaluate whether a permutation test for the synthetic control estimator is robust to heteroskedasticity.

Consider the model in Abadie et al (2010),

$$Y_{jt} = \alpha_{1t} d_{it} + \gamma_t + \beta_t Z_j + \lambda_t \mu_j + \eta_{jt}^{SC} \tag{16}$$

where d_{jt} is an indicator variable that equals one if j is the treated region and $t > t^\star$ (post-intervention period), and Z_j is a vector of observed covariates for region j. The unobserved residual is $u_{jt} = \lambda_t \mu_j + \eta_{jt}^{SC}$. They assume that the η_{jt}^{SC} are *i.i.d* cross j and t, and that η_{jt}^{SC} are mean indepedent of $\{Z_j, \mu_j\}_{j=1}^N$. Suppose that we construct our Synthetic Control estimator using the lag of the outcome variables from periods one

[23] For more details, see Abadie et al. (2010).

to $t^* - k$. Therefore, we leave the last k pre-treatment outcomes out from the minimization problem to estimate the synthetic control weights.

We again consider the setting in Canay et al. (2014) in that the distribution of the data is approximately symmetric and we can work with asymptotic results. Following Canay et al. (2014), we need a test statistic that has the same limiting distribution regardless of the permutation. We propose the following test statistic:

$$t^{SC} = \frac{\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}(Y_{1t} - \hat{Y}_{1t}^N)^2}{\frac{1}{T-t^*+k}\sum_{t=t-k+1}^{T}(Y_{1t} - \hat{Y}_{1t}^N)^2} \tag{17}$$

Differently from Abadie et al. (2010), we include in the denominator only the pre-treatment lags not included in the estimation of the synthetic control weights and we also include the post-treatment periods. We do not include in the test statistic lag outcomes included in the estimation of the weights. As in Abadie et al. (2010), suppose that there are $(w_2^*, ..., w_{t^*}^*)$ such that $\sum_{j=2}^{N} w_j^* Y_{jt} = Y_{1t}$ for $t \leq t^*$ and $\sum_{j=2}^{N} w_j^* Z_j = Z_1$. We show in Appendix A.4 that, if $\sum_{t=1}^{t^*-k} \lambda_t' \lambda_t$ is nonsingular, then the predicted error is dominated by the transient shocks η_{jt}^{SC}:

$$Y_{1t} - \sum_{j=2}^{N_0+1} w_j^* Y_{jt} \to \eta_{1t}^{SC} - \sum_{j=2}^{N_0+1} w_j^* \eta_{jt}^{SC} \tag{18}$$

In Appendix A.4 we show that, if the error term η_{jt}^{SC} is normally distributed and independent of $\{Z_j, \mu_j\}_{j=1}^{N}$, then the test statistic t^{SC} has asymptotically the same distribution in all permutations as $t^* - k \to \infty$ even if $var(\eta_{1t}^{SC}) \neq var(\eta_{jt}^{SC})$. If we assume only that the first and second moments of η_{jt}^{SC} are independent of $\{Z_j, \mu_j\}_{j=1}^{N}$ (instead of assuming normality and independence), then we can still show that this test statistic has the same expected value regardless of the permutation. The variance of this test statistic, however, will depend on the variance of the treated group unless the forth moment of the distribution of $\frac{\eta_{jt}^{SC}}{\sqrt{var(\eta_{jt}^{SC})}}$ is equal to 3 (which is the case under normality).[24]

We recommend leaving out from the test statistic the squared predicted errors from lags included in the estimation of the synthetic control weights. Although it is possible to show that test statistics that include lags used as predictors would still be asymptotically valid (assuming normality and that the approximation error goes to zero), lags not included in this minimization problem should provide a better approximation to the post-treatment squared predicted error. We show in our simulations in Section 6.2 that a permutation test

[24] We can show that the inference method proposed by Abadie et al. (2010) also corrects for heteroskedasticity under the assumptions of normality and $t^* - k \to \infty$. We also show in Appendix A.4 that the expected value of their test statistic might depend on the variance of the treated group unless the forth moment of the distribution of $\frac{\eta_{jt}^{SC}}{\sqrt{var(\eta_{jt}^{SC})}}$ is equal to 3.

using our method provided better heteroskedasticity correction than alternatives that include the squared predicted error of lags used as predictors in the test statistic. In particular, this suggests that it is not a good idea to include all outcome lags as predictors, as is commonly used in synthetic control applications (see Kaul et al. (2015)).[25] Note also that a test that relies on only on a few excluded lags might be underpowered, as the test statistic would be highly dependent on the transient shocks in these periods. On the other hand, excluding many lags might imply in a higher approximation error in the synthetic control estimation.

One important assumption is that the error term η_{jt}^{SC} is i.i.d. across time for each group j. Note that the assumption that errors are serially uncorrelated in a synthetic control model is weaker than this assumption in a DID model, since the synthetic control model includes the term $\lambda_t \mu_j$. The main assumption, therefore, is that the unobserved common factors already capture all possible serial correlation in the synthetic control model. Another assumption is that the variance of η_{jt}^{SC} is constant across t. This assumption might be problematic if, for example, there is variation in the number of observations per group across time, as argued in Section 3.

Finally, it is also important to note that the permutation graphical analyses in Abadie et al. (2010) would also suffer from the heteroskedasticity problem we highlight in this paper.[26] An easy way to make these graphs more accurate in the presence of heteroskedasticity is to divide each placebo estimate by the squared root of its pre-treatment MSPE and multiply it by the squared root of the pre-treatment MSPE of the treated group (always leaving out pre-treatment lags used in the estimation the synthetic control weights).

6 Monte Carlo Evidence

In this section we provide Monte Carlo evidence of different hypothesis testing methods in DID. We also simulate difference inference methods for synthetic control models in Section 6.2. We assume that the underlying data generating process (DGP) is given by:

$$Y_{ijt} = \nu_{jt} + \epsilon_{ijt} \tag{19}$$

In most simulations, we estimate a DID model given by equation 6 where only $j = 1$ is treated and $T = 2$, and then we test the null hypothesis of $\alpha = 0$ using different hypothesis testing methods. We consider variations in the DGP along three dimensions:

[25] They also show that using all outcome lags as separate predictors renders all other covariates irrelevant.

[26] Figures 4 to 7 in Abadie et al. (2010).

1. The number of groups: $N_0 + 1 \in \{25, 50, 100, 400\}$.

2. The intra-group correlation: ν_{jt} and ϵ_{ijt} are drawn from normal random variables. We hold constant the total variance $var(\nu_{jt} + \epsilon_{ijt}) = 1$, while changing $\rho = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\epsilon^2} \in \{.01\%, 1\%, 4\%\}$.

3. The number of observations within group: we draw for each group j the number of observations per period from a discrete uniform random variable with range $[\underline{M}, \overline{M}] \in \{[50, 200], [200, 800], [50, 950]\}$.[27]

For each case, we simulated 100,000 estimates. We present rejection rate results for inference using robust standard errors in the individual-level OLS regression, CT, DL, and for the cluster residual bootstrap with and without our heteroskedasticity correction. We do not include in the simulations methods that allow for unrestricted heteroskedasticity. As explained in Section 2.1, these methods do not work well when there is only one treated group. We also do not include MacKinnon and Webb (2015a) method in the simulations because their method collapses to CT when there is only one treated group.

6.1 Inference in DID Models

6.1.1 Test Size

We present in Table 1 results from simulations using 400 groups (one treated and 399 controls) for different numbers of observations per group and for different values of the intra-group correlations. In panel A, we present results when the number of individual observations per group varies from 50 to 200. Column 1 shows that average rejection rates for a test with 5% significance using robust standard errors in the individual level DID regression. The rejection rate is slightly higher than 5% when the intra-group correlation $\rho = 0.01\%$ (5.4%), but increases sharply for larger values of the intra-group correlation. Rejection rate is 19% when $\rho = 1\%$ and 42% when $\rho = 4\%$. When we use DL, CT or cluster residual bootstrap without correction, average rejection rate is always around 5% (columns 3, 5, and 7). However, this average rejection rate hides an important variation with respect to the number of observations in the treated group (M_1).

In Figure 1.A, we show rejection rates for cluster residual bootstrap without correction conditional on the size of the treated group.[28] The rejection rate is around 14% when the treated group is in the first decile of number of observations per group, while it is only 0.8% when the treated group is in the 10th decile. Note also that this distortion in rejection rates is not confined to the extremes of the distribution of group sizes.

[27] In the Monte Carlo simulations, we always consider the case $M(j,t) = M_j$. In the simulations with real datasets in Section 7, there is variation in $M(j,t)$ across t.

[28] Results for DL and CT are similar.

For example, the rejection rate is 3% when the treated group is in the 6th decile of number of observations per group. We summarize this variation in rejection rates by looking at the absolute difference in rejection rates for each decile of M_1 relative to the average rejection rate. Then we average these absolute differences across deciles. We present these results in columns 4, 6, and 8 for the methods without heteroskedasticity correction. Conditional on the number of observations of the treated group, these methods present an average variation in the rejection rates of 3.4-3.9 percentage points for a 5% significance test.

We present rejection rates by decile of the treated group for cluster residual bootstrap without correction when $\rho = 1\%$ and when $\rho = 4\%$ in Figures 1.B and 1.C, respectively. As expected, this variation in rejection rates becomes less relevant when the intra-group correlation becomes stronger. This happens because the aggregation from individual to group x time averages induces less heteroskedasticity in the residuals when a larger share of the residual is correlated within group. Still, even when $\rho = 4\%$ the difference in rejection rates by number of observations in the treated group remains relevant. The rejection rate is around 6.5% when the treated group is in the first decile of number of observations per group, while it is 4.2% when the treated group is in the 10th decile. The average absolute difference in rejection rates for DL, CT and for the residual bootstrap without correction is around 0.7 percentage points in this scenario.

Given that inference using these methods is problematic when there is variation in the number of observations per group, we consider our residual bootstrap method with heteroskedasticity correction derived in Section 2.3. We present rejection rates by decile of the treated group when the intra-group correlation is 0.01%, 1% and 4% in Figures 1.D to 1.F. Average rejection rates using our method are always around 5% and, more importantly, there is no variation with respect to the number of observations in the treated group. These results are also presented in columns 9 and 10 of Table 1. The average absolute difference in rejection rates is only around 0.1-0.2 percentage points, regardless of the value of the intra-group correlation.

In panel B of Table 1 we present the simulation results when the number of observations per group increases from $[50, 200]$ to $[200, 800]$. We increase the number of observations per group while holding the ratio between the number of observations in different groups constant. Note that increasing the number of observations per group worsens the over-rejection problem of inference relying in robust OLS standard errors. When we consider DL, CT and residual bootstrap without correction, increasing the number of observations per group ameliorates the problem of (over-) under-rejecting the null when M_1 is (small) large relative to the number of observations in the control groups. In particular, when $\rho = 4\%$ the average absolute difference in rejection rates across deciles of M_1 is only 0.3 percentage points. However, increasing the number of observations has no detectable effect when the intra-group correlation is 0.01%. This happens because in this case the individual component of the residual becomes more relevant. Therefore, the ratio between the variance of W_1 and the variance of W_j becomes less sensitive with respect to the number of observations per group. As explained in Section 2, in the extreme case with $\rho = 0$, heteroskedasticity would still be a problem even when $M \rightarrow \infty$.

In panel C of Table 1, we present the simulation results when the number of observations vary from 50 to 950. Therefore, the average number of observations remains constant, but we have more variation in M relative to the simulations in panel B. As expected, more variation in the number of observations per group worsens the inference problem we highlight in CT, DL and residual bootstrap without correction. On the contrary, our residual bootstrap with heteroskedasticity correction remains accurate irrespective of the variation in the number of observations per group.

As presented in Section 2.3, our method works asymptotically when $N_0 \rightarrow \infty$. This assumption is important for two reasons. First, as in any other cluster bootstrap method, a small number of groups implies a small number of possible distinct pseudo-samples. In this case, the bootstrap distribution will not be smooth even with many bootstrap replications (Cameron et al. (2008)). In order to mitigate this problem, we apply the insight of wild cluster bootstrap to our method, so that we can generate more variation in the bootstrap samples, as explained in Section 2.3. Additionally, our method requires that we estimate $var(W_j|M_j)$ using the group x time aggregate data so that we can apply our heteroskedasticity correction. If there are only a few groups, then our estimator of $var(W_j|M_j)$ will be less precise. In particular, it might be the case that $\widehat{var(W_j|M_j)} < 0$ for some j, which implies that we would not be able to normalize the residual of observation j. When $\widehat{var(W_j|M_j)} < 0$ for some j, we used the following rule: if $\hat{A} < 0$, then we used $\widehat{var(W_j|M_j)} = \frac{1}{M_j}$, as $\hat{A} < 0$ would suggest that there is not not a large intra-group or serial correlation problem. If $\hat{B} < 0$, then we used $\widehat{var(W_j|M_j)} = 1$, as $\hat{B} < 0$ would suggest that there is not much heteroskedasticity. It is important to note that asymptotically this rule would not be relevant, since $var(W_j|M_j) > 0$ for all M. We had $\widehat{var(W_j|M_j)} > 0$ for all j in more than 99.97% of our simulations with $N = 400$. However, when there are fewer control groups, the function $var(W_j|M_j)$ will be estimated with less precision.

We present in Tables 2 to 4 and in Figures 2 to 4 the simulation results when the total number of groups are 100, 50 and 25. Average rejection rates are always lower than 5.3% when the total number of groups is 100 or 50, which is reasonably close to the correct size of the test. More importantly, the average absolute difference in rejection rates is always lower than 0.5 percentage points, suggesting that there is not much variation in rejection rates depending on the size of the treated group. These results are confirmed in Figures

2 and 3. When we have 25 groups, then average rejection rates are slightly higher, at around 5.5%, and we start to have more variation depending on the size of the treated group. As shown in Figure 4, there is some distortion in rejection rates when the treated group is in the first decile of group size. Still, our method provides reasonably accurate hypothesis testing with 25 groups. In particular, our method provides substantial improvement relative to alternative methods when the intra-group correlation is not too strong.

Note that with a chi-squared distribution for ν_{jt} it would not be true that $W_j|M_j$ is identically distributed up to a scale parameter. As a robustness check, we consider a DGP where $\nu_{jt} \sim \chi^2(1)/\sqrt{2}$ instead of being normally distributed.[29] The simulation results under this alternative DGP are presented in Table 5 for $N \in \{25, 50, 100, 400\}$ and $\rho \in \{0.01\%, 1\%, 4\%\}$. In all simulations we consider M_j uniformly distributed between 50 and 200. We consider inference using residual bootstrap with and without heteroskedasticity correction. Still, our simulation results suggest that our method still provides reliable inference in this setting. The simulation results are very similar to the case where ν_{jt} is normally distributed. In particular, our method provides substantial improvement relative to a bootstrap without correction when the intra-group correlation is not too strong. In Section 7 we provide evidence that our inference method also provides substantial improvement relative to alternative methods in simulations with real datasets, where we do not have control on the DGP.

6.1.2 Test Power

We have focused so far on Type I error. We saw in Section 6.1.1 that our method is efficient in providing tests that reject the null with the correct size when the null is true. We are interested now in whether our tests have power to detect effects when the null hypothesis is false. We run the same simulations as in Section 6.1.1, with the difference that we now add an effect of β standard deviations for observation $\{ijt\}$ when $d_{jt} = 1$. Given that we know the DGP in our Monte Carlo simulations, we can calculate the variance of $\hat{\alpha}$ given the parameters of the model and generate a t-statistic $t = \frac{\hat{\alpha}}{\sigma_{\hat{\alpha}}}$. Note that with two periods and one treated group, with $N_0 \rightarrow \infty$, the DID OLS estimator is asymptotically equivalent to the GLS estimator where the full structure of the variance/covariance matrix is known. Therefore, since the errors in our DGP are normally distributed, we know that a test based on this t-statistic is the uniformly most powerful test (UMP) for this particular case. We then compare the power of the bootstrap with our heteroskedasticity correction with the power of the UMP test.

In Figure 5, we present power results for different intra-group correlation parameters and for different

[29] We divide by $\sqrt{2}$ so that $var(\nu_{jt}) = 1$. This makes these simulations more comparable to the case $\nu_{jt} \sim N(0, 1)$.

distributions of group sizes when there are 400 groups (1 treated and 399 control groups) separately when the treated group is above and below the median of number of observations per group. The most important feature in these graphs is that, for this particular DGP, the power of our method converges to the power of the UMP test when we have many control groups in all intra-group correlation and group size scenarios. It is also interesting to note that the power is higher when the treated group is larger. This is reasonable, since the main component of the variance of the DID estimator with few treated and many control groups comes from the variance of the treated groups. The difference in power for above- and below-median treated groups vanishes when the intra-group correlation increases. This happens because a higher intra-group correlation makes the model less heteroskedastic, so the size of the treated group would be less related to the precision of the estimator. Finally, the power of the test decreases with the intra-group correlation which reflects that, for a given number of observations per group, a higher intra-group correlation implies more volatility in the group x time regression.

When we have 25 groups (1 treated and 24 control), then the power of our method is slightly lower than the power of the UMP test (Figure 6). This is partially explained by fact that we need to estimate the function $var(W_j|M_j)$ and, with a finite number of control groups, this function would not be precisely estimated. Still, the power of our method is relatively close to the power of the UMP test, especially when the intra-group correlation is not high.

6.2 Inference in Synthetic Controls

An alternative estimation method when there is only one treated group and the number of pre-treatment periods is large is the synthetic control estimator. We run 100,000 simulations using the DGP presented in equation 19 with 20 pre-intervention periods (t^*) and 30 periods in total (T). We consider the case with $N = 20$, $M \in [50, 950]$ and $\rho = 0.01$.[30] In Figure 7.A, we present rejection rates when we construct the synthetic control estimator using all outcome lags as predictors and use the test statistic suggested in Abadie et al. (2010) in the permutation test. Interestingly, there is substantial variation in rejection rates depending on the size of the treatment group. Rejection rates vary from 1% when the treated group is in the first decile of M to 8.5% when it is in the last decile. In Figure 7.B, we also use the test statistic suggested in Abadie et al. (2010), but we estimate the synthetic control weights using only the first 17 pre-treatment periods. There is still substantial variation in rejection rates, with only a slight improvement relative to Figure 7.A. On the

[30] With a higher ρ or a lower variance in M, the heteroskedasticity problem is less noticeable. However, qualitative results are the same.

contrary, when we use the test statistic we propose, leaving out lags used as predictors, there is virtually no variation in rejection rates depending on the size of the treated group (Figure 7.C). These results suggests that, when the variance of the treated group is higher, the squared predicted error of lags not used as predictors is more than proportionally higher than the squared predicted error for post-treatment periods. This implies that the adjustment in the test statistic when we divide by the pre-treatment MSPE would not be accurate. Therefore, these results suggest that it is important to exclude the predicted squared errors of lags used in the estimation of the synthetic control weights so that the test statistic in all permutation have asymptotically the same distribution.

7 Simulations with Real Datasets

The results presented in Section 6 suggest that heteroskedasticity generated by variation in group sizes invalidates inference methods that rely on homoskedasticity such as DL, CT and cluster residual bootstrap, while our method performs well in correcting for heteroskedasticity when there are 25 or more groups. However, a natural question that arises is whether these results are "externally valid." In particular, we want to know (i) whether heteroskedasticity generated by variation in group sizes is a problem in real datasets with large number of observations, and (ii) whether our method works in real datasets, where we do not have control over the DGP. More specifically, our DGP implies that the *real* variance of W_j would have exactly the relationship $var(W_j|M_j) = A + \frac{B}{M_j}$, which might not be the case in real datasets. To illustrate the magnitude of the heteroskedasticity problem and to test the accuracy of our method, we conduct simulations of placebo interventions using two different real datasets: the American Community Survey (ACS) and the Current Population Survey (CPS).[31]

We consider two different group levels for the ACS based on the geographical location of residence: Public Use Microdata Areas (PUMA) and states. Simulations using placebo interventions at the PUMA level would be a good approximation to our assumption that N_1 is small while $N_0 \to \infty$. Simulations using placebo interventions at the state level would mimic situations of DID designs that are commonly used in applied work where the treatment unit is a state, with a dataset that includes a very large number of observations per group x time cell. We also consider the CPS for simulations with more than two periods. As shown in Bertrand et al. (2004), this dataset exhibits an important serial correlation in the residuals, so we want to check whether our method method is efficient in correcting for that.

[31] We created our ACS extract using IPUMS (Ruggles et al. (2015)).

We use the ACS dataset for the years 2005 to 2013, and the CPS Merged Outgoing Rotation Groups for the years 1979 to 2014. We extract information on employment status and earnings for women between ages 25 and 50, following Bertrand et al. (2004). We present in Table 6 the distribution of number of observations per group x cell for the PUMA-level ACS (column 1), for the state-level ACS (column 2) and for the state-level CPS (column 3). There are, on average, 778 observations in each PUMA x time cell in the ACS. This number, however, hides an important heterogeneity in cell sizes. The 10th percentile of PUMA x time cell sizes is 174, while the 90th percentile is 1,418. There is also substantial heterogeneity in state x time cell sizes in the ACS. While the average cell size is 10,138, the 10th percentile is 1,250, while the 90th percentile is 21,099. Finally, the state x time cells in the CPS have substantially fewer observations compared to the ACS. While the average cell size is 771, the 10th percentile is 392, while the 90th percentile is 1709.

For the ACS simulations, we consider pairs of two consecutive years and estimate placebo DID regressions using one of the groups (PUMA or state) at a time as the treated group. Note that this differs from Bertrand et al. (2004) simulations, as they randomly selected half of the states to be treated. In each simulation, we test the null hypothesis that the "intervention" has no effect ($\alpha = 0$) using robust standard errors, and bootstrap with and without our heteroskedasticity correction. Since we are looking at placebo interventions, if the inference method is correct, then we would expect to reject the null roughly 5% of the time for a test with 5% significance level. For each pair of years, the number of PUMAs that appear in both years ranges from 427 to 982, leading to 5,188 regressions in total. For the state-level simulations, we have $51 \times 8 = 408$ regressions (we include Washington, D.C.). For the CPS simulations, we used 2, 4, 6 or 8 consecutive years always using the first half of the years as pre-treatment and the other half as post-treatment. This leads to 1479 to 1785 regressions, depending on the number of years used in each regression.

7.1 American Community Survey (ACS) Results

In Panel A of Table 7, we present results from simulations using the PUMA-level treatments using the ACS. In column 1, we show rejection rates using OLS robust standard errors in the individual-level DID regression. Rejection rates for a 5% significance test are 7.2% when the outcome variable is employment, and 8.1% when it is log wages. This over-rejection suggests that there is important intra-group correlation that the robust individual-level standard error does not take into account. In column 3 of Table 7, we present results for the bootstrap without the heteroskedasticity correction (results for DL and CT are simular). As in the Monte Carlo simulations, average rejection rates without correction are very close to 5%. However, there is substantial variation when we look at rejection rates conditional on the size of the treated group.

We present in column 4 of Table 7 the difference in rejection rates when the number of observations in the treated group is above and below the median.[32] For both outcome variables, the rejection rate is 8 percentage points lower when the treated group has a group size above the median. This implies a rejection rate of almost 9% when the treated group is below the median, and slightly lower than 1% when the treated group is above the median. In columns 5 and 6 of Table 7, we present the rejection rates using bootstrap with our heteroskedasticity correction. For both outcomes, average rejection rate has the correct size of 5% and, more importantly, there is virtually no difference between rejection rates when the treated group is above or below the median. Therefore, our method was successful in correcting for the heteroskedasticity problem even in a setting where we do not have control over the DGP.

We present in Panel B of Table 7 the results for state-level simulations. The most striking result in this table is that rejection rates using bootstrap without correction still depend on the size of the treated group. This happens in a dataset with, on average, more than 10,000 observations per group x time cell. In particular, the rejection rate in the simulations with log wages as the outcome variable is zero when the treated group is below the median, and 10% when the treated group is above the median. We present rejection rates using bootstrap with our heteroskedasticity correction in columns 5 and 6. Average rejection rates are around 5%, and we cannot reject that there is no difference in rejection rates above and below the median. However, this test of our method is under-powered, since we estimate rejection rates in the state-level models based on only 408 simulations. In order to provide more precision to estimate the rejection rates of our method, we simulate DID placebo regressions randomly selecting 50 PUMAs in each simulation, which generates many more placebo estimates. These results are presented in panel C of Table ACS. We also present results DID placebo regressions randomly selecting 25 PUMAs in each simulation in Panel D of Table 7. Remarkably, our method still provides hypothesis testing with correct size regardless of the size of the treated group when $N = 50$ and when $N = 25$.

7.2 Current Population Survey (CPS) Results

We present the simulation results using the CPS in Table 8. Panel A presents rejection rates of DID models using 2 years of data, while Panels B, C and D present rejection rates using respectively 4, 6 and 8 years. Inference with OLS robust standard errors on the individual-level model becomes worse when we include more years of data in the model (column 1). This result is consistent with the findings in Bertrand et al.

[32] Given that we have a limited number of simulations, we do not calculate the average absolute difference in rejection rates across deciles, as we do in the Monte Carlo simulations. For the PUMA-level simulations, there are only approximately 500 simulations for each decile. For the state-level simulations there would be only around 40 simulations for each decile.

(2004). The key point is that the panel structure of the CPS Merged Outgoing Rotation Groups generates serial correlation in the errors. We present rejection rates for the residual bootstrap without correction in columns 3 and 4. The average rejection rates are close to 5% irrespective of the number of periods, which was expected given that this method takes serial correlation into account by looking at a linear combination of the residuals (as in CT). However, since this linear combination of the residuals is heteroskedastic, rejection rates based on this method vary with the size of the treated group. We present rejection rates using bootstrap with our heteroskedasticity correction in columns 5 and 6. As in the ACS simulations, we cannot reject that rejection rates have the correct size on average and that rejection rates do not depend on the size of the treated group in all simulations. Therefore, our method is efficient in correcting for heteroskedasticity in a scenario that serial correlation is important without the need to specify the structure of the serial correlation.

7.3 Power with Real Data Simulations

We saw in Sections 7.1 and 7.2 that our method provides tests with correct size in simulations with the ACS and the CPS. We now present in Figure 8 power results from simulations with these datasets. Figure 8.A shows power results using the ACS with state-level treatment. When the treated group is above the median, our method is able to detect an effect size of 0.06 log points with probability greater than 70%. When the treated group is below median, we are only able to attain this power for effects greater than 0.1 log points. This again reflects that the variance of $\hat{\alpha}$ is higher when the treated group is smaller. Figures 8.B to 8.E present results for simulations using the CPS with different numbers of time periods. The power in the CPS simulations is considerably lower than in the ACS simulations. The power to reject an effect of 0.06 log points when the treated group is above the median ranges from 26% to 41%, depending on the number of periods used in the simulations. This happens because the ACS has a much larger number of observations than the CPS. Even though we have only one treated group in all simulations, the larger number of observations in the ACS implies that the group x time variance of the error would be smaller.[33]

As opposed to the power results presented in Section 6.1.2, we do not know the true variance of $\hat{\alpha}$, so it is not possible to compare the power of our method with the power of the UMP test. Still, results from the Monte Carlo simulations suggest that the power of our method should be very close to the power of a UMP test.

[33] For some CPS simulations, the power when the treated group is below median crosses the power when the treated group is above median when the effect size is large. This happens because a large effect size would imply that $\widehat{W}_1^2$ (which is calculated from a model with H_0 imposed) would be large, which would bias our estimate of $var(W_j|M_j)$. Note that this does not invalidate the method, since $\widehat{var(W_j|M_j)}$ is consistent under the null. Also, this distortion only appears when the power of the test was already above 90%.

8 Conclusion

This paper shows that usual inference methods used in DID models might not perform well in the presence of heteroskedasticity when the number of treated groups is small. Then we derive an alternative inference method that corrects for heteroskedasticity when there are few treated groups (or even just one) and many control groups. With few pre-treatment periods, our method requires knowledge on how the heteroskedasticity was generated. We focus on the example of variation in group sizes, in which it is possible to derive the heteroskedasticity as a function of the number of observations per group under very mild assumptions on the errors. However, our model is more general and can be applied in any situation in which we are able to assume a structure on the variance of a linear combination of the errors, W_j. It is important to note that there is no heteroskedasticity-robust inference method in DID when there is only one treated group. Therefore, although our method is not robust to any form of unknown heteroskedasticity, it provides an important improvement relative to existing methods that rely on homoskedasticity. With many pre-treatment periods, we provide an alternative inference method that relies on strict stationarity and ergodicity of the time series instead of the assumption on how the heteroskedasticity was generated. We also provide valid inference methods for linear factor models and synthetic controls, two estimation methods that have been recently developed as alternatives to DID when there are many pre-treatment periods.

References

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

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

Andrews, D. W. K., "End-of-Sample Instability Tests," *Econometrica*, 2003, *71* (6), 1661–1694.

Angrist, J.D. and J.S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2009.

Assuncao, J. and B. Ferman, "Does affirmative action enhance or undercut investment incentives? Evidence from quotas in Brazilian Public Universities," *Unpublished Manuscript*, February 2015, *Can be found (as of Feb. 2015), at https://dl.dropboxusercontent.com/u/12654869/Assuncao%20and%20Ferman022015.pdf.*

Bai, Jushan, "Panel Data Models With Interactive Fixed Effects," *Econometrica*, 2009, *77* (4), 1229–1279.

Barndorff-Nielsen, O. E., "Conditionality Resolutions," *Biometrika*, 1980, *67* (2), 293–310.

_ , "On a formula for the distribution of the maximum likelihood estimator," *Biometrika*, 1983, *70*, 343–65.

_ , "On Conditionality Resolution and the Likelihood Ratio for Curved Exponential Models," *Scandinavian Journal of Statistics*, 1984, *11* (3), 157–170.

Bell, R. M. and D. F. McCaffrey, "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples," *Survey Methodology*, 2002, *28* (2), 169–181.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?," *Quarterly Journal of Economics*, 2004, p. 24975.

Brewer, Mike, Thomas F. Crossley, and Robert Joyce, "Inference with Difference-in-Differences Revisited," IZA Discussion Papers 7742, Institute for the Study of Labor (IZA) November 2013.

Cameron, A.C., J.B. Gelbach, and D.L. Miller, "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 2008, *90* (3), 414–427.

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

Casella, G. and C. Goutis, "Frequentist Post-Data Inference," *International Statistical Review*, 1995, *63*, 325–344.

Conley, Timothy G. and Christopher R. Taber, "Inference with "Difference in Differences with a Small Number of Policy Changes," *The Review of Economics and Statistics*, February 2011, *93* (1), 113–125.

Cox, D. R., "Some Problems Connected with Statistical Inference," *Ann. Math. Statist.*, 06 1958, *29* (2), 357–372.

_ , "Local Ancillarity," *Biometrika*, 1980, *67*, 279–86.

Cox, D.R. and D.V. Hinkley, *Theoretical Statistics*, Taylor & Francis, 1979.

Donald, Stephen G. and Kevin Lang, "Inference with Difference-in-Differences and Other Panel Data," *The Review of Economics and Statistics*, May 2007, *89* (2), 221–233.

Efron, David V. Hinkley Bradley, "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information," *Biometrika*, 1978, *65* (3), 457–482.

Fisher, R. A., "Two New Properties of Mathematical Likelihood," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 1934, *144* (852), 285–307.

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

Fraser, D.A.S., *The structure of inference* Wiley series in probability and mathematical statistics, Wiley, 1968.

Gobillon, Laurent and Thierry Magnac, "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls," PSE Working Papers halshs-00849071, HAL July 2013.

Hansen, Christian B., "Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects," *Journal of Econometrics*, October 2007, *140* (2), 670–694.

Hausman, Jerry and Guido Kuersteiner, "Difference in difference meets generalized least squares: Higher order properties of hypotheses tests," *Journal of Econometrics*, June 2008, *144* (2), 371–391.

Hinkley, D. V., "Likelihood as Approximate Pivotal Distribution," *Biometrika*, 1980, *67* (2), 287–292.

Ibragimov, Rustam and Ulrich K. Mller, “Inference with Few Heterogenous Clusters,” 2013.

Imbens, Guido W. and Michal Kolesar, “Robust Standard Errors in Small Samples: Some Practical Advice,” Working Paper 18478, National Bureau of Economic Research October 2012.

Kaul, Ashok, Stefan Klobner, Gregor Pfeifer, and Manuel Schieler, “Synthetic Control Methods: Never Use All Pre-Intervention Outcomes as Economic Predictores,” 2015.

Lehmann, E.L. and J.P. Romano, *Testing Statistical Hypotheses* Springer Texts in Statistics, Springer New York, 2008.

Liang, Kung-Yee and Scott L. Zeger, “Longitudinal data analysis using generalized linear models,” *Biometrika*, 1986, *73* (1), 13–22.

MacKinnon, James G. and Matthew D. Webb, “Differences-in-Differences Inference with Few Treated Clusters,” 2015.

__ **and** __ , “Wild Bootstrap Inference for Wildly Different Cluster Sizes,” Working Papers 1314, Queen’s University, Department of Economics February 2015.

Mcullagh, P., “Local Sufficiency,” *Biometrika*, 1984, *71*, 233–44.

__ , “Conditional Inference and Cauchy Models,” *Biometrika*, 1992, *79* (2), 247–259.

Moulton, Brent R., “Random group effects and the precision of regression estimates,” *Journal of Econometrics*, August 1986, *32* (3), 385–397.

Pengyuan, Wang, Traskin Mikhail, and Small Dylan S., “Robust Inferences from a Before-and-After Study with Multiple Unaffected Control Groups,” *Journal of Causal Inference*, 2013, *1* (2), 209–234.

Pitman, E. J. G., “The Estimation of the Location and Scale Parameters of a Continuous Population of any Given Form,” *Biometrika*, 1938, *30* (3-4), 391–421.

Rosenbaum, Paul R., “Covariance Adjustment in Randomized Experiments and Observational Studies,” *Statist. Sci.*, 08 2002, *17* (3), 286–327.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek, “Integrated Public Use Microdata Series: Version 6.0 [Machine-readable database].,” 2015.

Webb, Matthew D., "Reworking Wild Bootstrap Based Inference for Clustered Errors," Working Papers 1315, Queen's University, Department of Economics November 2014.

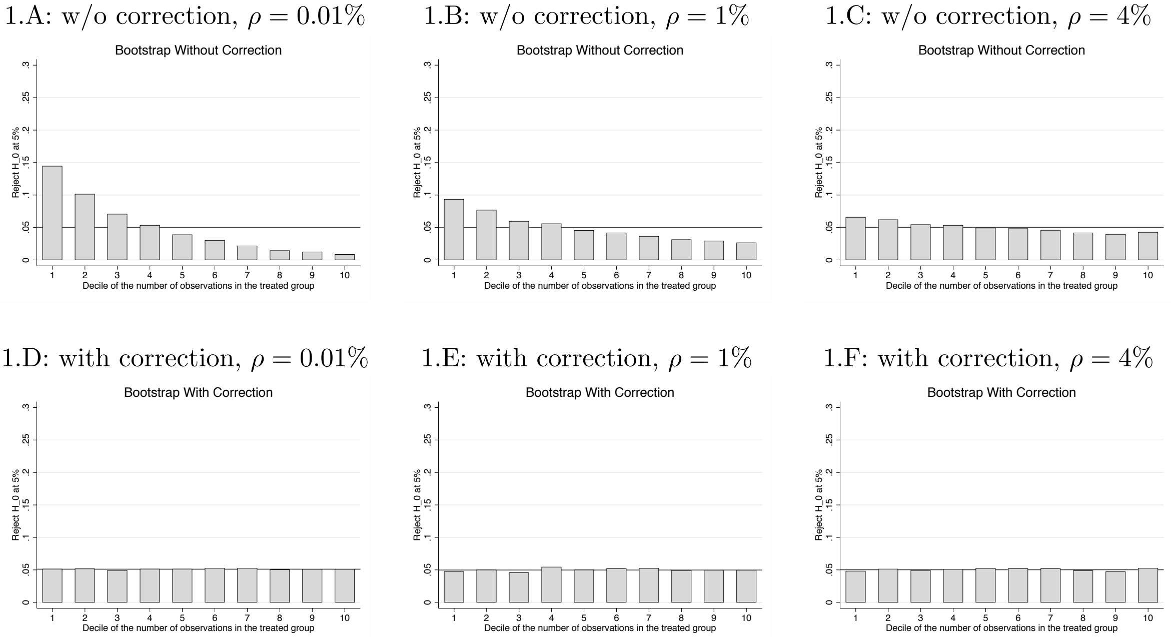
White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, May 1980, *48* (4), 817–838.

Wooldridge, Jeffrey M., "Cluster-Sample Methods in Applied Econometrics," *American Economic Review*, 2003, *93* (2), 133–138.

Yates, F., "Tests of Significance for 2 2 Contingency Tables," *Journal of the Royal Statistical Society. Series A (General)*, 1984, *147* (3), 426–463.

Young, Alwyn, "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results," 2015.

Figure 1: **Rejection Rates in MC Simulations by Decile of M_1, $N = 400$**



1.A: w/o correction, $\rho = 0.01\%$

1.B: w/o correction, $\rho = 1\%$

1.C: w/o correction, $\rho = 4\%$

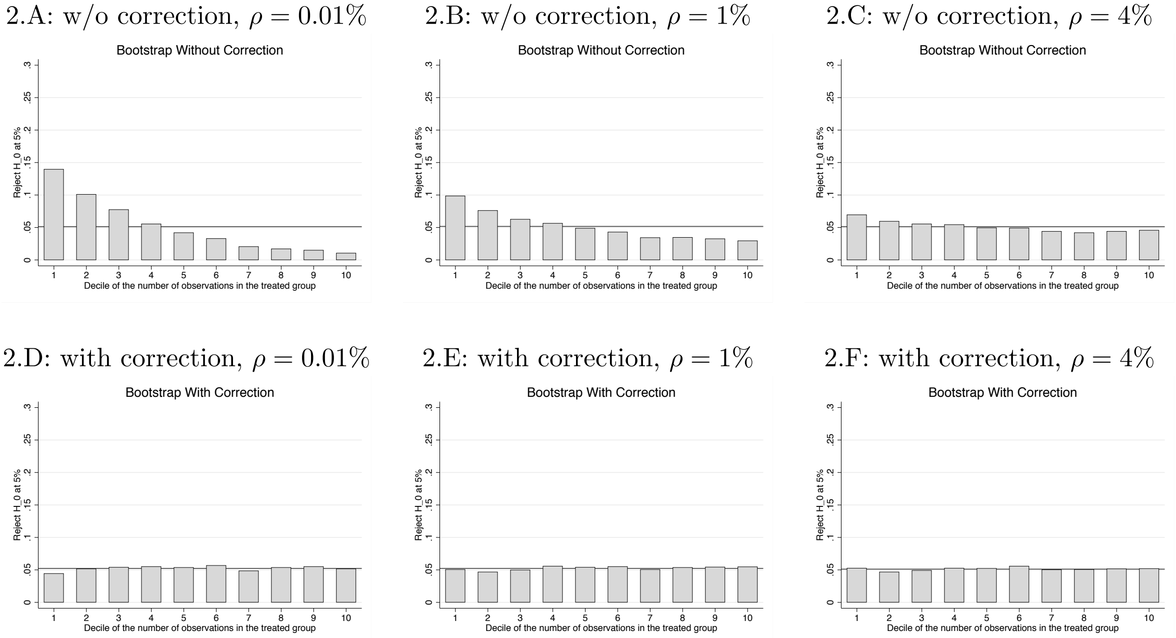
1.D: with correction, $\rho = 0.01\%$

1.E: with correction, $\rho = 1\%$

1.F: with correction, $\rho = 4\%$

Notes: These figures present the rejection rates conditional on the decile of the number of observation of the treated group when $N = 400$ and $M \in [50, 200]$. These rejection rates are based on Monte Carlos simulations explained in Section 6. Figures 1.A to 1.C present results using the residual bootstrap without correction, while Figures 1.D to 1.F present results using the residual bootstrap method with our heteroskedasticity correction, as explained in Section 2.3.

Figure 2: **Rejection Rates in MC Simulations by Decile of M_1, $N = 100$**


2.A: w/o correction, $\rho = 0.01\%$

2.B: w/o correction, $\rho = 1\%$

2.C: w/o correction, $\rho = 4\%$

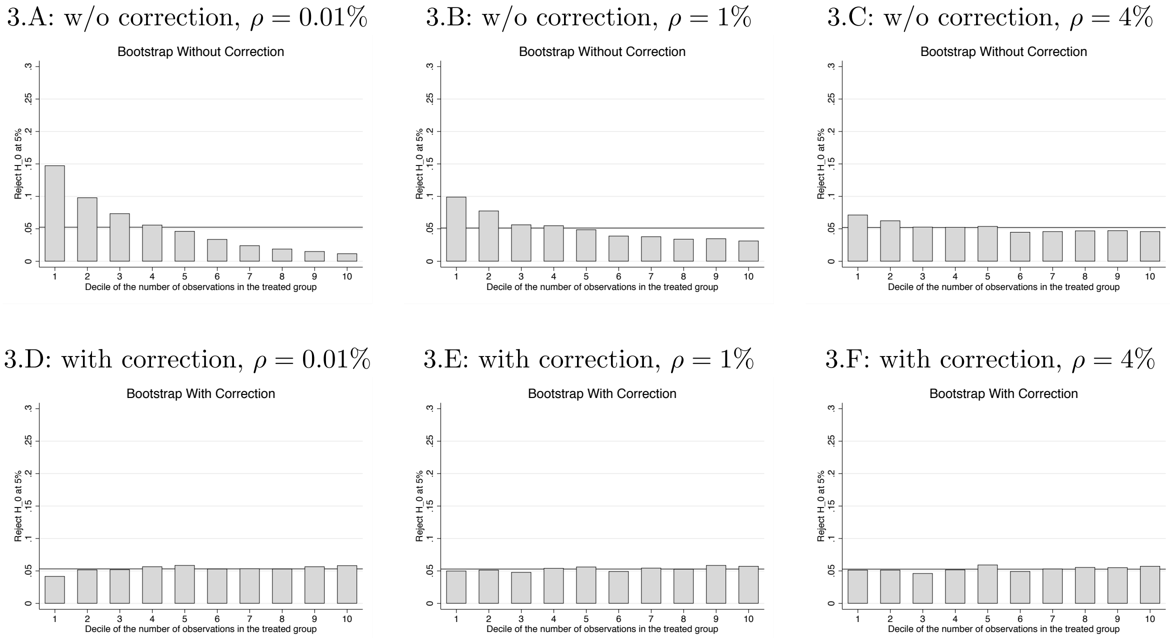
2.D: with correction, $\rho = 0.01\%$

2.E: with correction, $\rho = 1\%$

2.F: with correction, $\rho = 4\%$

Notes: These figures present the rejection rates conditional on the decile of the number of observation of the treated group when $N = 100$ and $M \in [50, 200]$. These rejection rates are based on Monte Carlos simulations explained in Section 6. Figures 2.A to 2.C present results using the residual bootstrap without correction, while Figures 2.D to 2.F present results using the residual bootstrap method with our heteroskedasticity correction, as explained in Section 2.3.

Figure 3: **Rejection Rates in MC Simulations by Decile of M_1, $N = 50$**



3.A: w/o correction, $\rho = 0.01\%$

3.B: w/o correction, $\rho = 1\%$

3.C: w/o correction, $\rho = 4\%$

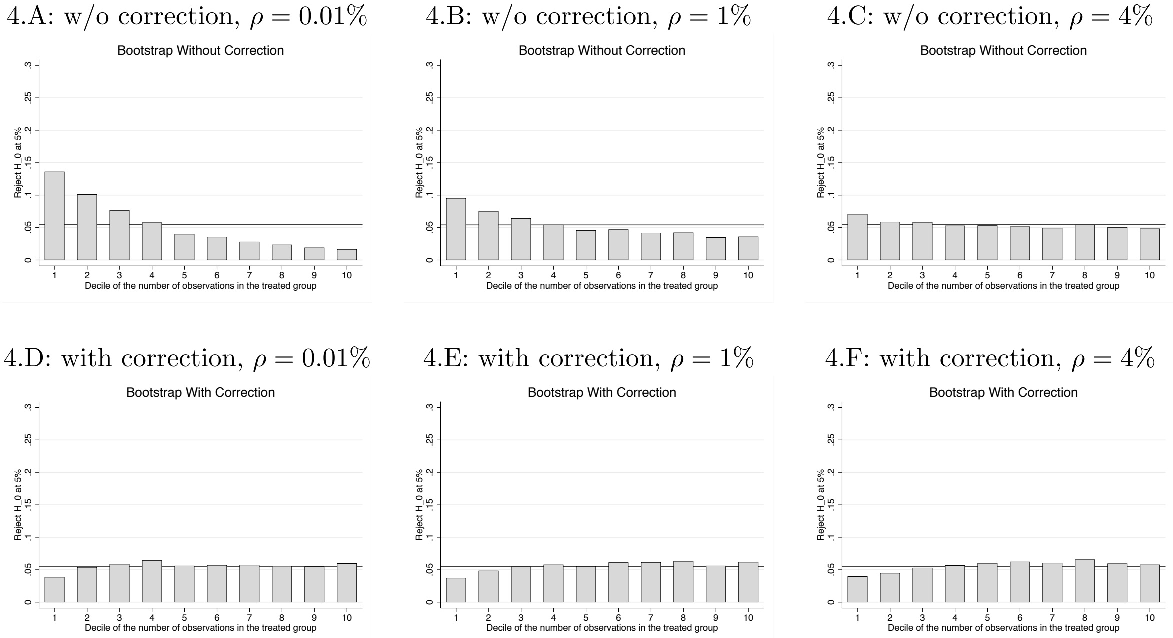
3.D: with correction, $\rho = 0.01\%$

3.E: with correction, $\rho = 1\%$

3.F: with correction, $\rho = 4\%$

Notes: These figures present the rejection rates conditional on the decile of the number of observation of the treated group when $N = 50$ and $M \in [50, 200]$. These rejection rates are based on Monte Carlos simulations explained in Section 6. Figures 3.A to 3.C present results using the residual bootstrap without correction, while Figures 3.D to 3.F present results using the residual bootstrap method with our heteroskedasticity correction, as explained in Section 2.3.

Figure 4: **Rejection Rates in MC Simulations by Decile of M_1, $N = 25$**



4.A: w/o correction, $\rho = 0.01\%$

4.B: w/o correction, $\rho = 1\%$

4.C: w/o correction, $\rho = 4\%$

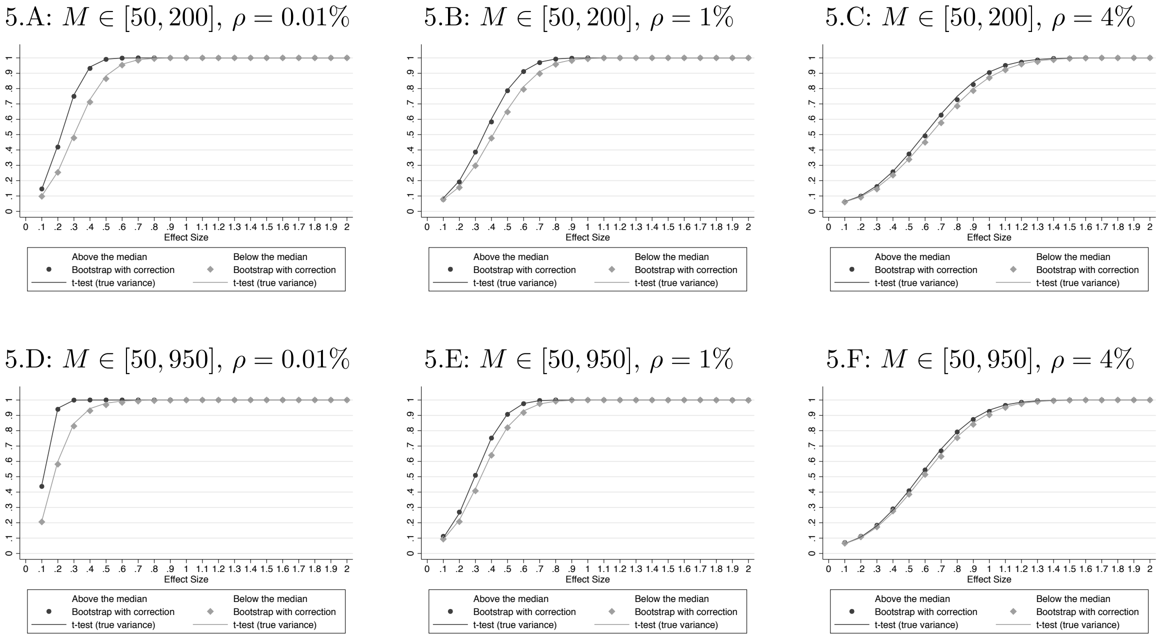
4.D: with correction, $\rho = 0.01\%$

4.E: with correction, $\rho = 1\%$

4.F: with correction, $\rho = 4\%$

Notes: These figures present the rejection rates conditional on the decile of the number of observation of the treated group when $N = 25$ and $M \in [50, 200]$. These rejection rates are based on Monte Carlos simulations explained in Section 6. Figures 4.A to 4.C present results using the residual bootstrap without correction, while Figures 4.D to 4.F present results using the residual bootstrap method with our heteroskedasticity correction, as explained in Section 2.3.

Figure 5: **Test Power by Treated Group Size - Monte Carlo Simulations with** $N = 400$


5.A: $M \in [50, 200]$, $\rho = 0.01\%$

5.B: $M \in [50, 200]$, $\rho = 1\%$

5.C: $M \in [50, 200]$, $\rho = 4\%$

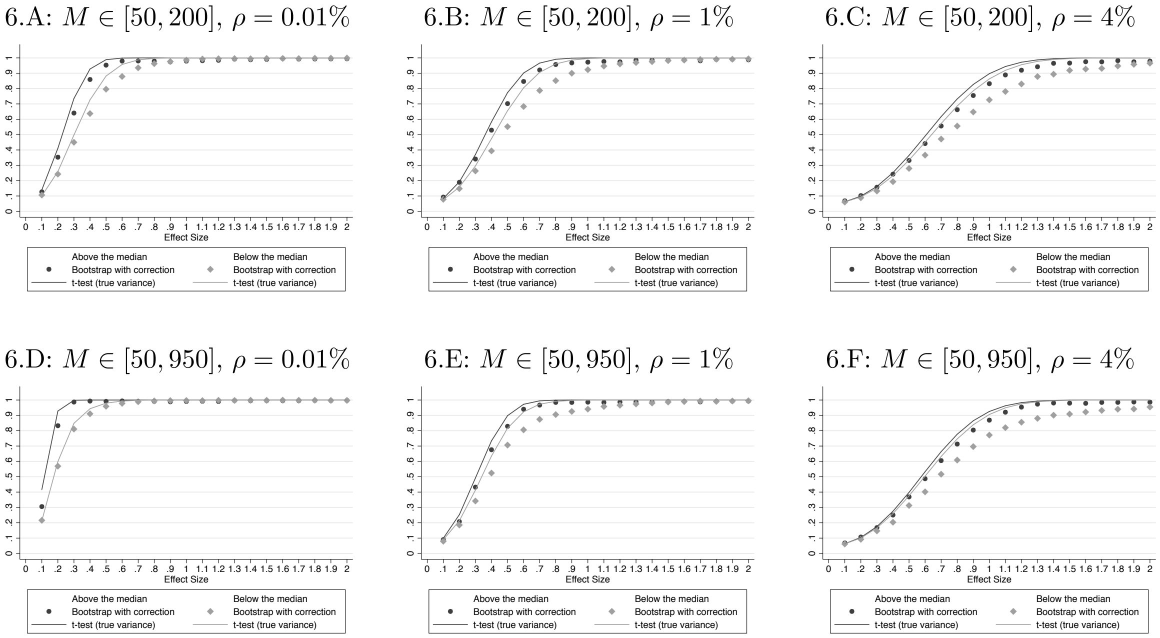
5.D: $M \in [50, 950]$, $\rho = 0.01\%$

5.E: $M \in [50, 950]$, $\rho = 1\%$

5.F: $M \in [50, 950]$, $\rho = 4\%$

Notes: These figures present the power of the bootstrap with heteroskedasticity correction as a function of the effect size separately when the treated group is above and below the median of group size. The standard deviation of the individual level observation is equal to one across the different scenarios. Therefore, the effect size is in standard deviation terms. Results are based on simulations with total number groups $N = 400$.

Figure 6: **Test Power by Treated Group Size - Monte Carlo Simulations with $N = 25$**



6.A: $M \in [50, 200]$, $\rho = 0.01\%$

6.B: $M \in [50, 200]$, $\rho = 1\%$

6.C: $M \in [50, 200]$, $\rho = 4\%$

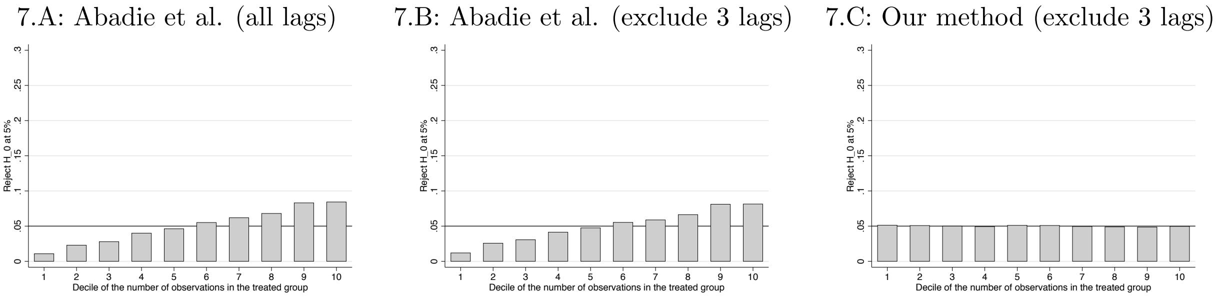
6.D: $M \in [50, 950]$, $\rho = 0.01\%$

6.E: $M \in [50, 950]$, $\rho = 1\%$

6.F: $M \in [50, 950]$, $\rho = 4\%$

Notes: These figures present the power of the bootstrap with heteroskedasticity correction as a function of the effect size separately when the treated group is above and below the median of group size. The standard deviation of the individual level observation is equal to one across the different scenarios. Therefore, the effect size is in standard deviation terms. Results are based on simulations with total number groups $N = 25$.

Figure 7: **Inference with Synthetic Control**



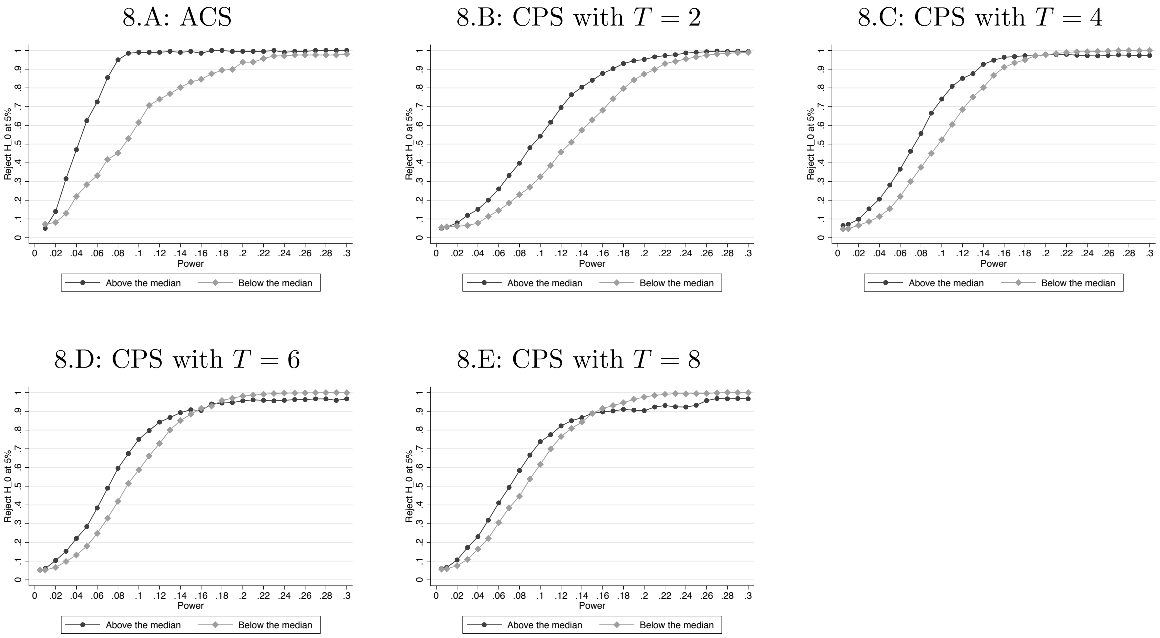
7.A: Abadie et al. (all lags)

7.B: Abadie et al. (exclude 3 lags)

7.C: Our method (exclude 3 lags)

Notes: These figures present rejection rates from Monte Carlo simulations using the synthetic control method. The DGP is described in Section 6.2. Figure 7.A presents rejection rates using the permutation test suggested in Abadie et al. (2010) when we include all outcome lags as predictors. In Figures 7.B and 7.C we estimate the synthetic control using pre-treatment lags from 1 to 17 as predictors. Figure 7.B presents rejection rates using Abadie et al. (2010) test statistic while Figure 7.C presents rejection rates using our proposed test statistic.

Figure 8: **Test Power by Treated Group Size - Simulations with Real Dataset**



8.A: ACS

8.B: CPS with $T = 2$

8.C: CPS with $T = 4$

8.D: CPS with $T = 6$

8.E: CPS with $T = 8$

Notes: These figures present the power of the bootstrap with heteroskedasticity correction for simulations using real datasets. Results are presented separately when the treated group is above and below the median of group size. The outcome variable is log wages, and effect sizes are measured in log points. Figure 8.A presents results using the ACS, while Figures 8.B to 8.E present results using the CPS with varying number of periods.

Table 1: **Rejection Rates in MC Simulations with $N_0 + 1 = 400$**

	Inference Method									
	Robust OLS		Donald and Lang		Conley and Taber		Bootstrap w/o correction		Bootstrap with correction	
ρ	Mean (1)	Absolute Difference (2)	Mean (3)	Absolute Difference (4)	Mean (5)	Absolute Difference (6)	Mean (7)	Absolute Difference (8)	Mean (9)	Absolute Difference (10)
Panel A: $M \in [50, 200]$										
0.01%	0.054	0.002	0.053	0.039	0.050	0.036	0.049	0.034	0.051	0.001
1%	0.192	0.036	0.050	0.019	0.050	0.018	0.049	0.017	0.050	0.002
4%	0.420	0.059	0.049	0.007	0.050	0.006	0.050	0.007	0.050	0.002
Panel B: $M \in [200, 800]$										
0.01%	0.057	0.002	0.053	0.036	0.051	0.034	0.049	0.034	0.049	0.002
1%	0.415	0.065	0.051	0.008	0.049	0.006	0.050	0.008	0.050	0.002
4%	0.661	0.051	0.049	0.004	0.051	0.003	0.050	0.003	0.050	0.002
Panel C: $M \in [50, 950]$										
0.01%	0.057	0.003	0.054	0.061	0.051	0.057	0.050	0.057	0.051	0.002
1%	0.396	0.098	0.051	0.019	0.051	0.019	0.049	0.018	0.050	0.001
4%	0.637	0.093	0.051	0.006	0.049	0.006	0.050	0.006	0.049	0.002

Note: This table presents results from Monte Carlo simulations with 400 groups, as explained in Section 6. In all simulations, only one group is treated. Each line presents simulation for different values of intra-group correlation, while each panel presents results for different numbers of observations per group. We consider 5 inference methods: hypothesis testing using robust standard errors from the individual level regression, DL, CT, cluster residual bootstrap without correction, and cluster residual bootstrap with our heteroskedasticity correction. For the bootstrap methods, we imposed H_0, and we used the wild bootstrap idea of randomizing whether we multiply the residuals by 1 or -1. For each inference method, we report the average rejection rate for a 5% significance level test. We also report a measure of how rejection rates depend on the number of observations in the treated group, which we call "absolute difference". To construct this measure, we calculate the absolute difference in rejection rates for each decile of M_1 relative to the average rejection rate, and then we average these absolute differences across deciles. We run 100,000 simulations for each $M \times \rho \times N_0$ scenario. The standard error for the average rejection rates is around 0.07 percentage points, while the standard error for the absolute difference is around 0.04-0.07 percentage points.

Table 2: **Rejection Rates in MC Simulations with $N_0 + 1 = 100$**

	Inference Method									
	Robust OLS		Donald and Lang		Conley and Taber		Bootstrap w/o correction		Bootstrap with correction	
ρ	Mean	Absolute Difference	Mean	Absolute Difference	Mean	Absolute Difference	Mean	Absolute Difference	Mean	Absolute Difference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: $M \in [50, 200]$									
0.01%	0.054	0.003	0.054	0.036	0.049	0.032	0.051	0.034	0.052	0.003
1%	0.193	0.032	0.052	0.017	0.049	0.018	0.052	0.017	0.052	0.002
4%	0.418	0.062	0.052	0.008	0.047	0.007	0.051	0.007	0.051	0.002
	Panel B: $M \in [200, 800]$									
0.01%	0.057	0.001	0.052	0.037	0.049	0.032	0.050	0.033	0.050	0.002
1%	0.415	0.058	0.050	0.008	0.049	0.008	0.052	0.007	0.052	0.002
4%	0.658	0.049	0.050	0.004	0.048	0.003	0.052	0.002	0.053	0.002
	Panel C: $M \in [50, 950]$									
0.01%	0.057	0.002	0.057	0.060	0.049	0.053	0.050	0.054	0.052	0.003
1%	0.400	0.095	0.050	0.019	0.049	0.018	0.050	0.017	0.051	0.002
4%	0.636	0.089	0.049	0.006	0.048	0.005	0.052	0.006	0.051	0.001

Note: This table presents results from Monte Carlo simulations with 100 groups, as explained in Section 6. In all simulations, only one group is treated. Each line presents simulation for different values of intra-group correlation, while each panel presents results for different numbers of observations per group. We consider 5 inference methods: hypothesis testing using robust standard errors from the individual level regression, DL, CT, cluster residual bootstrap without correction, and cluster residual bootstrap with our heteroskedasticity correction. For the bootstrap methods, we imposed H_0, and we used the wild bootstrap idea of randomizing whether we multiply the residuals by 1 or -1. For each inference method, we report the average rejection rate for a 5% significance level test. We also report a measure of how rejection rates depend on the number of observations in the treated group, which we call "absolute difference". To construct this measure, we calculate the absolute difference in rejection rates for each decile of M_1 relative to the average rejection rate, and then we average these absolute differences across deciles. We run 100,000 simulations for each $M \times \rho \times N_0$ scenario. The standard error for the average rejection rates is around 0.07 percentage points, while the standard error for the absolute difference is around 0.04-0.07 percentage points.

Table 3: **Rejection Rates in MC Simulations with $N_0 + 1 = 50$**

	Inference Method									
	Robust OLS		Donald and Lang		Conley and Taber		Bootstrap w/o correction		Bootstrap with correction	
ρ	Mean (1)	Absolute Difference (2)	Mean (3)	Absolute Difference (4)	Mean (5)	Absolute Difference (6)	Mean (7)	Absolute Difference (8)	Mean (9)	Absolute Difference (10)
				Panel A: $M \in [50, 200]$						
0.01%	0.052	0.003	0.054	0.035	0.046	0.030	0.052	0.033	0.053	0.003
1%	0.192	0.037	0.051	0.017	0.046	0.014	0.051	0.016	0.053	0.003
4%	0.420	0.057	0.050	0.006	0.045	0.005	0.052	0.006	0.053	0.003
				Panel B: $M \in [200, 800]$						
0.01%	0.057	0.002	0.053	0.034	0.047	0.029	0.051	0.031	0.052	0.003
1%	0.415	0.060	0.049	0.007	0.047	0.006	0.052	0.006	0.052	0.003
4%	0.663	0.047	0.049	0.002	0.047	0.002	0.051	0.002	0.052	0.003
				Panel C: $M \in [50, 950]$						
0.01%	0.056	0.002	0.057	0.060	0.046	0.048	0.050	0.052	0.052	0.004
1%	0.398	0.099	0.051	0.019	0.047	0.017	0.051	0.015	0.051	0.004
4%	0.635	0.089	0.050	0.006	0.046	0.005	0.051	0.003	0.051	0.005

Note: This table presents results from Monte Carlo simulations with 50 groups, as explained in Section 6. In all simulations, only one group is treated. Each line presents simulation for different values of intra-group correlation, while each panel presents results for different numbers of observations per group. We consider 5 inference methods: hypothesis testing using robust standard errors from the individual level regression, DL, CT, cluster residual bootstrap without correction, and cluster residual bootstrap with our heteroskedasticity correction. For the bootstrap methods, we imposed H_0, and we used the wild bootstrap idea of randomizing whether we multiply the residuals by 1 or -1. For each inference method, we report the average rejection rate for a 5% significance level test. We also report a measure of how rejection rates depend on the number of observations in the treated group, which we call "absolute difference". To construct this measure, we calculate the absolute difference in rejection rates for each decile of M_1 relative to the average rejection rate, and then we average these absolute differences across deciles. We run 100,000 simulations for each $M \times \rho \times N_0$ scenario. The standard error for the average rejection rates is around 0.07 percentage points, while the standard error for the absolute difference is around 0.04-0.07 percentage points.

Table 4: **Rejection Rates in MC Simulations with $N_0 + 1 = 25$**

	Inference Method									
	Robust OLS		Donald and Lang		Conley and Taber		Bootstrap w/o correction		Bootstrap with correction	
ρ	Mean (1)	Absolute Difference (2)	Mean (3)	Absolute Difference (4)	Mean (5)	Absolute Difference (6)	Mean (7)	Absolute Difference (8)	Mean (9)	Absolute Difference (10)
					Panel A: $M \in [50, 200]$					
0.01%	0.052	0.002	0.053	0.033	0.078	0.038	0.053	0.032	0.055	0.004
1%	0.193	0.032	0.051	0.016	0.079	0.020	0.053	0.015	0.055	0.005
4%	0.424	0.055	0.050	0.006	0.079	0.008	0.054	0.005	0.056	0.006
					Panel B: $M \in [200, 800]$					
0.01%	0.056	0.002	0.053	0.031	0.077	0.037	0.051	0.029	0.056	0.006
1%	0.417	0.060	0.049	0.009	0.078	0.008	0.054	0.005	0.056	0.006
4%	0.664	0.048	0.050	0.005	0.079	0.003	0.055	0.001	0.054	0.007
					Panel C: $M \in [50, 950]$					
0.01%	0.057	0.003	0.056	0.055	0.076	0.059	0.047	0.045	0.056	0.004
1%	0.403	0.091	0.052	0.015	0.077	0.019	0.052	0.015	0.056	0.007
4%	0.643	0.084	0.052	0.004	0.080	0.006	0.054	0.005	0.055	0.007

Note: This table presents results from Monte Carlo simulations with 25 groups, as explained in Section 6. In all simulations, only one group is treated. Each line presents simulation for different values of intra-group correlation, while each panel presents results for different numbers of observations per group. We consider 5 inference methods: hypothesis testing using robust standard errors from the individual level regression, DL, CT, cluster residual bootstrap without correction, and cluster residual bootstrap with our heteroskedasticity correction. For the bootstrap methods, we imposed H_0, and we used the wild bootstrap idea of randomizing whether we multiply the residuals by 1 or -1. For each inference method, we report the average rejection rate for a 5% significance level test. We also report a measure of how rejection rates depend on the number of observations in the treated group, which we call “absolute difference”. To construct this measure, we calculate the absolute difference in rejection rates for each decile of M_1 relative to the average rejection rate, and then we average these absolute differences across deciles. We run 100,000 simulations for each $M \times \rho \times N_0$ scenario. The standard error for the average rejection rates is around 0.07 percentage points, while the standard error for the absolute difference is around 0.04-0.07 percentage points.

Table 5: **Rejection Rates in MC Simulations with $\nu_{jt} \sim \chi^2(1)/\sqrt{2}$**

	Bootstrap w/o correction		Bootstrap with correction	
ρ	Mean (1)	Absolute Difference (2)	Mean (3)	Absolute Difference (4)
	Panel A: $N = 400$			
0.01%	0.050	0.035	0.050	0.002
1%	0.050	0.013	0.050	0.002
4%	0.050	0.003	0.050	0.003
	Panel B: $N = 100$			
0.01%	0.051	0.035	0.053	0.002
1%	0.052	0.013	0.052	0.003
4%	0.051	0.004	0.051	0.003
	Panel A: $N = 50$			
0.01%	0.051	0.033	0.053	0.003
1%	0.052	0.012	0.053	0.004
4%	0.053	0.003	0.053	0.004
	Panel B: $N = 25$			
0.01%	0.051	0.029	0.055	0.005
1%	0.052	0.012	0.056	0.007
4%	0.055	0.003	0.056	0.007

Note: This table replicates the Monte Carlo simulation results presented in Table 1 to 4 with $\nu_{jt} \sim \chi^2(1)/\sqrt{2}$. All simulations consider $M \in [50, 200]$.

Table 6: **Number of Observations per Group x Time cell**

	ACS		CPS
	PUMA	State	State
	(1)	(2)	(3)
Average	778.12	10,137.79	771.23
1%	129	883	119
5%	157	1,037	355
10%	174	1,250	392
25%	218	2,527	464
50%	338	7,205	546
75%	703	11,509	775
90%	1,418	21,099	1,709
95%	2,469	32,961	1,937
99%	9,555	62,752	3,297

Note: This Table presents the distribution of number of observations per groups in the simulations with real datasets (Section 7). Column 1 presents information for PUMA-level ACS simulations, column 2 presents information for state-level ACS simulations, while column 3 presents information for state-level CPS simulations.

Table 7: **Simulations with the ACS Survey**

	Inference Method					
	Robust OLS		Bootstrap w/o correction		Bootstrap with correction	
Outcome Variable	Mean (1)	Diff (2)	Mean (3)	Diff (4)	Mean (5)	Diff (6)
Panel A: ACS with PUMA level interventions						
Employment	0.072***	0.010	0.050	-0.082***	0.049	-0.003
	(0.004)	(0.008)	(0.003)	(0.006)	(0.003)	(0.006)
Log(wages)	0.081***	0.000	0.050	-0.086***	0.050	0.002
	(0.004)	(0.008)	(0.003)	(0.006)	(0.003)	(0.007)
Panel B: ACS with state level interventions						
Employment	0.064	0.003	0.044	-0.087***	0.051	-0.013
	(0.011)	(0.021)	(0.016)	(0.028)	(0.011)	(0.020)
Log(wages)	0.081**	-0.021	0.054	-0.106**	0.056	-0.032
	(0.015)	(0.031)	(0.022)	(0.042)	(0.015)	(0.030)
Panel C: ACS with PUMA level interventions, $N = 50$						
Employment	0.072***	0.001	0.045*	-0.072***	0.050	-0.005
	(0.004)	(0.007)	(0.003)	(0.005)	(0.003)	(0.005)
Log(wages)	0.084***	-0.001	0.046	-0.073***	0.052	-0.001
	(0.004)	(0.008)	(0.003)	(0.005)	(0.003)	(0.005)
Panel D: ACS with PUMA level interventions, $N = 25$						
Employment	0.069***	0.009	0.041***	-0.062***	0.050	-0.007
	(0.004)	(0.007)	(0.003)	(0.004)	(0.002)	(0.004)
Log(wages)	0.082***	0.000	0.042***	-0.064***	0.051	-0.001
	(0.004)	(0.008)	(0.003)	(0.004)	(0.003)	(0.005)

Note: This table presents rejection rates for the simulations using ACS data. For each pair of consecutive years, we run a DID regression using one group as treated and the other groups as a control. The outcome variable is employment status or log(wages) for women aged between 25 and 40. Then we test the hypothesis that the effect of the "intervention" is equal to zero using different inference methods: hypothesis testing using robust standard errors from individual level DID model, bootstrap without and bootstrap with our heteroskedasticity correction. Panel A reports results when groups are defined as PUMAs, while Panel B reports results when groups are defined as states. In Panels C and D we present results with PUMA-level treatments using 50 and 25 randomly selected PUMAs (for each PUMA x year we simulated with 5 different randomly chosen sets of control PUMAs, leading to 25,940 simulations). We report average rejection rate and the difference in rejection rates when the size of the treated group is above or below the median. Given that we have a limited number of simulations, we do not calculate the average absolute difference in rejection rates across deciles, as we do in the Monte Carlo simulations. We present in brackets standard errors for the rejection rates. Standard errors are clustered at the treated group level. For average rejection rates (columns 1, 3, and 5), * means that we reject at 10% that the average rejection rate is equal to 5%, while for the differences in rejection rates (columns 2, 4, and 6) * means that we reject at 10% that rejection rate for M_1 above and below the median are equal. ** means that we reject at 5%, while *** means that we reject at 1%.

Table 8: **Simulations with the CPS Survey**

	Inference Method					
			Bootstrap		Bootstrap	
	Robust OLS		w/o correction		with correction	
Outcome	Mean	Diff	Mean	Diff	Mean	Diff
Variable	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: 2 years			
Employment	0.047	-0.003	0.046	-0.043***	0.053	0.004
	(0.007)	(0.011)	(0.007)	(0.012)	(0.007)	(0.012)
Log(wages)	0.066***	-0.011	0.046	-0.047***	0.052	-0.002
	(0.006)	(0.012)	(0.007)	(0.011)	(0.006)	(0.012)
			Panel B: 4 years			
Employment	0.062*	0.016	0.046	-0.042***	0.052	-0.012
	(0.007)	(0.013)	(0.007)	(0.013)	(0.007)	(0.013)
Log(wages)	0.102***	0.024	0.048	-0.041***	0.050	0.009
	(0.011)	(0.023)	(0.007)	(0.013)	(0.008)	(0.015)
			Panel C: 6 years			
Employment	0.087***	0.001	0.053	-0.046***	0.054	-0.018
	(0.009)	(0.017)	(0.007)	(0.014)	(0.006)	(0.014)
Log(wages)	0.143***	0.059*	0.047	-0.044***	0.051	0.000
	(0.014)	(0.035)	(0.008)	(0.014)	(0.009)	(0.017)
			Panel C: 8 years			
Employment	0.135***	0.044	0.043	-0.040***	0.046	-0.009
	(0.013)	(0.028)	(0.008)	(0.015)	(0.007)	(0.014)
Log(wages)	0.207***	0.043	0.045	-0.029*	0.049	0.006
	(0.015)	(0.036)	(0.009)	(0.016)	(0.010)	(0.017)

Note: This table presents rejection rates for the simulations using CPS data. In each simulation, we run a DID regression using one group as treated and the other groups as a control. The outcome variable is employment status or log(wages) for women aged between 25 and 40. Then we test the hypothesis that the effect of the "intervention" is equal to zero using different inference methods: hypothesis testing using robust standard errors from individual level DID model, bootstrap without and bootstrap with our heteroskedasticity correction. Panel A reports results of DID models using 2 consecutive years of data, while Panels B and C report results of DID models using respectively 4 and 6 consecutive years of data. We report average rejection rate and the difference in rejection rates when the size of the treated group is above or below the median. Given that we have a limited number of simulations, we do not calculate the average absolute difference in rejection rates across deciles, as we do in the Monte Carlo simulations. We present in brackets standard errors for the rejection rates. Standard errors are clustered at the treated group level. For average rejection rates (columns 1, 3, and 5), * means that we reject at 10% that the average rejection rate is equal to 5%, while for the differences in rejection rates (columns 2, 4, and 6) * means that we reject at 10% that rejection rate for M_1 above and below the median are equal. ** means that we reject at 5%, while *** means that we reject at 1%.

A Supplemental Appendix: Inference in Differences-in-Differences with Different Group Sizes

A.1 Proof of the Main Results

This supplemental appendix contains the main theorems and proofs of the paper ”Inference in Differences-in-Differences with Different Group Sizes”. We use the same notation as in the main paper. Let $M(j,t)$ be the number of observations in group j, time t.

The aggregated model is:

$$Y_{jt} = \alpha d_{jt} + \theta_j + \beta X_{jt} + \gamma_t + \eta_{jt} \tag{20}$$

where X_{jt} is a kx1 vector of covariates. For simplicity, we start with the case that $\beta = 0$ and then extend to the case with covariates.

We assume T periods of time $(t = 1, .., T)$ and N_1 treated groups and N_0 control groups in such a way that $N_0 + N_1 = N$. Consider the restricted model in which we impose the null hypothesis, $H_0 : \alpha = \alpha_0$,

$$Y_{jt} = \alpha_0 d_{jt} + \theta_j + \gamma_t + \eta_{jt}$$

We will work with a linear combination of the residuals of this regression,

$$\widehat{W}_j^R = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T} \widehat{\eta}_{jt}^R - \frac{1}{t^*}\sum_{t=1}^{t^*} \widehat{\eta}_{jt}^R$$

We can calculate the DID coefficient $\widehat{\alpha}$ based on a linear combination of $\widehat{W}_j^R$. Define the operator $\triangledown Y_j = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T} Y_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} Y_{jt}$. We can write $\widehat{\alpha}$ as:

$$\widehat{\alpha} = \frac{1}{N_1}\sum_{j=1}^{N_1} \triangledown Y_1 - \frac{1}{N_0}\sum_{j=N_1+1}^{N} \triangledown Y_j$$

Since $\widehat{Y}_{jt} = \alpha_0 d_{jt} + \widehat{\theta}_j + \widehat{\gamma}_t$, then $\triangledown \widehat{Y}_j^R = \alpha_0 + \frac{1}{T-t^*}\sum_{t=t^*+1}^{T} \widehat{\gamma}_t - \frac{1}{t^*}\sum_{t=1}^{t^*} \widehat{\gamma}_t$ for $j = 1, ..., N_1$ and $\triangledown \widehat{Y}_j^R = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T} \widehat{\gamma}_t - \frac{1}{t^*}\sum_{t=1}^{t^*} \widehat{\gamma}_t$ for $j = N_1 + 1, ..., N$.

Therefore:

$$\widehat{\alpha} - \alpha_0 = \frac{1}{N_1}\sum_{j=1}^{N_1} \widehat{W}_j^R - \frac{1}{N_0}\sum_{j=N_1+1}^{N} \widehat{W}_j^R$$

We define W_j as a linear combination of the error terms,

$$W_j = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T} \eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} \eta_{jt}$$

We impose assumptions about the behavior of W_j. We assume that T is fixed. In the leading example of the paper, we assume that the heteroskedaticity is generated by variation of the groups' sample size. In this appendix, we deal with the general case, and then specialize to this example.

Assumption 1 (Independence): $\{W_j, X_j\}$ is i.i.d. across $j \in \{1, ..., N_1\}$, i.i.d. across $j \in \{N_1 + 1, ..., N\}$ and independently distributed across $j \in \{1, ..., N\}$.

Assumption 2 (Distribution): $W_j | X_j, d_j \overset{d}{=} W_j | \tilde{X}_j$, where $\tilde{X}_j$ is a subset of X_j.

Assumption 3 (Heteroskedasticity): $W_j|\tilde{X}_j$ has the same distribution across $\tilde{X}_j$ up to a scale parameter. That is, $\frac{W_j}{\sqrt{var(W_j|\tilde{X}_j)}}|\tilde{X}_j$ does not depend on $\tilde{X}_j$.

Assumption 4(Exogeneity): $E[W_j|X_j, d_j] = E[W_j|X_j] = 0$.

For the leading example in the paper, the conditional variance of W_j on X_j only depends on M_j and it is given by:

$$\begin{aligned} Var\left[W_j | M_{j1}, ..., M_{jT}\right] &= A + \widetilde{B}\left(\frac{1}{(T-t^*)^2}\sum_{t=t^*+1}^{T}\frac{1}{M(j,t)} + \frac{1}{(t^*)^2}\sum_{t=1}^{t^*}\frac{1}{M(j,t)}\right) \\ &= A + \widetilde{B} \cdot h\left(M(j,t)\right) \end{aligned}$$

where A and $\widetilde{B}$ are constants, and $h\left(M(j,t)\right) \equiv \frac{1}{(T-t^*)^2}\sum_{t=t^*+1}^{T}\frac{1}{M(j,t)} + \frac{1}{(t^*)^2}\sum_{t=1}^{t^*}\frac{1}{M(j,t)}$. For simplicity, in the paper, we work with the case in which $M(j,t) = M_j$. In this case, the variance simplifies to $Var\left[W_j | M_j\right] = A + \frac{B}{M_j}$ for a constant B.

Under assumptions 1, 2, 3 and 4 the variance of this DID estimator is

$$Var\left[\widehat{\alpha} - \alpha_0 | M_j\right] = A\left(\frac{N_1 + N_0}{N_1 N_0}\right) + \widetilde{B}\left(\frac{1}{N_1^2}\sum_{j=1}^{N_1} h\left(M(j)\right) + \frac{1}{N_0^2}\sum_{j=N_1+1}^{N} h\left(M(j)\right)\right) \tag{21}$$

In our leading example, we assume that the number of individuals in each group is fixed and does not vary with N_0. As $N_0 \to \infty$,

$$\widehat{\alpha} - \alpha_0 \to \frac{1}{N_1}\sum_{j=1}^{N_1}\widehat{W}_j^R$$

$$Var\left[\widehat{\alpha} | M_j\right] \to \frac{A}{N_1} + \widetilde{B}\left(\frac{1}{N_1^2}\sum_{j=1}^{N_1} h\left(M(j)\right)\right)$$

In general, if we know the variance of $W_j|\tilde{X}_j$, we could re-scale the residuals $\widehat{W}_j^R$ and use a cluster residual bootstrap on the re-scaled residuals even if the model is heteroskedastic. The idea is to normalize $\widehat{W}_j^R$ such that $\widehat{W}_j^{norm} = \widehat{W}_j^R \cdot \sqrt{\frac{1}{Var\left[W_j | X_j\right]}}$, and then generate a bootstrap sample using the re-scaled residuals $\widetilde{\widehat{W}}_{j,b} = \widehat{W}_{j,b}^{norm} \cdot \sqrt{Var\left[W_j | X_j\right]}$, and use the residuals $\widetilde{\widehat{W}}_{j,b}$ to estimate $\widehat{\alpha}_b - \alpha_0$,

$$\widehat{\alpha}_b - \alpha_0 = \frac{1}{N_1}\sum_{j=1}^{N_1}\widetilde{\widehat{W}}_{j,b} - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\widetilde{\widehat{W}}_{j,b}$$

where b indicates each re-sampling, $b = 1, ..., \mathcal{B}$. In each re-sampling, we calculate $\widehat{\alpha}_b$. We reject H_0 at level α if and only if $\widehat{\alpha} - \alpha_0 < (\widehat{\alpha}_b - \alpha_0)\left[\frac{\alpha}{2}\right]$ or $\widehat{\alpha} - \alpha_0 > (\widehat{\alpha}_b - \alpha_0)\left[1 - \frac{\alpha}{2}\right]$, where $(\widehat{\alpha}_b - \alpha_0)[q]$ denotes the qth quantile of the distribution of $\{(\widehat{\alpha}_1 - \alpha_0), ..., (\widehat{\alpha}_{\mathcal{B}} - \alpha_0)\}$.

Let's $\tilde{X}$ be the matrix with $\tilde{X}_j$ for $j = 1, ..., N$

Theorem 1 *Define $d^*_{1-\frac{\alpha}{2}}$ and $d^*_{\frac{\alpha}{2}}$ as the $(1-\frac{\alpha}{2})$th and $\frac{\alpha}{2}$th quantile of the empirical distribution of $(\widehat{\alpha}_b - \alpha_0)$ given X, for $b = 1, ..., \mathcal{B}$. Assuming that we know the variance of $W_j|\tilde{X}_j$, under assumptions 1, 2, 3 and 4* $\Pr\left[d^*_{1-\frac{\alpha}{2}} \leq \widehat{\alpha} - \alpha_0 \leq d^*_{\frac{\alpha}{2}} \middle| \alpha_0, X\right] \to_p 1 - \alpha$.

Proof. We divide this proof in two parts. Define $\Gamma_j(w) \equiv \Pr\left[\sum_{j=1}^{N_1} W_j < w|\tilde{X}_j\right]$ and $\widehat{\Gamma}_{j,b}(w) = \Pr\left[\sum_{j=1}^{N_1}\widehat{W}_{j,b}^R < w|\tilde{X}_j, b\right]$. First we show that $\widehat{\Gamma}_{j,b}(w)$ converges in probability to $\Gamma_j(w)$ uniformly on any compact subset of the support of W, as $N_0 \to \infty$ and $\mathcal{B} \to \infty$. Then, we show that $\Pr\left[d^*_{1-\frac{\alpha}{2}} \leq \widehat{\alpha} - \alpha_0 \leq d^*_{\frac{\alpha}{2}} \middle| \alpha_0, X\right] \to_p 1 - \alpha$.

Since under our assumptions, $W_j^{'}|\tilde{X}_j$s are independent across j and have the same distribution except by the variance, we can write

$$\begin{aligned}\Gamma_j\left(w\right) &= \Pr\left[\sum_{j=1}^{N_1} W_j < w|\tilde{X}_j\right] \\ &= \int ... \int 1\left\{\sum_{j=1}^{N_1} W_j < w\right\} dF_1\left(W_1|\tilde{X}_1\right) \cdot dF_2\left(W_2|\tilde{X}_2\right) \cdot ... \cdot dF_{N_1}|\tilde{X}_{N_1}\left(W_{N_1}\right)\end{aligned}$$

and

$$\begin{aligned}\widehat{\Gamma}_{j,b}\left(w\right) &= \Pr\left[\sum_{j=1}^{N_1} \widehat{W}^R_{j,b} < w_{j,b}|\tilde{X}_{j,b}\right] \\ &= \int ... \int 1\left\{\sum_{j=1}^{N_1} \widehat{W}^R_j < w\right\} d\widehat{F}_1\left(W^R_1|\tilde{X}_1\right) \cdot d\widehat{F}_2|\tilde{X}_2\left(W^R_2\right) \cdot ... \cdot d\widehat{F}_{N_1}\left(W^R_{N_1}|\tilde{X}_{N_1}\right)\end{aligned}$$

In order to estimate this distribution, we use $\widehat{F}_j\left(.\right)$ which is the empirical CDF obtained using the re-scaled residuals $\widetilde{W}_{j,b} = \widehat{W}^R_{j,b} \cdot \sqrt{\frac{Var\left[W_j|X_j\right]}{Var\left[W_{j,b}|X_{j,b}\right]}}$,

$$\begin{aligned}\widehat{F}_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right) &= \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} 1\{\widetilde{W}_{j,b} < w_{j,b}\} \\ &= \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} 1\left\{\widehat{W}^R_{j,b} \cdot \sqrt{\frac{Var\left[W_j\right|X_j]}{Var\left[W_{j,b}\right|X_{j,b}]}} < w_j \cdot \sqrt{\frac{Var\left[W_j\right|X_j]}{Var\left[W_{j,b}\right|X_{j,b}]}}\right\}\end{aligned}$$

where $w_{j,b} = w_j \cdot c_{jb}$, with $c_{jb} = \sqrt{\frac{Var\left[W_j|X_j\right]}{Var\left[W_{j,b}|X_{j,b}\right]}}$ .In this case, c_{jb} is a constant.

Define $\widehat{F}^*_{j,b|\tilde{X}_{j,b}}\left(w_{j,bj}\right) = \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} 1\{W_{j,b} < w_{j,b}\}$. Note that

$$\begin{aligned}\sup_{w_j\in\Theta}\left|\widehat{F}_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right) - \Gamma_j\left(w\right)\right| &= \sup_{w_j\in\Theta}\left|\widehat{F}_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right) - \widehat{F}^*_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right) + \widehat{F}^*_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right) - F_{j|\tilde{X}_j}\left(w\right)\right| \\ &\leq \sup_{w_j\in\Theta}\left|\widehat{F}_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right) - \widehat{F}^*_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right)\right| + \sup_{w_j\in\Theta}\left|\widehat{F}^*_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right) - F_{j|\tilde{X}_j}\left(w\right)\right|\end{aligned}$$

Define ι_T as a vector Tx1 of $1's$ and ι_N as a vector Tx1 of $1's$. and note that,

$$\begin{aligned}\widehat{\eta}^R_{jt} &= y_{jt} - \widehat{\theta}_j - \widehat{\gamma}_t \\ &= \widetilde{\widetilde{y}}_{jt} = \widetilde{\widetilde{\eta}}_{jt}\end{aligned}$$

where $\widetilde{\widetilde{y}}_{jt} = (1-P_T)(1-P_N)y_{jt}$ and $\widetilde{\widetilde{\eta}}_{jt} = (1-P_T)(1-P_N)\eta_{jt}$, where $P_T = \iota_T\left(\iota'_T\iota_T\right)^{-1}\iota'_T$ and $P_N = \iota_N\left(\iota'_N\iota_N\right)^{-1}\iota'_N$. As $N_0 \to \infty$, $\widetilde{\widetilde{\eta}}_{jt} \to (1-P_T)\eta_{jt}$, and we can show that

$$\widehat{W}^R_j \to \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\eta_{jt}$$

$$\begin{aligned}\sup_{w_j\in\Theta}\left|\widehat{F}_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right)-\widehat{F}^{*}_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right)\right| &= \sup_{w_j\in\Theta}\left|\frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}}\left(1\{\widehat{W}^{R}_{j,b}<w_{j,b}\}-1\{W^{R}_{j,b}<w_{j,b}\}\right)\right| \\ &\leq \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}}\sup_{w_j\in\Theta}\left|1\{\widehat{W}^{R}_{j,b}<w_{j,b}\}-1\{W^{R}_{j,b}<w_{j,b}\}\right| \\ &= o\left(1\right)\end{aligned}$$

Now, we work with the second term.

$$\begin{aligned}\sup_{w_j\in\Theta}\left|\widehat{F}^{*}_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right)-\Gamma_j\left(w\right)\right| \leq & \sup_{w_j\in\Theta}\left|\widehat{F}^{*}_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right)-F_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right)\right|+ \\ & \sup_{w_j\in\Theta}\left|F_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right)-F_{j|\tilde{X}_{j}}\left(w\right)\right|\end{aligned}$$

where $F_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right)$ is the cumulative distribution function of $W_{j,b}$. Note that $W_{j,b}$ are independent across j, have the same distribution and the same variance that equals de variance of W_j. By the Glivenko-Cantelli Theorem,

$$\sup_{w_j\in\Theta}\left|\widehat{F}^{*}_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right)-F_{j,b|\tilde{X}_{j,b}}\left(w_{j,b}\right)\right|=o_p\left(1\right)$$

In addition,

$$\begin{aligned}F_{j,b}\left(w_{j,b}|\tilde{X}_{j,b}\right) &= \Pr\left[W_{j,b}\leq w_{j,b}\right] \\ &= \Pr\left[W_j\cdot c_{jb}\leq w_j\cdot c_{jb}\right] \\ &= F_{j|\tilde{X}_j}\left(w_j\right)\end{aligned}$$

Note that

$$\begin{aligned}\sup_{w_j\in\Theta}\left|\Gamma_j\left(w\right)-\widehat{\Gamma}_{j,b}\left(w\right)\right| \leq & \sup_{w_j\in\Theta}\left|\Gamma_j\left(w\right)-\widehat{\Gamma}_{j}\left(w\right)\right| \\ & + \sup_{w_j\in\Theta}\left|\widehat{\Gamma}_j\left(w\right)-\widehat{\Gamma}_{j,b}\left(w\right)\right|\end{aligned}$$

where $\widehat{\Gamma}_j\left(w\right)=\int ...\int 1\left\{\sum_{j=1}^{N_1}W_j^R<w\right\}d\widehat{F}_1\left(W_1^R|\tilde{X}_1\right)\cdot d\widehat{F}_2\left(W_2^R|\tilde{X}_2\right)\cdot ...\cdot d\widehat{F}_{N_1}\left(W_{N_1}^R|\tilde{X}_{N_1}\right)$. By the results above,

$$\sup_{w_j\in\Theta}\left|\Gamma_j\left(w\right)-\widehat{\Gamma}_j\left(w\right)\right|=o(1)$$

$$\sup_{w_j\in\Theta}\left|\widehat{\Gamma}_j\left(w\right)-\widehat{\Gamma}_{j,b}\left(w\right)\right|=o_p\left(1\right)$$

Now, we show that $\Pr\left[d^{*}_{1-\frac{\alpha}{2}}\leq\widehat{\alpha}-\alpha_0\leq d^{*}_{\frac{\alpha}{2}}\middle|\alpha_0,X\right]\to_p 1-\alpha$. As $N_0\to\infty$,

$$\widehat{\alpha}-\alpha_0\to\frac{1}{N_1}\sum_{j=1}^{N_1}\widehat{W}_j^R \text{ and } \widehat{\alpha}_b-\alpha_0=\frac{1}{N_1}\sum_{j=1}^{N_1}\widetilde{W}_{j,b}$$

Using the results above, we can show that

$$\begin{aligned}\Pr\left[d^*_{1-\frac{\alpha}{2}} \leq \widehat{\alpha} - \alpha_0 \leq d^*_{\frac{\alpha}{2}} \middle| \alpha_0, X\right] &= \Pr\left[d^*_{1-\frac{\alpha}{2}} \leq \widehat{\alpha}_b - \alpha_0 \leq d^*_{\frac{\alpha}{2}} \middle| \alpha_0, X\right] + o_p\left(1\right) \\ &= 1 - \alpha\end{aligned}$$

■

The approach proposed to estimate $\widetilde{W}_{j,b}$ is unfeasible since we do not the variances of W_j's. Theorem 2 shows that if we have a consistent estimator of $\sqrt{\frac{Var[W_j|X_j]}{Var[W_{j,b}|X_{j,b}]}}$, we can construct $\widetilde{\widehat{W}}_{j_b} = \widehat{W}^R_{jb} \cdot \sqrt{\frac{Var[\widehat{W_j}|X_j]}{Var[\widehat{W_{j,b}}|X_{j,b}]}}$, and use the approach proposed above.

Theorem 2 *Define $d^*_{1-\frac{\alpha}{2}}$ and $d^*_{\frac{\alpha}{2}}$ as the $(1-\frac{\alpha}{2})$th and $\frac{a}{2}$th quantile of the empirical distribution of $(\widehat{\alpha}_b - \alpha_0)$ given X, for $b = 1, ..., \mathcal{B}$. If for each j $\sqrt{\frac{Var[\widehat{W_j}|X_j]}{Var[\widehat{W_{j,b}}|X_{j,b}]}}$ is a consistent estimator for $\sqrt{\frac{Var[W_j|X_j]}{Var[W_{j,b}|X_{j,b}]}}$, under assumptions 1, 2, 3 and 4*

$$\Pr\left[d^*_{1-\frac{\alpha}{2}} \leq \widehat{\alpha} - \alpha_0 \leq d^*_{\frac{\alpha}{2}} \middle| \alpha_0, X\right] \to_p 1 - \alpha$$

Proof. Now, we do not know the variance of $\not{W}_j$.In this case, we define $\widehat{F}_{j|\tilde{X}_j}\left(\widehat{w}_j\right) = \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} 1\{\widetilde{\widehat{W}}_{j,b} < w_j\}$

$$\begin{aligned}\sup_{w_j\in\Theta}\left|\widehat{F}_{j|\tilde{X}_j}\left(w_j\right) - \Gamma_j\left(w\right)\right| &= \sup_{w_j\in\Theta}\left|\widehat{F}_{j|\tilde{X}_j}\left(\widehat{w}_j\right) - \widehat{F}_{j|\tilde{X}_j}\left(w_j\right) + \widehat{F}_{j|\tilde{X}_j}\left(w_j\right) - \widehat{F}^*_{j|\tilde{X}_j}\left(w_j\right) + \widehat{F}^*_{j|\tilde{X}_j}\left(w_j\right) - \Gamma_j\left(w\right)\right| \\ &\leq \sup_{w_j\in\Theta}\left|\widehat{F}_{j|\tilde{X}_j}\left(\widehat{w}_j\right) - \widehat{F}_{j|\tilde{X}_j}\left(w_j\right)\right| + \sup_{w_j\in\Theta}\left|\widehat{F}^*_{j|\tilde{X}_j}\left(\widehat{w}_j\right) - \widehat{F}^*_{j|\tilde{X}_j}\left(w_j\right)\right| + \sup_{w_j\in\Theta}\left|\widehat{F}^*_{j|\tilde{X}_j}\left(w_j\right) - \Gamma_j\left(w\right)\right|\end{aligned}$$

We show in the previous theorem that $\sup_{w_j\in\Theta}\left|\widehat{F}^*_{j|\tilde{X}_{jb}}\left(\widehat{w}_j\right) - \widehat{F}^*_{j|\tilde{X}_j}\left(w_j\right)\right| = o(1)$ and $\sup_{w_j\in\Theta}\left|\widehat{F}^*_{j|\tilde{X}_j}\left(w_j\right) - \Gamma_j\left(w\right)\right| = o_p\left(1\right)$. We only need to work with the first term,

$$\begin{aligned}\sup_{w_j\in\Theta}\left|\widehat{F}_{j|\tilde{X}_j}\left(\widehat{w}_j\right) - \widehat{F}_{j|\tilde{X}_j}\left(w_j\right)\right| &= \sup_{w_j\in\Theta}\left|\frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} 1\{W_{j,b} < w_j \cdot \widehat{c}_{jb}\} - \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} 1\{W_{j,b} < w_j \cdot c_{jb}\}\right| \\ &\leq \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} \sup_{w_j\in\Theta}\left|1\{W_{j,b} < w_j \cdot \widehat{c}_{jb}\} - 1\{W_{j,b} < w_j \cdot c_{jb}\}\right| \\ &\to_p 0 \text{ since } \widehat{c}_{jb} \to_p c_{j_b}.\end{aligned}$$

■

For our leading example, we propose a consistent estimator of $\sqrt{\frac{Var[\widehat{W_j}|M_j]}{Var[\widehat{W_{j,b}}|M_{j,b}]}}$ based on an ordinary least squares estimator. We estimate a linear regression that relates $\left(\widehat{W}^R_j\right)^2$ with $\frac{1}{M_j}$ and constant. We obtain $\widehat{A}$ as the least squares coefficient associated with the constant, and $\widehat{B}$ as the coefficient associated with $\frac{1}{M_j}$. We use A and B to construct a consistent estimator for the $Var[W^R_j|M_j\right]$,

$$Var\left[\widehat{W^R_j}\middle| M_j\right] = \widehat{A} + \frac{\widehat{B}}{M_j}$$

We use these two estimator to estimate the ratio $\widehat{c}_{jb} \equiv \sqrt{\frac{Var\left[\widehat{W^R_j}\middle|M_j\right]}{Var\left[\widehat{W^R_{j,b}}\middle|M_j\right]}}$. Theorem 3 shows that is $\widehat{c}_{jb}$ is a consistent estimator for $\sqrt{\frac{Var[W_1|M_1]}{Var[W_j|M_j]}}$.

Theorem 3 *Under assumptions 1, 2, 3 and 4, for our leading example, $\widehat{c}_j$ is a consistent estimator for* $c_{jb} = \sqrt{\frac{Var\left[W_{j,b}\middle|M_{j,b}\right]}{Var\left[W_j\middle|M_j\right]}}$.

Proof.

$$Var\left[W_j^R\middle| M_j\right] = A + \frac{B}{M_j} \text{ and } \mathbb{E}\left[W_{jt}\middle| M_j\right] = 0$$

So we can write

$$\mathbb{E}\left[\left(W_j^R\right)^2\middle| M_j\right] = A + \frac{B}{M_j}$$

or

$$\left(W_j^R\right)^2 = A + \frac{B}{M_j} + \omega$$

where $\mathbb{E}\left[\omega\middle| M_j\right] = 0$. In this case, we estimate A and B by ordinary least squares, we obtain consistent estimators as $N_0 \to \infty$. Since M_j does not vary with N_0, $\widehat{g}\left(M_j\right) \to_p g\left(M_j\right)$. ■

A.2 Extension: Two or more treated periods

So far, we consider that that treatment happens only at t^*. Now, we extend the proofs to the case that treatment happens in different periods. We assume that there N_0 control groups, N_1 treated groups that started treatment at t_1^*,....,N_K treated groups that started treatment at t_K^*. We will say that $j \in N_0$ to refer to a group j that belongs to the control group and $j \in N_k$ with $k > 0$ to refer to a treated group j that started treatment at t_k^*. In this case, $N = N_0 + \sum_{k=1}^{K} N_k$.

First, we show that in this case, we can write the estimator $\widehat{\alpha}$ as a linear combination of $\frac{1}{N_k}\sum_{j\in N_k}\nabla^k Y_j - \frac{1}{N_0}\sum_{j\in N_0}\nabla^k Y_j$, where $\nabla^k Y_j = \frac{1}{T-t_k^*}\sum_{t>t_k^*} Y_{jt} - \frac{1}{t_k^*}\sum_{t\le t_k^*} Y_{jt}$.

Define $\widetilde{d}_{jt}$ as

$$\widetilde{d}_{jt} = d_{jt} - \frac{1}{N}\sum_{j'=1}^{N} d_{j't} - \frac{1}{T}\sum_{t'=1}^{T} d_{jt'} + \frac{1}{N}\frac{1}{T}\sum_{j'=1}^{N}\sum_{t'=1}^{T} d_{j't'}$$

By the Frisch-Waugh-Lovell theorem we know that

$$\hat{\alpha} = \frac{\sum_j\sum_t \widetilde{d}_{jt} Y_{jt}}{\sum_j\sum_t \widetilde{d}_{jt}^2}$$

We will first analyze the denominator. For $j \in N_0$, we have that:

$$\widetilde{d}_{jt} = 0 - \frac{1}{N}\sum_{k=1}^{K} 1[t > t_k^*] \times N_k - 0 + \frac{1}{NT}\sum_{k=1}^{K}(T - t_k^*)N_k$$

Since $\widetilde{d}_{jt}^2$ does not depend on j, then

$$\sum_{j\in N_0} \widetilde{d}_{jt}^2 = \frac{N_0}{N^2}\left[\frac{1}{T^2}\left(\sum_{k=1}^{K}(T - t_k^*)N_k\right)^2 + \left(\sum_{k=1}^{K} 1[t > t_k^*]N_k\right)^2 - \frac{2}{T}\left(\sum_{k=1}^{K}(T - t_k^*)N_k\right)\left(\sum_{k=1}^{K} 1[t > t_k^*]N_k\right)\right]$$

Since N_k with $k > 0$ is fixed, as $N_0 \to \infty$

$$\sum_{j\in N_0} \widetilde{d}_{jt}^2 \overset{N_0\to\infty}{\longrightarrow} 0$$

For $j \in N_k$ with $k > 0$ we have

$$\widetilde{d}_{jt} = d_{jt} - \frac{1}{N}\sum_{k'=1}^{K} 1[t > t_{k'}^*] \times N_{k'} - \frac{1}{T}(T - t_k^*) + \frac{1}{NT}\sum_{k'=1}^{K}(T - t_{k'}^*)N_{k'}$$

When $N_0 \to \infty$:

$$\begin{aligned}
\widetilde{d}_{jt}^2 &= \left(d_{jt} - \frac{1}{T}(T - t_k^*)\right)^2 \\
&= d_{jt}^2 + \frac{1}{T^2}(T - t_k^*)^2 - 2\frac{d_{jt}}{T}(T - t_k^*) \qquad (22)
\end{aligned}$$

Therefore:

$$\begin{aligned}
\sum_{t=1}^{T} \widetilde{d}_{jt}^2 &= t_k^* \left(\frac{1}{T^2}(T - t_k^*)^2\right) + (T - t_k^*)\left(1 + \frac{1}{T^2}(T - t_k^*)^2 - 2\frac{1}{T}(T - t_k^*)\right) \\
&= \frac{1}{T^2}\left[t_k^*(T - t_k^*)^2 + (T - t_k^*)(t_k^*)^2\right] = \frac{t_k^*(T - t_k^*)}{T} \qquad (23)
\end{aligned}$$

This implies that the denominator becomes (as $N_0 \to \infty$):

$$\sum_{j=1}^{N}\sum_{t=1}^{T} \widetilde{d}_{jt}^2 = \frac{1}{T}\sum_{k=1}^{K} N_k \left[t_k^*(T - t_k^*)\right] \qquad (24)$$

Now we will analyze the numerator. For a $j \in N_0$, we have that:

$$\begin{aligned}
\widetilde{d}_{jt} &= \frac{1}{NT}\sum_{k=1}^{K}(T - t_k^*)N_k - \frac{1}{N}\sum_{k=1}^{K} 1[t > t_k^*] \times N_k \\
&= \frac{1}{N}\sum_{k=1}^{K} N_k \left[\frac{(T - t_k^*)}{T} - 1[t > t_k^*]\right] \qquad (25)
\end{aligned}$$

$$\begin{aligned}
\sum_t d_{jt}Y_{jt} = \frac{1}{N}&\left(N_1\frac{(T - t_1^*)}{T} + \ldots + N_K\frac{(T - t_K^*)}{T}\right)\sum_{t \le t_1^*} Y_{jt} + \\
&\frac{1}{N}\left(N_1\frac{(-t_1^*)}{T} + N_2\frac{(T - t_2^*)}{T} + \ldots + N_K\frac{(T - t_K^*)}{T}\right)\sum_{t_1^* < t \le t_2^*} Y_{jt} + \\
&\frac{1}{N}\left(N_1\frac{(-t_1^*)}{T} + N_2\frac{(-t_2^*)}{T} + N_3\frac{(T - t_3^*)}{T} + \ldots + N_K\frac{(T - t_K^*)}{T}\right)\sum_{t_2^* < t \le t_3^*} Y_{jt} + \\
&\vdots \\
&\frac{1}{N}\left(N_1\frac{(-t_1^*)}{T} + N_2\frac{(-t_2^*)}{T} + \ldots + N_K\frac{(-t_K^*)}{T}\right)\sum_{t > t_K^*} Y_{jt} \\
&= -\frac{N_1}{N}\left(\frac{t_1^*}{T}\sum_{t > t_1^*} Y_{jt} - \frac{(T - t_1^*)}{T}\sum_{t \le t_1^*} Y_{jt}\right) - \ldots - \frac{N_K}{N}\left(\frac{t_K^*}{T}\sum_{t > t_K^*} Y_{jt} - \frac{(T - t_K^*)}{T}\sum_{t \le t_K^*} Y_{jt}\right) \qquad (26)
\end{aligned}$$

As $N_0 \to \infty$:

$$\sum_{j \in N_0}\sum_t d_{jt}Y_{jt} = \sum_{j \in N_0}\sum_{k=1}^{K} -\frac{N_k}{N}\left(\frac{t_k^*}{T}\sum_{t > t_k^*} Y_{jt} - \frac{(T - t_k^*)}{T}\sum_{t \le t_k^*} Y_{jt}\right)$$

For $j \in N_k$ with $k > 0$ as $N_0 \to \infty$: we have:

$$\widetilde{d}_{jt} = d_{jt} - \frac{1}{T}(T - t_k^*) \qquad (27)$$

Then (when $N_0 \to \infty$),

$$\sum_t d_{jt} Y_{jt} = \frac{t_k^*}{T} \sum_{t>t_k^*} Y_{jt} - \frac{(T-t_k^*)}{T} \sum_{t \le t_k^*} Y_{jt}$$

Therefore,

$$\hat{\alpha} \overset{N_0 \to \infty}{\longrightarrow} \sum_{k=1}^{K} \frac{N_k[t_k^*(T-t_k^*)]}{\sum_{k'=1}^{K} N_{k'} \left[t_{k'}^*(T-t_{k'}^*)\right]} \left[\frac{1}{N_k} \sum_{j \in N_k} \nabla^k Y_j - \frac{1}{N_0} \sum_{j \in N_0} \nabla^k Y_j \right]$$

Note that

$$\sum_{k=1}^{K} \frac{N_k[t_k^*(T-t_k^*)]}{\sum_{k'=1}^{K} N_{k'} \left[t_{k'}^*(T-t_{k'}^*)\right]} \left[\frac{1}{N_k} \sum_{j \in N_k} \nabla^k Y_j - \frac{1}{N_0} \sum_{j \in N_0} \nabla^k Y_j \right]$$
$$= \alpha_0 + \sum_{k=1}^{K} \frac{N_k[t_k^*(T-t_k^*)]}{\sum_{k'=1}^{K} N_{k'} \left[t_{k'}^*(T-t_{k'}^*)\right]} \left[\frac{1}{N_k} \sum_{j \in N_k} \widehat{W}_j^R - \frac{1}{N_0} \sum_{j \in N_0} \widehat{W}_j^R \right]$$

In our leading example,

$$Var\left[\widehat{\alpha} - \alpha_0 \middle| M_j\right] = \sum_{k=1}^{K} \left(\frac{N_k[t_k^*(T-t_k^*)]}{\sum_{k'=1}^{K} N_{k'} \left[t_{k'}^*(T-t_{k'}^*)\right]} \right)^2 \left[A \left(\frac{N_k + N_0}{N_k N_0} \right) + \widetilde{B} \left(\frac{1}{N_k^2} \sum_{j \in N_k} h\left(M\left(j\right)\right) + \frac{1}{N_0^2} \sum_{j \in N_0} h\left(M\left(j\right)\right) \right) \right]$$

As before, we assume that the number of individuals in each group is fixed and does not vary with N_0. As $N_0 \to \infty$,

$$Var\left[\widehat{\alpha} - \alpha_0 \middle| M_j\right] \to \sum_{k=1}^{K} \left(\frac{N_k[t_k^*(T-t_k^*)]}{\sum_{k'=1}^{K} N_{k'} \left[t_{k'}^*(T-t_{k'}^*)\right]} \right)^2 \left[\frac{A}{N_k} + \widetilde{B} \left(\frac{1}{N_k^2} \sum_{j \in N_k} h\left(M\left(j,t\right)\right) \right) \right]$$

In the general case, if we knew the variance of W_j, then we could use the normalized $\widehat{W}_j^R \left(\widehat{W}_j^{norm} = \widehat{W}_j^R \cdot \sqrt{\frac{1}{Var\left[W_j^R \middle| M_j\right]}} \right)$, generate a bootstrap sample using the re-scaled residuals $\widetilde{\widetilde{W}}_{j,b} = \widehat{W}_{j,b}^{norm} \cdot \sqrt{Var\left[W_j^R \middle| M_j\right]}$, and use the residuals $\widetilde{\widetilde{W}}_{j,b}$ to estimate $\widehat{\alpha}_b - \alpha_0$,

$$\widehat{\alpha} - \alpha_0 = \sum_{k=1}^{K} \frac{N_k[t_k^*(T-t_k^*)]}{\sum_{k'=1}^{K} N_{k'} \left[t_{k'}^*(T-t_{k'}^*)\right]} \left[\frac{1}{N_k} \sum_{j \in N_k} \widetilde{\widetilde{W}}_{j,b} - \frac{1}{N_0} \sum_{j \in N_0} \widetilde{\widetilde{W}}_{j,b} \right]$$

and use the same hypothesis test stated in the previous section. If the variance of W_j is unknown, then we can estimate this variance using theorem 3.

A.3 Extension: Model with Covariates

In this case, we work with the aggregate model that includes the covariates:

$$Y_{jt}{=}\, \alpha d_{jt}{+}\theta_j{+}\beta X_{jt}{+}\gamma_t{+}\eta_{jt} \tag{28}$$

where $X_{jt}$is a kx1vector of covariates. We consider the case in which treatment happens only at t^*.

As before, by applying the Frisch-Waugh-Lovell theorem we know that:

$$\hat{\alpha} = \frac{\sum_j \sum_t \widetilde{d}_{jt} \left(Y_{jt} - \widehat{\beta} X_{jt} \right)}{\sum_j \sum_t \widetilde{d}_{jt}^2}$$

and,

$$\hat{\alpha} \overset{N_0 \to \infty}{\longrightarrow} \sum_{k=1}^{K} \frac{N_k[t_k^*(T-t_k^*)]}{\sum_{k'=1}^{K} N_{k'} \left[t_{k'}^*(T-t_{k'}^*)\right]} \left[\frac{1}{N_k} \sum_{j \in N_k} \nabla^k Y_j^* - \frac{1}{N_0} \sum_{j \in N_0} \nabla^k Y_j^* \right]$$

In this case, under H_0 and for the treatment group,

$$\begin{aligned}\nabla^k Y_j =& \frac{1}{T-t_k^*}\sum_{t>t_k^*}\left(Y_{jt}-\widehat{\beta}X_{jt}\right)-\frac{1}{t_k^*}\sum_{t\leq t_k^*}\left(Y_{jt}-\widehat{\beta}X_{jt}\right)\\ &=\alpha_0+\left(\widehat{\beta}-\beta\right)\left(\frac{1}{T-t_k^*}\sum_{t>t_k^*}X_{jt}-\frac{1}{t_k^*}\sum_{t\leq t_k^*}X_{jt}\right)\\ &+\frac{1}{T-t_k^*}\sum_{t>t_k^*}\gamma_t-\frac{1}{t_k^*}\sum_{t\leq t_k^*}\gamma_t\\ &+\frac{1}{T-t_k^*}\sum_{t>t_k^*}\eta_{jt}-\frac{1}{t_k^*}\sum_{t\leq t_k^*}\eta_{jt}\end{aligned}$$

By proposition 1 in Conley and Taber (2011), $\widehat{\beta}\rightarrow_p\beta$. Therefore:

$$\hat{\alpha}\overset{N_0\rightarrow\infty}{\longrightarrow}\alpha_0+\sum_{k=1}^{K}\frac{N_k[t_k^*(T-t_k^*)]}{\sum_{k'=1}^{K}N_{k'}\left[t_{k'}^*(T-t_{k'}^*)\right]}\left[\frac{1}{N_k}\sum_{j\in N_k}W_j-\frac{1}{N_0}\sum_{j\in N_0}W_j\right]$$

and we have the same result as in the previous section.

A.4 Synthetic Controls

Now, we work with the proofs of section 5. We want to show that the test statistic $t^{SC}=\frac{\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}(Y_{1t}-\hat{Y}_{1t}^N)^2}{\frac{1}{T-t^*+k}\sum_{t=t-k+1}^{T}(Y_{1t}-\hat{Y}_{1t}^N)^2}$ proposed for the synthetic control estimator is asymptotically symmetric when $t^*-k\rightarrow\infty$. Using Abadie et al (2010) derivations, for $t>t^*-k$ we have that:[34]

$$\begin{aligned}Y_{1t}-\hat{Y}_{1t}^N&=\alpha_{1t}d_{it}+\sum_{j=2}^{N_0+1}w_j^*\sum_{s=1}^{t^*-k}\lambda_t\left(\sum_{n=1}^{t^*-k}\lambda_n^{'}\lambda_n\right)^{-1}\lambda_s^{'}\left(\eta_{js}^{SC}-\eta_{1s}^{SC}\right)+\\ &+\sum_{j=2}^{N_0+1}w_j^*\left(\eta_{jt}^{SC}-\eta_{1t}^{SC}\right)\end{aligned}\tag{29}$$

assuming that $\left\{w_j^*\right\}_{j=2}^{N_0+1}$ that satisfies the equalities $Y_{jt}{=}\sum_{j=2}^{N_0+1}w_j^*Y_{jt}$ for $t=1,...,t^*{-}k$ and $Z_j{=}\sum_{j=2}^{N_0+1}w_j^*Z_j$.

By the Cauchy-Schwartz Inequality

$$\begin{aligned}\left|\sum_{s=1}^{t^*-k}\lambda_t\left(\sum_{n=1}^{t^*-k}\lambda_n^{'}\lambda_n\right)^{-1}\lambda_s^{'}\left(\eta_{js}^{SC}-\eta_{1s}^{SC}\right)\right|^2&\leq\sum_{s=1}^{t^*-k}\left|\lambda_t\left(\sum_{n=1}^{t^*-k}\lambda_n^{'}\lambda_n\right)^{-1}\lambda_s^{'}\right|^2\sum_{s=1}^{t^*-k}\left|\eta_{js}^{SC}-\eta_{1s}^{SC}\right|^2\\ &\leq\frac{C^*}{(t^*-k)^2}\sum_{s=1}^{t^*-k}\left|\eta_{js}^{SC}-\eta_{1s}^{SC}\right|^2\rightarrow 0\text{ as }t^*-k\rightarrow\infty\end{aligned}$$

For the second innequality, we use the bound proposed in the appendix of Abadie et al. (2010). As in Abadie et al. (2010),, $E\left[\left|\sum_{j=2}^{N_0+1}w_j^*\sum_{s=1}^{t^*-k}\lambda_t\left(\sum_{n=1}^{t^*-k}\lambda_n^{'}\lambda_n\right)^{-1}\lambda_s^{'}\left(\eta_{js}^{SC}-\eta_{1s}^{SC}\right)\right|\right]\rightarrow 0$ as $t^*{-}k\rightarrow\infty$.

[34] Differently from Abadie et al. (2010), we do not include periods t^*-k+1 until t^* in the second term of the right hand side of expression 29. We do this modification because we leave these periods out from the estimation of weights. Therefore, it is as if we were treating these periods as a post-treatment.

Therefore, under the null that $\alpha_{1t} = 0$, we have that:

$$t^{SC} \to \frac{\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\left[\sum_{j=2}^{N_0+1} w_j^*\left(\eta_{jt}^{SC} - \eta_{1t}^{SC}\right)\right]^2}{\frac{1}{T-t^*+k}\sum_{t=t^*-k+1}^{T}\left[\sum_{j=2}^{N_0+1} w_j^*\left(\eta_{jt}^{SC} - \eta_{1t}^{SC}\right)\right]^2} \tag{30}$$

Assuming that η_{jt} is i.i.d. across t and have the same distribution up to a variance parameter across j, we can write $\eta_{jt} = a_j \times \tilde{\eta}_{jt}$, where $\tilde{\eta}_{jt}$ is i.i.d. across t and j. Note that we can re-write the predicted error for a given t as:

$$\sum_{j=2}^{N_0+1} w_j^*\left(\eta_{jt}^{SC} - \eta_{1t}^{SC}\right) = \sum_{j=1}^{N} \tilde{a}_j \tilde{\eta}_{jt}$$

for $\widetilde{a}_j \equiv a_j w_j^*$.

We want to show that the distribution t^{SC} does not depend on $(\tilde{a}_1, ..., \tilde{a}_N)$. Assuming η_{jt} is indepedent of (Z_j, μ_j) and is normally distributed with mean zero and variance a_j, so that $\sum_{j=1}^{N} \tilde{a}_j \tilde{\eta}_{jt} \sim N(0, \sum_{j=1}^{N} \tilde{a}_j^2)$, we have:

$$\begin{aligned} t^{SC} &\to \frac{\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\left[\sum_{j=1}^{N} \tilde{a}_j \tilde{\eta}_{jt}\right]^2}{\frac{1}{T-t^*+k}\sum_{t=t^*-k+1}^{T}\left[\sum_{j=1}^{N} \tilde{a}_j \tilde{\eta}_{jt}\right]^2} = \frac{\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\left[\sum_{j=1}^{N} \frac{\tilde{a}_j \tilde{\eta}_{jt}}{\sqrt{\sum_{j=1}^{N} \tilde{a}_j^2}}\right]^2}{\frac{1}{T-t^*+k}\sum_{t=t^*-k+1}^{T}\left[\sum_{j=1}^{N} \frac{\tilde{a}_j \tilde{\eta}_{jt}}{\sqrt{\sum_{j=1}^{N} \tilde{a}_j^2}}\right]^2} \\ &= \frac{\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} X_t}{\frac{1}{T-t^*+k}\sum_{t=t^*-k+1}^{T} X_t} \end{aligned} \tag{31}$$

where X_t, $t = t^* - k + 1, ..., T$ are i.i.d. chi-squared random variables. Therefore, the distribution of t^{SC} does not depend on $\tilde{a}_j$.

If we relax the normality and independence assumptions, then it is not possible to guarantee that the distribution of t^{SC} will not depend on $(\tilde{a}_1, ..., \tilde{a}_N)$. In this case, we need to assume that $E\left[\eta_{jt} \middle| Z_j, \mu_j\right] = 0$ and $E\left[\eta_{jt}^2 \middle| Z_j, \mu_j\right] = a_j^2$. Under these assumptions, we can show that $E[t^{SC}]$ does not depend on $(\tilde{a}_1, ..., \tilde{a}_N)$. This comes from the fact that t^{SC} has the form $\frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{m} Y_i}$, where $m > n$ and $Y_1, ..., Y_m$ are i.i.d. random variables. We also have that the first order approximation of the variance of t^{SC} is given by:[35]

$$var(t^{SC}) = \left[\frac{1}{T-t^*} - \frac{1}{T-t^*+k}\right]\left[2 + \frac{\sum_{j=1}^{N} a_j^4}{\left(\sum_{j=1}^{N} a_j^2\right)^2}\left(E[\tilde{\eta}_{jt}^4] - 3\right)\right]$$

Also note that the second order approximation for the expected value of the Abadie et al. (2010) test statistic is given by:

$$E(t^{SC}) = 3 + \frac{\sum_{j=1}^{N} a_j^4}{\left(\sum_{j=1}^{N} a_j^2\right)^2}\left(E[\tilde{\eta}_{jt}^4] - 3\right) \tag{32}$$

[35] We use the formula for the first order approximation of the variance given by $var\left(\frac{X}{Y}\right) = \frac{[E(X)]^2}{[E(Y)]^2} \times \left(\frac{var(X)}{[E(X)]^2} - 2\frac{cov(X,Y)}{E(X)E(Y)} - \frac{var(Y)}{[E(Y)]^2}\right)$ for random variables X and Y.