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

Ferman, Bruno and Pinto, Cristine

Sao Paulo School of Economics - FGV, Sao Paulo School of Economics - FGV

6 November 2015

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

Bruno Ferman†
Sao Paulo School of Economics - FGV

Cristine Pinto‡
Sao Paulo School of Economics - FGV

November, 2015

Abstract

Differences-in-Differences (DID) is one of the most widely used identification strategies in applied economics. However, inference in DID models when there are few treated groups is still an open question. We show that usual inference methods used in DID models might not perform well when there are few treated groups and residuals are heteroskedastic. In particular, when there is variation in the number of observations per group, we show that inference methods designed to work when there are few treated groups would 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 would 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) dataset to show that this problem is relevant even in datasets with large number of observations per group. Then we derive alternative inference methods that provide accurate hypothesis testing in situations of few treated groups and many control groups in the presence of heteroskedasticity (including the case of only one treated group). The main assumption is that we know how the heteroskedasticity is generated, which is the case when it is generated by variation in the number of observations per group. Finally, we also show that an inference method for the Synthetic Control Estimator proposed by Abadie et al. (2010) can correct for the heteroskedasticity problem, and derive conditions under which this inference method provides accurate hypothesis testing.

Keywords: differences-in-differences; inference; heteroskedasticity; clustering; few clusters; bootstrap

JEL Codes: C12; C21; C33

*We would like to thank Josh Angrist, Sergio Firpo, Bernardo Guimaraes, Lance Lochner, Vladimir Ponczek, Andre Portela, Vitor Possebom, Rodrigo Soares, Chris Taber, and Gabriel Ulyssea 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 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 standard errors (CRSE) due to Liang and Zeger (1986) at the group level.[1][2] By clustering at the group level, we allow for unrestricted correlation in the within group residuals. More specifically, we allow not only for correlation in the residuals of observations in the same group x time, but also for correlation in the residuals of observations in the same group at different time periods.[3] One important advantage of the CRSE is that it 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. CRSE take heteroskedasticity into account by essentially estimating the standard errors separately for the treated and for the control groups. Bertrand et al. (2004) show that CRSE and pairs-bootstrap at the group level work well when the number of groups is large. When there are only a small number of groups, it might still be possible to obtain tests with correct size even with unrestricted heteroskedasticity, especially when there is not much imbalance in the number of treated and control groups (Cameron et al. (2008), Brewer et al. (2013), Imbens and Kolesar (2012), Bell and McCaffrey (2002), 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) and Brewer et al. (2013)). The problem is that, with a small number of treated groups, the variance component related to the treated group would be severely underestimated. In the polar case where there is only one treated group, the estimate of this component would be identical to zero.[4]

[1] 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 residuals of individuals within a group can be correlated while residuals of individuals in different groups are uncorrelated.

[2] For example, Bedard and Do (2005), Choi (2011), and Pettersson-Lidbom (2012).

[3] 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.

[4] Another alternative presented by Bertrand et al. (2004) is to collapse the pre and post information. This approach would

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) deal with the case when the number of treated and control groups are small. They use small sample inference procedures under the assumption that residuals in the group x time DID aggregate model are normal, homoskedastic, and serially uncorrelated. Conley and Taber (2011) 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. The main idea of their method is to use information on the residuals of the control groups to estimate the distribution of the DID estimator under the null. Residuals-bootstrap provides another alternative when there are few treated clusters (Cameron et al. (2008)). In residuals-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 variance is homoskedastic, so that we can use information on the variance 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, residuals 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 averages. In a recent paper, MacKinnon and Webb (2015a) propose an alternative method for the case of few treated groups under heteroskedasticity. Their main idea is a permutation test where they compare t-statistics calculated using CRSE. This method works well when there are enough treated and control groups. However, it will fail when there are very few treated groups. In particular, their method will be almost the same as Conley and Taber (2011) method when there is only one treated group. The main problem with this method is that the CRSE will be underestimated when there are very few treated groups.

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 the component of the variance related to the treated groups would be underestimated. 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 there is variation in the number of observations per group, these methods would tend to (under-) over-reject the null hypothesis when the number of observations of the treated groups is (large) small relative to the number of observations of the control groups. The main

take care of the auto-correlation problem. However, in order to allow for heteroskedasticity, one would have to use robust standard errors. In this case, this method would also fail when there are few treated groups.

idea is that variation in the number of observations per group would invalidate the assumption that residuals are i.i.d. across groups, because larger groups would have lower variance. The intuition of this result was already exposed in Assuncao and Ferman (2015) in an application of Conley and Taber (2011) method.[5] Here we formalize this idea and derive conditions under which this problem would be more or less relevant. In particular, we show that this problem becomes more severe when the intra-cluster correlation is smaller and when there are fewer observations per group. Then we 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 number of observations per group. This happens because when the intra-cluster correlation goes to 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 (which is one of the justifications used by Donald and Lang (2007), pp. 224).

We then derive alternative methods for inference when there are only few treated groups (including the case of only one treated group) that take into account the fact that residuals are inherently heteroskedastic when there is variation in the number of observations per group. The main assumption is that we know how the heteroskedasticity is generated, which is the case when it 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 derive the distribution of the residuals for the treated groups. Our simulations show that these corrections imply in hypothesis testing with correct sizes when the number of control groups is large. We also provide a refinement of our method using residuals-bootstrap on a pivotal statistics with our heteroskedasticity correction that provided more accurate hypothesis testing when the number of control states is not that large (for example, with 1 treated and 24 control states) in our simulations.

Finally, we show that Synthetic Control, an alternative estimation method for the case of one treated group proposed by Abadie et al. (2010), can provide accurate hypothesis testing even in presence of heteroskedasticity. This happens because, under some circumstances, one of the inference methods proposed in Abadie et al. (2010) turns out to correct for the presence of heteroskedasticity by using information from

[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 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.*"

the pre-treatment period. We derive the conditions under which this method provides accurate hypothesis testing. One important scenario that Abadie et al. (2010) does not correct for heteroskedasticity (and our method does) is when there is only one pre-treatment period.

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 alternative inference methods that are valid in this scenario, and present the conditions under which the inference method for Synthetic Control proposed by Abadie et al. (2010) provide accurate hypothesis testing in the presence of heteroskedasticity. In Section 3 we perform Monte Carlo simulations to examine the performance of existing inference methods and to compare that to the performance of our corrected inference methods. In Section 4 we compare the different inference methods by simulating placebo laws in a real dataset with a large number of observations: the American Community Survey (ACS). We conclude in Section 5.

2 Empirical Model

2.1 A Review of Existing Methods

We consider a group x time DID aggregate model:

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

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 effects for group j, while γ_t is a time fixed effect; η_{jt} is a group x time random variable that might be correlated over time, but uncorrelated across groups. Depending on the application, groups might stand for, for example, 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's 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 $N_1\to\infty$ and $N_0\to\infty$. As shown in Conley and Taber (2011), 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 $N_0\to\infty$.

The variance of the DID estimator, under the assumption that η_{jt} are independent across j, will be 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 comparison and the variance of the control groups comparison. We allow for any kind of auto-correlation between η_{jt} and $\eta_{jt'}$.

When there are many treated and control groups, Bertrand et al. (2004) suggest that CRSE at the group level works well, as this method allows for unrestricted auto-correlation in the residuals η_{jt}, and for heteroskedasticity in the residuals. The CRSE has a very intuitive formula in the DID framework:[6]

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

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

With CRSE we calculate each component of the variance of the DID estimator separately. In other words, we use the treated groups residuals to calculate the component related to the treated groups, and

[6]The clustered-robust variance matrix 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).

the control groups residuals to calculate the component related to the control groups. While CRSE are 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)). Finally, it is also problematic to implement heteroskedasticity-robust bootstrap methods such as pairs-bootstrap and wild bootstrap when there are few treated groups. In pairs-bootstrap, there will be a high probability that the bootstrap sample will not include a treated unit. In wild bootstrap, the idea is to generate variation in the residuals of each j by randomizing whether its residual will be $\hat{\eta}_{jt}$ or $-\hat{\eta}_{jt}$. However, if we go again to the extreme case with only one treated, then $\widehat{W_1} = 0$. Therefore, the wild bootstrap would not generate variation in the treated group.

It is clear then that the inference problem in DID models with few treated groups lies essentially on how to estimate the component of the DID estimator variance related to the treated group using $\hat{\eta}_{jt}$. Alternative methods use information on the control groups residuals in order to estimate the component of the variance related to the treated groups. These methods, however, rely on specific assumptions on the residuals. Donald and Lang (2007) assume that the group x time residuals are normal, homoskedastic, and serially uncorrelated. Under these assumptions, the variance of $\hat{\alpha}$ becomes:

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

where $var(\eta_{jt}) = \sigma_\eta^2$ and p is the proportion of treated groups. Therefore, under these assumptions, one could easily recover the variance of $\hat{\alpha}$ by estimating σ_η^2 using the estimated residuals $\hat{\eta}_{jt}$. As suggested by Donald and Lang (2007), if NT is small, then one should compare the test statistic $t = \hat{\alpha}/\sqrt{var(\hat{\alpha})}$ to the student-t distribution instead of calculating the critical values based on the normal distribution. The assumption that residuals are serially uncorrelated, however, might be unappealing in DID applications (Bertrand et al. (2004)).

Conley and Taber (2011) provide an interesting alternative inference method that allows for unrestricted auto-correlation in the residuals. The main idea of their method is to use information on the residuals of the control groups to estimate the distribution of the DID estimator under the null. In the simpler case with only one treated group, $\hat{\alpha} - \alpha$ would converge to W_1 when $N_0 \to \infty$. 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 Conley and

Taber (2011) relax the assumption of no auto-correlation, it requires that residuals are i.i.d. across groups (as do Donald and Lang (2007)), so that $\{\widehat{W}_j\}_{j=2}^{N_0+1}$ approximates the distribution of W_1 when $N_0 \to \infty$.

Finally, residuals-bootstrap methods resample the residuals while holding the regressors constant throughout the pseudo-samples. Therefore, 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 the residuals are homoskedastic. It is important to note that bootstrap alternatives with asymptotic refinements that focus on pivotal test statistics would not work well in situations of few treated groups. This happens because these methods require a consistent estimator of the variance. However, with N_1 fixed, the heteroskedasticity-robust methods to estimate the variance would not work properly.

2.2 The Heteroskedasticity Problem

As seen in Section 2.1, CRSE in DID models with few treated groups severely underestimate the variance of $\hat{\alpha}$. Alternative methods such as Donald and Lang (2007), Conley and Taber (2011) and residuals-bootstrap require strong distributional assumptions on the residuals. In particular, they require homoskedasticity. In this section, we show that these methods might not perform well in the presence of heteroskedasticity. In particular, 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.

We start with an 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 random effect (possibly correlated over time), and ϵ_{ijt} is an individual level residual. The main feature that defines a "group" in this setting is the assumption that residual $(\nu_{jt} + \epsilon_{ijt})$ of two individuals in the same group might be correlated, while residuals 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 corrections will require weaker assumptions in the residuals, as will be presented in Section 2.3.

In this case, 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 residuals 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:

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

for constants A and B, regardless of the auto-correlation of ν_{jt}. 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, $var(W_j) = A + \frac{B}{M_j}$.

This heteroskedasticity in the residuals 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 (overestimate) 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 residual would dominate. In this case, if we assume that the group x time random effect ν_{jt} is i.i.d., then $\frac{var(W_j)}{var(W_j')} \to 1$, which implies that control groups residuals would be a good approximation for the distribution of the treated groups residuals even when the number of observations in each group is different. This is one of the main rationales used in Donald and Lang (2007) 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 $\frac{var(W_j)}{var(W_{j'})}$ would remain constant (unless $\frac{M_j}{M_{j'}} \to 1$), which implies that the aggregate model would still be heteroskedastic even asymptotically. Therefore, Conley and Taber (2011), Donald and Lang (2007), and residuals-bootstrap methods 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, unless the intra-group correlation is large.

2.3 Corrected Inference Method

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. CRSE estimate this component of the variance without using information from the control groups. While this approach has the appealing property of allowing for unrestricted heteroskedasticity in the residuals, it is unfeasible when the number of treated groups is small. On the other extreme, other methods method surpass the problem of few treated groups by using information from the control groups. The problem with these approaches is that they require that residuals are homoskedastic.

In this section, we derive inference methods that use information from the control groups to estimate the variance of the treated groups while allowing for heteroskedasticity. The main assumption is that we know how the heteroskedasticity is generated, which is the case when 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 derive the distribution of the residuals for the treated groups. While we motivate our methods based on heteroskedasticity generated by variation in the number of groups, it is important to note that our methods are more general. The main assumption will be that we know the structure of the heteroskedasticity.

We derive first an extension of Conley and Taber (2011) method that corrects for heteroskedasticity. For ease of exposition, we consider the simpler case with only $j = 1$ treated, although our methods can be extended for any number of treated groups. In Theorem 1 in Appendix A, we show that, if we knew the variance of each random variable W_j, then we could re-scale each observed $\hat{W}_j$ by $\tilde{W}_j = \hat{W}_j \sqrt{\frac{var(W_1)}{var(W_j)}}$ so that all $\tilde{W}_j$ have the same variance as W_1, and use Conley and Taber (2011) approach with the re-scaled

residuals. The main assumption we need is that $\{\eta_{j1}, ..., \eta_{jT}\}$ is independent across j and have the same distribution up to the variance parameter. We also assume that $var(W_j)$ is a function that depends only on M_j, $G(M_j)$. Our proposed inference methods consist in estimating the variance of W_j as a function of the number of observations in group j (M_j), and then re-scaling the residuals used to estimate the distribution of W_1. Therefore, given an estimate $\widehat{G(M)}$, one would simply have to calculate $\tilde{W}_j = \hat{W}_j \sqrt{\frac{\widehat{G(M_1)}}{\widehat{G(M_j)}}}$, and then reject the null if the point estimate $\hat{\alpha}$ is (lower) greater than the (5^{th}) 95^{th} percentile of the distribution of $\{\tilde{W}_j\}_{j=2}^{N_0+1}$, for a test with 10% significance level.[7] In Theorem 2 in Appendix A, we show that this approach works asymptotically when $N_0 \to \infty$ if we have a consistent estimator for $G(M)$.

We propose a consistent estimator for function $G(M)$ using group x time aggregate data. We assume that $var(W_j) = A + \frac{B}{M_j}$, for constants A and B. The structure of the residuals we assumed in Section 2.2 imply this structure. However, this assumption is more general. In particular, it is important to note that we do not have to make any assumption on the auto-correlation of η_{jt}. Given this assumption, we can run a regression of $\hat{W}_j^2$ on $\frac{1}{M_j}$ and a constant, and then use the predicted $\widehat{G(M_j)}$. We show in Theorem 3 in Appendix A that this estimator is consistent. 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 group x time averages.

One important point is that this method should only provide an accurate hypothesis testing procedure when N_0 is large enough. Therefore, we consider a pivotal test statistics and use residuals-bootstrap with our heteroskedasticity correction to recover its distribution, which should provide a better finite sample approximation as suggested in the literature (see Davison and Hinkley (1997), Cameron et al. (2008), and Cameron and Miller (2015)). To calculate the pivotal statistic, we use the finite N_0 formula for $var(\hat{\alpha})$, which is given by:

$$\begin{aligned} var(\hat{\alpha}) &= var(W_1) + \frac{1}{(N_0)^2} \sum_{j=2}^{N_0+1} var(W_j) \\ &= G(M_1) + \frac{1}{(N_0)^2} \sum_{j=2}^{N_0+1} G(M_j) \end{aligned} \tag{10}$$

Given our estimates $\hat{\alpha}$ and $\widehat{G()}$, we calculate $\hat{s} = \frac{\hat{\alpha}}{\sqrt{\widehat{var(\hat{\alpha})}}}$ and use bootstrap to approximate this distribution. More specifically, we calculate from the aggregate DID regression the predicted values of Y_{jt} and η_{jt},

[7] If we want a test with the null $H_0 : \alpha = \alpha_0$, we would simply have to compare $\hat{\alpha} - \alpha_0$ (instead of $\hat{\alpha}$) to the distribution $\{\tilde{W}_j\}_{j=2}^{N_0+1}$.

so that we can calculate $\nabla\hat{Y}_j = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\hat{Y}_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\hat{Y}_{jt}$ and $\hat{W}_j = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\eta_{jt}$. Then we normalize the residuals so that they have variance equal to one, $\widetilde{\widehat{W}}_j = \frac{\widehat{W}_j}{\sqrt{\widehat{G(M_j)}}}$. By bootstrap, we obtain $\mathcal{B}$ resamples of the aggregate residuals $\hat{W}_j$. In each each bth sample, we re-scale the bootstrapped residual, so that they have the same variance structure as the original sample, $\widetilde{\widehat{W}}^*_{j,b} = \widetilde{\widehat{W}}_{j,b}\sqrt{\widehat{G(M_j)}}$. We calculate the bootstrap estimate as:

$$\hat{\alpha}_b = \left(\nabla\hat{Y}_1 + \widetilde{\widehat{W}}^*_{1,b}\right) - \frac{1}{N_0}\sum_{j=2}^{N_0+1}\left(\nabla\hat{Y}_j + \widetilde{\widehat{W}}^*_{j,b}\right) \tag{11}$$

We then re-estimate the $G(M)$ function using the bootstrapped residuals $\widetilde{\widehat{W}}^*_{j,b}$, and calculate the bootstrapped variance of $\hat{\alpha}_b$ using formula 10. Our bootstrapped test statistics will be given by:

$$\hat{s}_b = \frac{\hat{\alpha}_b - \hat{\alpha}}{\sqrt{\widehat{var(\hat{\alpha}_b)}}} \tag{12}$$

If s is (greater) lower than the (95th) 5th percentile of the bootstrap distribution, then we reject the null hypothesis at 10% significance level. We also consider a residual-bootstrap with heteroskedasticity correction on the parameter $\hat{\alpha}$.

2.4 Alternative Inference Methods

An alternative inference method for the case of few treated groups under heteroskedasticity was proposed by MacKinnon and Webb (2015a). Their main idea is a permutation test where they compare t-statistics (rather than the estimator itself). This method works well when there are enough treated and control groups. However, it will fail when there are very few treated groups because they need to estimate the variance of the estimator to construct the t-statistic. The problem is that the heteroskedasticity-robust methods to estimate the variance would be biased with only a few treated groups. In particular, their method collapses to Conley and Taber (2011) method when there is only one treated group. The reason is that the CRSE 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. Therefore, there would be no correction relative to a permutation test on the estimator itself. In contrast, our method works even when there is only one treated group.

2.5 Alternative Estimation Methods - Synthetic Control

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. 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 $\hat{Y}_{1t}^N = \sum_{j=2}^{N_0+1} \hat{\omega}_j Y_{jt}$, where the weights $\hat{\omega}_j$ are estimated so that the differences between actual and estimated pre-treatment outcomes (Y_{1t} and $\hat{Y}_{1t}^N$) and covariates (X_{1t} and $\hat{X}_{1t}^N$) are minimized.[8] In the Synthetic Control approach, we need to decide which variables we want to include to estimate the weights $\hat{\omega}_j$. Particularly important for our application, one can either include the Y_{jt} for all pre-treatment t, or can leave some of the pre-treatment Y_{jt} out.

The inference method suggested in Abadie et al. (2010) is a permutation test where we estimate placebo regressions using each of the control units as a placebo treatment. In essence, this is the same as what Conley and Taber (2011) method does 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 we divide by the pre-treatment MSPE, we correct for the fact that the Synthetic Control estimators for this placebo group would always be large.

It turns out that, in some cases, looking at this ratio provides proper hypothesis testing under heteroskedasticity. For simplicity, consider that we have 3 periods, two before the treatment and one after the treatment. Suppose that we construct our Synthetic Control estimator using only the outcome variable in period 1. Under the Synthetic Control assumptions, when we consider the j unit as the placebo group, then the difference $Y_{j1} - Y_{j1}^N$ will be close to zero, since the weights used to construct Y_{j1}^N were chosen to minimize this difference, while $Y_{jt} - Y_{jt}^N$ would be approximately the residual η_{jt}, for $j = 2, 3$.[9] Therefore, when we look at the post/pre-intervention RMSE ratio, it will be close to $\frac{var(\eta_{j3})}{var(\eta_{j2})}$. Under our assumption that $\{\eta_{jt}\}_{t=1}^T$ is identically distributed across j up to the variance parameter, this ratio would be constant

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

[9] The difference $Y_{j1} - Y_{j1}^N$ will not, in general, be identical to zero because we require that Y_{j1}^N be a convex combination of the outcomes of the other groups.

for all j. This is why Abadie et al. (2010) inference method corrects the information from the control groups residuals so that they become comparable to the treated group residuals.[10]

However, this approach would not work properly if there is only one pre-treatment period. In this case, one would have to estimate the weights using the single pre-treatment period, which implies that the denominator would not be the variance of η_{jt}. We could still calculate the RMSE ratio, since $Y_{j1} - Y_{j1}^N$ will not be identical to zero. However, this division would not re-scale the numerator correctly. The same problem applies when we have more than one pre-treatment period but include all pre-treatment periods to estimate the weights. It is also important to note that Abadie et al. (2010) placebo graphical analyses (Figures 4 to 7 in Abadie et al. (2010)) would still suffer from the heteroskedasticity problem we highlight in this paper. An easy way to fix to this problem is to divide each placebo estimate by the squared root of its pre-treatment RMSE and multiply it by the squared root of the the pre-treatment RMSE of the treated group.

3 Monte Carlo Evidence

In this section we provide Monte Carlo evidence of different hypothesis testing methods in DID. We also simulate the inference method for Synthetic Control models proposed by Abadie et al. (2010) in Section 3.2. We assume that the underlying data generating process (DGP) is given by:

$$Y_{ijt} = \nu_{jt} + \epsilon_{ijt} \tag{13}$$

In most of the simulations, we estimate a DID model given by equation 6 where $j = 1$ 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:

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

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

[10] Note that if we had more than one post period and/or more than one pre period not included in the estimation of ω_j, then the only modification is that we would have the sum of variances of η_{ij} in the numerator and in the denominator. Therefore, the ratios would remain constant, so that our rationale still applies.

period from a discrete uniform random variable with range $[\underline{M}, \overline{M}] \in \{[50, 200], [200, 800], [50, 950]\}$.[11]

For each case, we simulated 40,000 estimates. Note that we will 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. Since the estimated component of the variance related to the treated group is zero, these methods always severely over-reject the null.[12]

3.1 Inference in DID Models

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 10% significance using robust standard errors in the individual level DID regression. The rejection rate for a 10% significance level test is only slightly higher than 10% when the intra-group correlation is small (10.8% when $\rho = 0.01\%$), but increases sharply for larger values of the intra-group correlation. Rejection rate is almost 50% when $\alpha = 4\%$.

When we use Conley and Taber (2011) method, average rejection rate for a 10% significance level test is always around 10% (column 3). However, this average rejection rate hides an important heterogeneity with respect to the number of observations in the treated group (M_1). Column 4 shows the difference in rejection rates when the number of individual observations in the treated group is above the median compared to the case when it is below the median. When $\rho = 0.01\%$, the difference in rejection rates is around 11 percentage points. Therefore, although Conley and Taber (2011) method rejects the null on average in 10% of the cases, this happens because it over-rejects the null when the treated group is small while it under-rejects the null when the treated group is large. We show in more detail the relationship between rejection rates and the number of observations in the treated group in Figure 1.A for the case $\rho = 0.01\%$. Rejection rate is around 22% when the treated group is in the first quintile of number of observations per group, while it is only 4% when the treated group is in the fifth quintile. Note also that this distortion in rejection rates is not confined to the extremes of the distribution of group sizes. Rejection rates are 13.5% when the treated group is in the second quintile of number of observations per group, and 5.2% when it is in the fourth quintile. In columns 5 and 6 we show that Donald and Lang (2007) method suffer from exactly the same problem, despite the

[11] In the Monte Carlo simulations, we always consider the case $M(j,t) = M_j$.

[12] We also do not include MacKinnon and Webb (2015a) method in the simulations because their method collapses to Conley and Taber (2011) method when there is only one treated group.

fact that the distributional assumptions in their method are valid in our simulations, except for the fact that there is variation in the number of observations per group.

As expected, this heterogeneity 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, when $\rho = 4\%$ the difference in rejection rates by number of observations in the treated group remains relevant. In both Conley and Taber (2011) and Donald and Lang (2007) methods, the difference in rejection rates when the number of observations in the treated groups is above or below the median is around 2 percentage points. We present Conley and Taber (2011) rejection rates in more detail for the case $\rho = 4\%$ in Figure 1.B. Rejection rates are 11.8% when the treated group is in the first quintile of number of observations per group, while it is 8.5% when the treated group is in the fifth quintile.

Given that inference using Conley and Taber (2011) and Donald and Lang (2007) methods is problematic when there is variation in the number of observations per group, we consider our alternative inference methods that correct for the heteroskedasticity problem in the group x time regression. In columns 5 and 6 of Table 1 we present results from our correction when we estimate the $G(M)$ function using group x time data. We run an OLS regression of $\hat{W}^2$ on a constant and $\frac{1}{M_j}$, which provide us a consistent estimator of $G(M)$, and then use $\sqrt{\frac{\widehat{G(M_1)}}{\widehat{G(M_j)}}}$ to re-scale the residuals $\hat{W}_j$.[13] Average rejection rates using our method are only slightly higher than 10% (ranging from 10% to 10.8%) and, more importantly, rejection rates become homogeneous across the number of observations in the treated group. The inference method we propose provides a reasonably accurate hypothesis testing regardless of the value of the intra-group correlation. We present in Figures 1.C and 1.D rejection rates in more detail using our inference method for the cases $\rho = 0.01\%$ and $\rho = 4\%$, respectively. Rejection rates are always very close to 10% regardless of the quintile of M_1.

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. Intuitively, this happens because robust OLS standard errors do not take into account that the increase in the number of observations are not independent. When we consider Conley and Taber (2011) and Donald and Lang (2007) methods, increasing the number of observations per group ameliorates the problem of (over-

[13] In these simulations, we excluded the treated observation from the estimation of $G(M)$ since $\hat{W}_1^2 = 0$ by construction. Note, however, that this is not crucial, since the estimator of $G(M)$ remains consistent whether or not we include the treated observation.

) under-rejecting the null when M_1 is (small) large relative to the number of observations in the control groups. In particular, when $\rho = 4\%$ there is no significant difference in rejection rates between those with M_1 above and below the median. 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 $\rho = 0$, Conley and Taber (2011) and Donald and Lang (2007) methods would face this heteroskedasticity problem even when $M \to \infty$.

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 simulation in panel B. As expected, more variation in the number of observations per group worsens the inference problem we highlight in Conley and Taber (2011) and Donald and Lang (2007) methods. On the contrary, our proposed inference methods remain accurate irrespective of the variation in the number of observations per group.

We present in Tables 2 and 3 the simulation results when the total number of groups are, respectively, 100 and 50. Conley and Taber (2011) and Donald and Lang (2007) continue to face a problem of differential rejection rates when the treated group is small or large. In addition to this problem, Conley and Taber (2011) method also shows an average rejection rate higher than 10%. Conley and Taber (2011) method has a rejection rate of around 11% when $N = 100$ and around 12.5% when $N = 50$.[14] Donald and Lang (2007) method does not face this additional problem of over-rejection, although it is possible that this happens because the residuals in our simulations are normally distributed. While our correction method continues to solve the problem of differential rejection rates irrespectively of N_0, we face the problem of higher average rejection rates as do Conley and Taber (2011). These results highlight the importance of the number of control groups for our and Conley and Taber (2011) methods, as these results are only valid asymptotically when $N_0 \to \infty$. It is important to note that these methods over-reject the null even for numbers of groups that are considered large enough in the literature to conduct inference with CRSE (Bertrand et al. (2004), and Angrist and Pischke (2009)).

We consider, therefore, a pivotal test statistics and use residuals-bootstrap with our heteroskedasticity correction to recover its distribution, as explained in Section 2.3. While this method also relies on $N_0 \to \infty$,

[14]This problem does not arise because we introduced variation in M_j in our simulations. Conley and Taber (2011) would continue to face this problem even with constant M_j, which means that all of their assumptions would be valid.

there is evidence that it should provide a better finite sample approximation (see Davison and Hinkley (1997), Cameron et al. (2008), and Cameron and Miller (2015)). For the sake of comparison, we start presenting in Panel A of Table 4 rejection rates for the standard residuals-bootstrap without our heteroskedasticity correction. Note that it is not possible to bootstrap on a pivotal statistic because the CRSE will not be a consistent estimator of the variance when there is only one treated. Average rejection rates range from very close to 10% when N is large, to around 12.5% when $N = 25$. While these numbers do not look particularly bad, they hide exactly the same problem as Donald and Lang (2007) and Conley and Taber (2011) methods, with over-rejection when the treated group is small, and under-rejection when the treated group is large.

We then present in Panel B of Table 4 rejection rates of our residuals-bootstrap inference method with heteroskedasticity correction but without asymptotic refinement (where we bootstrap the distribution of $\hat{\alpha}$), while in Panel C we present rejection rates using a residuals-bootstrap with our heteroskedasticity correction and with asymptotic refinement (where we bootstrap the distribution of $\hat{s} = \hat{\alpha}/\sqrt{\widehat{var(\hat{\alpha})}}$). Rejection rates using our inference method with asymptotic refinement are always closer to 10% when compared to alternative methods. When $N = 400$, both methods (with and without asymptotic refinement) provide rejection rates virtually equal to 10%. For smaller N, there are important improvements in hypothesis testing when we use the method with asymptotic refinement. When $N = 100$, rejection rates are around 10.6% (compared to 11.4% without refinement), when $N = 50$, rejection rates are around 10.7% (compared to 12.2% without refinement), and when $N = 25$, rejection rates are around 11.1% (compared to 14.1% without refinement). In addition, our heteroskedasticity correction significantly improves the dependence of rejection rates with respect to the relative size of the treated group. When $N = 400$, there is virtually no difference in rejection rates for treated groups above and below the median number of observations. When $N = 25$, our method rejects slightly more when for larger treated groups when $\rho = 0.01\%$ (0.8 percentage points). This number, however, should be compared to the 16 percentage points difference in rejection rates with the residuals-bootstrap without correction. Therefore, our inference method with asymptotic refinement provides a significant improvement relative to alternative methods, providing reasonably good hypotheses testing even when the number of control groups is not that large.

3.2 Inference in Synthetic Controls

An alternative estimation method when there is only one treated group is to use the Synthetic Control Estimator. As explained in Section 2.5, one inference method suggested in Abadie et al. (2010) compares the ratio of post/pre-treatment RMSE of the Synthetic Control Estimator and compares it to the same ratio

when we use the control groups as placebo treatments. We present in Panel A of Table 5 rejection rates for the case with $T = 2$, with one pre and one post-intervention periods. Rejection rates are higher when the treated group is small when $\rho = 0.01\%$. This happens because the post-treatment RMSE used in the numerator is higher when the treated group is smaller, due to the heteroskedasticity generated by the variation in the number of observations per group. However, the pre-treatment RMSE used in the denominator is just an error term reflecting the fact that Y_{11}^N will not be identical to Y_{11} because we restrict to convex combinations of the control groups, so the ratio will be decreasing in M_1. When ρ is higher, then a given variation in the number of observations per group generates less heteroskedasticity, so this effect is weaker. Exactly the same pattern happens in Panel B, where we simulate a case with $T = 3$ with 2 pre-treatment periods, but include both Y_{j1} and Y_{j2} to estimate the weights.[15]

In Panel C, we consider again the case with $T = 3$ periods, but now we use only Y_{j1} to estimate the weights. In this case, the pre-treatment RMSE used in the denominator is higher when the treated group is smaller, since it includes the predicted error related to the pre-treatment period $t = 2$. As explained in Section 2.5, while both the numerator and the denominator decrease with M, the ratio will be constant under the assumption that the residuals are i.i.d. across groups up to the variance parameter (note that we allow for unrestricted auto-correlation across time within group). This implies that the difference in rejection rates for small and large groups is corrected using this inference method. The only detail is that rejection rates are slightly *lower* when the treated group is small. This happens because when the treated group is small, it will be more likely that it will not be possible to provide a good fit for the treated group. In this case, the pre-treatment RMSE will be larger. Again, this problem will be less relevant when ρ is larger, since this implies that variation in M generates less heteroskedasticity.

4 Simulations with Real Datasets

To illustrate the magnitude of this problem, we also conduct simulations of placebo interventions using a real dataset: the American Community Survey (ACS). We extract information on employment status earning for women between ages 25 and 50 for the years 2005 to 2013. We consider two different group levels based on the geographical local 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 $N_0 \to \infty$, while simulations using placebo interventions at the state level would mimic situations of

[15] When $N = 25$, average rejection rate is 12%. This, however, is just a consequence from the fact that we have only 25 estimates by changing the treated group.

DID designs that are commonly used in applied work, where one state is treated while all the other states are used as control.

We consider pairs of two consecutive years and estimate placebo DID regressions using one of the groups (PUMA or state) at a time. Note that this differs from Bertrand et al. (2004) simulations, since we are defining only one group to be treated, while they randomly selected half of the states to be treated. 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[16]. There are, on average, 730 observations in each PUMA x time cell. This number, however, hides an important heterogeneity in cell sizes. As presented in column 1 of Table 6, the 10th percentile of PUMA x time cell sizes is 171, while the 90th percentile is 1,337. For the state level simulations, we have $51 \times 8 = 408$ regressions (we include Washington, D.C.). Again, there is substantial heterogeneity in state x time cell sizes. As presented in column 2 of Table 6, while the average cell size is 10,138, the 10th percentile is 1,250, while the 90th percentile is 21,099.

For each placebo DID regression, we test the null hypothesis that the "intervention" has no effect ($\alpha = 0$) using robust standard errors, Conley and Taber (2011) method, Donald and Lang (2007) method, and our two corrected methods. Since we are looking at placebo interventions, if the hypothesis testing is correct, then we would expect to reject the null roughly 10% of the time for a test with 10% significance level. We present in Panel A of Table 7 rejection rates in simulations results using PUMAs as the group level, while in Panel B we present results using states as the group level. Results in columns 1 show that robust standard errors in the OLS individual level DID regression, that assume that all individual errors are independent, would tend to over-reject the null hypothesis. In particular, we reject the null at 10% significance level, on average, around 13%-14% of the time, in both the PUMA and the state level simulations.[17]

We present in columns 3 to 6 of Table 7 rejection rates for Conley and Taber (2011) and Donald and Lang (2007) inference methods. The results are very similar to our Monte Carlo simulations presented in Section 3. When we consider the PUMA level simulations, both methods over-reject the null when the treated group is small, and under-reject the null when the treated group is large. When we look at state level simulations, Conley and Taber (2011) method also over-rejects the null on average, which is again consistent with our Monte Carlo results. What is most remarkable, however, is that this problem of rejection rates varying with the size of the treated group is extremely relevant even in a dataset with a very large number of observations:

[16] Information on PUMA of residence is only available for ACS data after 2005.

[17] Clustered standard errors (whether at group or group x time level) perform very poorly in this situation. Rejection rates are always greater than 80% (results not shown). This was expected, since our simulations have only one treated group (Bertrand et al. (2004), and Wooldridge (2003)).

when we consider the state level simulations, average number of observations per group x time is greater than 10,000. When we consider the simulations with employment status as outcome variable, Conley and Taber (2011) method would have a rejection rate of 21% if the number of observations in the treated group is below the median, while it would have a rejection rate of 2% when it is above the median.

Given that the existing inference methods do not perform well in this situation, we now consider our corrected inference methods. When we apply our simpler correction method (columns 7 and 8 of Table 7) in the PUMA level simulations, the test is very accurate, rejecting around 10% of the time irrespectively of the treated group size. In the state level regressions ($N_0 + 1 = 51$), our simpler method present an average rejection rate of around 12%, which again is consistent with our Monte Carlo simulations. In columns 10 and 11 we present rejection rates with our residuals-bootstrap method with asymptotic refinement. With this method, we are able to achieve a rejection rate closer to 10%, and we cannot reject the null that there is no variation in rejection rates across M_1.

5 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. In particular, we show that, methods designed to work when there are few treated groups would tend to (under-) over-reject the null hypothesis when the number of observations of the treated groups is (large) small relative to the number of observations of the control groups. A notable exception is the inference method proposed by Abadie et al. (2010) for the Synthetic Control Estimator. This method takes heteroskedasticity into account provided that there is at least one pre-intervention period not included in the estimation of the Synthetic Control weights. Therefore, it is not possible to use this inference method to correct for heteroskedasticity when there is only one pre-treatement period. The inference methods we derive provide an alternative solution to the heteroskedasticity problem in DID models with few treated groups when the number of control groups is large. In particular, our methods work even when there is only one treated group and only one pre-treatment period. A refinement of our method using residuals-bootstrap also provided reasonably accurate hypothesis testing in our simulations when the number of control groups is around 25.

Finally, it is important to point out that our inference method for correcting for heteroskedasticity is more general than the main case we analyzed in this paper, in which the heteroskedasticity is generated by variation in the number of observations per group. In fact, as long as we are able to assume a structure of

the residual variances, we are able to apply our method. There are other applications where the variance of W_j might vary by group even when all groups have the same size. This would happen when, for example, Y_{ijt} is a binary variable and average Y_{jt} might be closer or farther away from 0.5 depending on j.

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.

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.*

Bedard, Kelly and Chau Do, "Are Middle Schools More Effective?: The Impact of School Structure on Student Outcomes," *Journal of Human Resources*, 2005, *40* (3).

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 Colin and Douglas L Miller, "A practitioners guide to cluster-robust inference," *Journal of Human Resources*, 2015, *50* (2), 317–372.

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.

Choi, Moonkyung Kate, "The impact of Medicaid insurance coverage on dental service use," *Journal of Health Economics*, 2011, *30* (5), 1020–1031.

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.

Davison, A.C. and D.V. Hinkley, *Bootstrap Methods and Their Application* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1997.

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.

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.

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.

Pettersson-Lidbom, Per, "Does the size of the legislature affect the size of government? Evidence from two natural experiments," *Journal of Public Economics*, 2012, *96* (3), 269–278.

van der Vaart, A. W., *Asymptotic statistics* Cambridge series in statistical and probabilistic mathematics, Cambridge (UK), New York (N.Y.): Cambridge University Press, 1998. Autre tirage : 2000 (dition broche), 2005, 2006, 2007.

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.

Figure 1: **Rejection Rates in Monte Carlo Simulations by Quintiles of M_1 ($H_0 : \alpha = 0$ at 10% significance level)**

Conley and Taber (2011) Method

Figure 1.A: $\rho = 0.01\%$

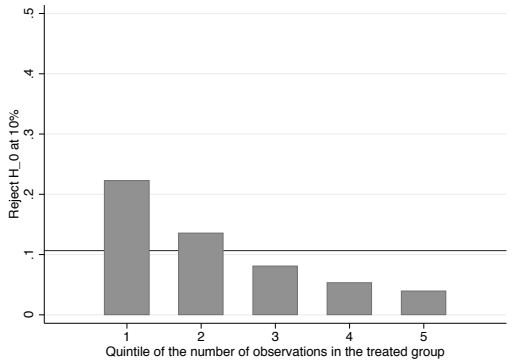
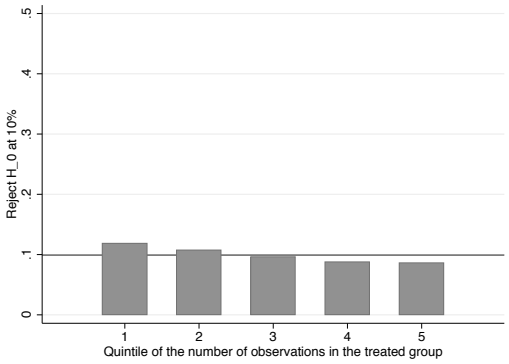


Figure 1.B: $\rho = 4\%$



Corrected Method

Figure 1.C: $\rho = 0.01\%$

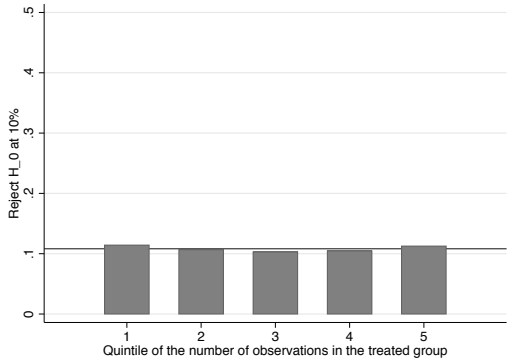
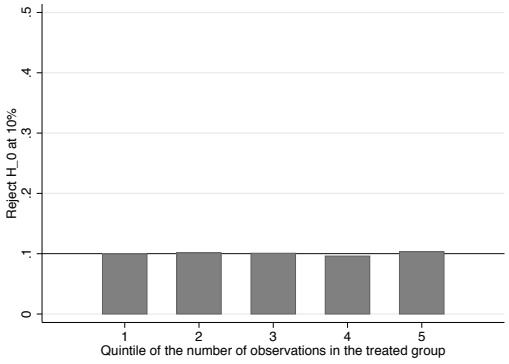


Figure 1.D: $\rho = 4\%$



Notes: These figures present the rejection rates by quintile of the number of observation of the treated group when$N_0 + 1 = 400$ and $M \in [50, 200]$. These rejection rates are based on Monte Carlos simulations explained in Section 3. Figures 1.A and 1.B present results using Conley and Taber (2011) inference method, while Figures 1.C and 1.D presents results using the corrected method proposed in this paper.

Table 1: **Rejection Rates in MC Simulations with** $N_0 + 1 = 400$ ($H_0 : \alpha = 0$ at 10% significance level)

	Inference Method							
	Robust OLS		Conley and Taber		Donald and Lang		Corrected Method	
ρ	Mean (1)	Diff (2)	Mean (3)	Diff (4)	Mean (5)	Diff (6)	Mean (7)	Diff (8)
Panel A: $M \in [50, 200]$								
0.01%	0.108	0.001	0.107	-0.111	0.102	-0.108	0.108	-0.001
1%	0.272	0.065	0.104	-0.052	0.100	-0.052	0.104	0.001
4%	0.493	0.112	0.099	-0.022	0.096	-0.022	0.100	-0.001
Panel B: $M \in [200, 800]$								
0.01%	0.109	0.000	0.104	-0.107	0.099	-0.103	0.106	0.000
1%	0.493	0.126	0.102	-0.021	0.100	-0.022	0.104	0.001
4%	0.716	0.073	0.100	-0.004	0.097	-0.005	0.101	0.003
Panel C: $M \in [50, 950]$								
0.01%	0.111	-0.010	0.102	-0.159	0.088	-0.145	0.101	-0.005
1%	0.474	0.187	0.100	-0.041	0.097	-0.040	0.102	0.003
4%	0.692	0.144	0.102	-0.011	0.099	-0.013	0.103	0.001

Note: This table presents results from Monte Carlo simulations with 400 groups, as explained in Section 3. 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. For each inference method we present the average rejection rate for a test with 10% significance level and the difference in rejection rates when the number of individual observations in the treated group (M_1) is above and when it is below the median. We run 40,000 simulations for each $M \times \rho \times N_0$ scenario. The standard error for the average rejection rates is around 0.16 percentage points, while the standard error for the difference in rejection rates between above and below median M_1 is around 0.3 percentage points.

Table 2: **Rejection Rates in MC Simulations with** $N_0 + 1 = 100$ ($H_0 : \alpha = 0$ at 10% significance level)

	Inference Method							
	Robust OLS		Conley and Taber		Donald and Lang		Corrected Method	
ρ	Mean (1)	Diff (2)	Mean (3)	Diff (4)	Mean (5)	Diff (6)	Mean (7)	Diff (8)
Panel A: $M \in [50, 200]$								
0.01%	0.130	0.009	0.107	-0.102	0.096	-0.100	0.113	0.004
1%	0.316	0.121	0.110	-0.054	0.102	-0.052	0.110	0.004
4%	0.506	0.049	0.113	-0.021	0.102	-0.022	0.115	-0.001
Panel B: $M \in [200, 800]$								
0.01%	0.110	-0.025	0.107	-0.091	0.097	-0.093	0.116	0.007
1%	0.486	0.135	0.111	-0.016	0.101	-0.015	0.112	0.007
4%	0.720	0.126	0.110	-0.004	0.101	0.000	0.113	0.002
Panel C: $M \in [50, 950]$								
0.01%	0.098	0.004	0.105	-0.157	0.088	-0.143	0.106	0.013
1%	0.460	0.212	0.112	-0.041	0.103	-0.042	0.115	0.001
4%	0.700	0.170	0.111	-0.011	0.102	-0.015	0.110	0.004

Note: This table presents results from Monte Carlo simulations with 100 groups, as explained in Section 3. 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. For each inference method we present the average rejection rate for a test with 10% significance level and the difference in rejection rates when the number of individual observations in the treated group (M_1) is above and when it is below the median. We run 40,000 simulations for each $M \times \rho \times N_0$ scenario. The standard error for the average rejection rates is around 0.16 percentage points, while the standard error for the difference in rejection rates between above and below median M_1 is around 0.3 percentage points.

Table 3: **Rejection Rates in MC Simulations with** $N_0 + 1 = 50$ ($H_0 : \alpha = 0$ at 10% significance level)

	Inference Method							
	Robust OLS		Conley and Taber		Donald and Lang		Corrected Method	
ρ	Mean (1)	Diff (2)	Mean (3)	Diff (4)	Mean (5)	Diff (6)	Mean (7)	Diff (8)
Panel A: $M \in [50, 200]$								
0.01%	0.098	0.008	0.119	-0.104	0.097	-0.095	0.126	0.006
1%	0.267	0.064	0.116	-0.053	0.098	-0.051	0.122	0.000
4%	0.490	0.108	0.116	-0.022	0.096	-0.018	0.120	0.002
Panel B: $M \in [200, 800]$								
0.01%	0.114	0.002	0.121	-0.103	0.101	-0.098	0.128	0.002
1%	0.498	0.103	0.122	-0.012	0.101	-0.013	0.128	0.011
4%	0.715	0.067	0.118	-0.006	0.100	-0.010	0.123	0.000
Panel C: $M \in [50, 950]$								
0.01%	0.106	0.005	0.126	-0.166	0.097	-0.147	0.130	0.002
1%	0.467	0.190	0.120	-0.049	0.100	-0.046	0.125	-0.005
4%	0.691	0.148	0.119	-0.008	0.100	-0.007	0.125	0.008

Note: This table presents results from Monte Carlo simulations with 50 groups, as explained in Section 3. 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. For each inference method we present the average rejection rate for a test with 10% significance level and the difference in rejection rates when the number of individual observations in the treated group (M_1) is above and when it is below the median. We run 40,000 simulations for each $M \times \rho \times N_0$ scenario. The standard error for the average rejection rates is around 0.16 percentage points, while the standard error for the difference in rejection rates between above and below median M_1 is around 0.3 percentage points.

Table 4: **Inference with Bootstrap Methods - MC Simulations** ($H_0 : \alpha = 0$ at 10% significance level)

	Total Number of Groups ($N_0 + 1$)							
	25		50		100		400	
ρ	Mean	Diff	Mean	Diff	Mean	Diff	Mean	Diff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Residuals-Bootstrap on $\hat{\alpha}$ w/o heteroskedasticity correction								
0.01%	0.127	-0.158	0.115	-0.168	0.110	-0.166	0.105	-0.169
4.00%	0.125	-0.008	0.115	-0.011	0.111	-0.019	0.104	-0.019
Panel B: Residuals-Bootstrap on $\hat{\alpha}$ w/ heteroskedasticity correction								
0.01%	0.142	-0.003	0.122	0.001	0.115	-0.002	0.104	0.000
4.00%	0.139	0.006	0.122	0.001	0.113	0.000	0.103	-0.001
Panel C: Residuals-Bootstrap on Pivotal Statistics w/ heteroskedasticity correction								
0.01%	0.112	0.008	0.106	0.002	0.106	-0.005	0.102	-0.001
4.00%	0.110	-0.005	0.108	-0.003	0.105	0.001	0.100	0.000

Note: This table presents results from Monte Carlo simulations for different number of groups and for different intra-group correlation parameters (ρ). In all simulations, only one group is treated. In each scenario, we run 100,000 simulations. In Panel A, we test the null hypothesis that $\alpha = 0$ with a 10% significance level with residuals-bootstrap without our heteroskedasticity correction to recover the distribution of the non-pivotal parameter $\hat{\alpha}$. In Panel B, we show results when we run a residuals-bootstrap with our heteroskedasticity correction to recover the distribution of the non-pivotal parameter $\hat{\alpha}$. In Panel C, we use a pivotal test statistics using residuals-bootstrap with our heteroskedasticity correction to recover its distribution. The test statistic is given by $\hat{s} = \hat{\alpha}/\sqrt{\widehat{var(\hat{\alpha})}}$. For each simulation, we bootstrap the residuals $\hat{W}_j$ in the group x time aggregate model, and calculate $s_b = (\hat{\alpha}_b - \hat{\alpha})/\sqrt{\widehat{var(\hat{\alpha}_b)}}$ 500 times to construct the distribution of the test statistic s. For each scenario, we present the average rejection rate for a test with 10% significance level and the difference in rejection rates when the number of individual observations in the treated group (M_1) is above and when it is below the median. The standard error for the average rejection rates is around 0.1 percentage points, while the standard error for the difference in rejection rates between above and below median M_1 is around 0.2 percentage points.

Table 5: **Inference with Synthetic Control - Monte Carlo Simulations**

	Total Number of Groups ($N_0 + 1$)			
	25		50	
ρ	Mean	Diff	Mean	Diff
	(1)	(2)	(3)	(4)
Panel A: $T = 2$, just-identified				
0.01%	0.120	-0.039	0.100	-0.042
4.00%	0.120	-0.003	0.100	-0.004
Panel B: $T = 3$, just-identified				
0.01%	0.120	-0.027	0.100	-0.048
4.00%	0.120	-0.001	0.100	-0.003
Panel B: $T = 3$, over-identified				
0.01%	0.120	0.008	0.100	0.004
4.00%	0.120	0.001	0.100	0.000

Note: This table presents rejection rates from Monte Carlo simulations using the inference proposed by Abadie et al. (2010) for the Synthetic Control Estimation for different number of groups and for different intra-group correlation parameters (ρ). In all simulations, only one group is treated. Panel A reports results for a scenario with 2 periods, one pre- and one post-treatment. We estimate the weights using Y_{j1} and M_j. Panel B reports results for a scenario with 3 periods, two pre- and one post-treatment. We estimate the weights using Y_{j1}, Y_{j2} and M_j. Panel C also reports results for a scenario with 3 periods using only Y_{j1} and M_j to estimate the weights. For each scenario, we present the average rejection rate for a test with 10% significance level and the difference in rejection rates when the number of individual observations in the treated group (M_1) is above and when it is below the median.

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

	Group Level	
	PUMA	State
	(1)	(2)
Average	729.91	10,137.79
1%	127	883
5%	154	1,037
10%	171	1,250
25%	212	2,527
50%	317	7,205
75%	626	11,509
90%	1,337	21,099
95%	2,333	32,961
99%	8,168	62,752

Note: This Table presents the distribution of number of observations per groups (PUMA or state) used in the simulations with the ACS dataset.

Table 7: **Simulations with the ACS Survey ($H_0 : \alpha = 0$ at 10% significance level)**

	Inference Method									
	Robust OLS		Conley and Taber		Donald and Lang		Corrected Method		Corrected Bootstrap	
Outcome Variable	Mean (1)	Diff (2)	Mean (3)	Diff (4)	Mean (5)	Diff (6)	Mean (7)	Diff (8)	Mean (10)	Diff (11)
					Panel A: ACS with PUMA level interventions					
Employment	0.133***	0.008	0.102	-0.149***	0.101	-0.147***	0.102	-0.005	0.101	-0.006
	(0.005)	(0.010)	(0.005)	(0.009)	(0.005)	(0.009)	(0.004)	(0.009)	(0.004)	(0.009)
Log(wages)	0.136***	0.002	0.102	-0.140***	0.102	-0.140***	0.102	0.005	0.105	0.011
	(0.005)	(0.009)	(0.005)	(0.009)	(0.005)	(0.009)	(0.005)	(0.009)	(0.005)	(0.009)
					Panel B: ACS with state level interventions					
Employment	0.130*	0.049	0.118	-0.192***	0.105	-0.168***	0.118	0.024	0.103	-0.016
	(0.016)	(0.032)	(0.024)	(0.038)	(0.023)	(0.038)	(0.019)	(0.036)	(0.017)	(0.032)
Log(wages)	0.137**	-0.024	0.118	-0.211***	0.091	-0.178***	0.125	0.030	0.089	0.015
	(0.019)	(0.039)	(0.028)	(0.046)	(0.026)	(0.044)	(0.021)	(0.043)	(0.018)	(0.037)

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, Conley and Taber (2011) inference method, Donald and Lang (2007) inference method, our correction method, and a residuals-bootstrao method on a pivotal statistic 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. We present in brackets standard errors for the rejection rates clustered at the treated group. For average rejection rates, * means that we reject at 10% that the average rejection rate is equal to 10%, while for the differences in rejection rates * 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%.

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

This supplemental appendix contains the main theorems and proof 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:

$$y_{jt} = \alpha d_{jt} + \theta_j + \gamma_t + \eta_{jt} \tag{14}$$

For now, we deal with the case with only $j = 1$ is treated, and two periods of time. We assume exogeneity of η_{jt} and a variance structure

The first assumption imposes independence of η_{jt} and the main right-hand side variable in the model. The second assumption states the first and second moments of $\eta_{j1-}\eta_{j0}$.

Assumption 1 (Independence and Distribution) : (η_{j1}, η_{j0}) is independent of (d_{j1}, d_{j0}) and is also independent across j. In addition, we assume that the distribution of the (η_{j1}, η_{j0}) only differs among the j by the variance.

Assumption 2[18] **(Exogeneity and Variance-Covariance Structure)**:

$$E[\eta_{j1} - \eta_{j0}] = 0$$

$$Var[\eta_{j1} - \eta_{j0}] = A + B\left(\frac{1}{M(j,1)} + \frac{1}{M(j,0)}\right)$$

where A and B are constants.

As noted in the main paper, in this model the DID estimator would be given by:

$$\widehat{\alpha} = \alpha + (\eta_{11} - \eta_{10}) - \frac{1}{N_0}\sum_{j=1}^{N_0+1}(\eta_{j1} - \eta_{j0})$$

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

$$Var[\widehat{\alpha}] = \left(\frac{N_0}{1+N_0}\right)A + \frac{B}{M(j,1)} + \frac{B}{M(j,0)} + \frac{1}{N_0^2}\sum_{j=1}^{N_0+1}\left[\frac{B}{M(j,1)} + \frac{B}{M(j,0)}\right] \tag{15}$$

As $N_0 \to \infty$,

$$\widehat{\alpha} - \alpha \to \eta_{11} - \eta_{10} \equiv W$$

$$Var[\widehat{\alpha}] \to A + \frac{B}{M(j,1)} + \frac{B}{M(j,0)}$$

We extend the main idea in Conley and Taber (2011) to the heteroskedasticity case, and use the predicted residuals from the control groups, $\widehat{W}_j = \widehat{\eta}_{j1} - \widehat{\eta}_{j0}$ to estimate the distribution of W. Because of the this heteroskedasticity, we would like to use $\widetilde{W}_j = \widehat{W}_j \cdot \sqrt{\frac{Var[W]}{Var[W_j]}}$ so that all $\widetilde{W}_j$ have the same variance as W.

We assume that the number of individuals in each group is fixed and does not vary with N_0. Denote $\Gamma(w_1) = \Pr[W_1 < w | t = 1, .., T]$ and $\widehat{\Gamma}(w_j) = 1\left\{\widetilde{W}_j < w\right\}$, where $\widetilde{W}_j = (\widehat{\eta}_{j1} - \widehat{\eta}_{j0}) \cdot \sqrt{\frac{Var[W]}{Var[W_j]}}$, for $j = 2, ..., N_0 + 1$.

[18] This assumption can be derived from assumptions about η_{jt} or about the unobservable terms in the individual-level model. However, this assumption is general, allowing serial correlation of the η_{jt}.

Theorem 1 shows that $\widehat{\Gamma}(w_j)$ converges uniformly on any compact subset of the support of W. The proof is similar to Conley and Taber (2011) proposition 2.

Theorem 1 *Under assumptions 1 and 2, $\widehat{\Gamma}(w_j)$ converges in probability to $\Gamma(w_1)$ uniformly on any compact subset of the support of W, as $N_0 \to \infty$.*

Proof. *Since under our assumptions, η_{jt} are independent across $j$and in the same family of distributions, we can write*

$$\begin{aligned}\Gamma(w_1) &= \Pr[W_1 < w | \, t = 1, .., T] \\ &= \int 1(W_1 < w) \, dF_1(W_1)\end{aligned}$$

and

$$\begin{aligned}\widehat{\Gamma}(w_1) &= \int 1(W_1 < w) \, d\widehat{F}_1(W_1) \\ &= \int 1(W_1 < w) \, d\widehat{F}_j^*(W_1)\end{aligned}$$

where

$$\widehat{F}_1^*(w_1) = \widehat{F}_1^*\left(w_m \cdot \sqrt{\frac{Var[W_1]}{Var[W_j]}}\right)$$

where $Var[W_1]$ and $Var[W_j]$ are unknown constants and $\widehat{F}_j^*(.)$ is the empirical CDF of the residuals from the control group normalized to have variance equal to the treatment group. In our case, we can take out the means and estimate the following model (as in C&T):

$$\widetilde{Y}_{jt} = \alpha \widetilde{d}_{jt} + \widetilde{\eta}_{jt}$$

The residual for a member of the control group is

$$\widetilde{\eta}_{jt} = \widetilde{Y}_{jt}$$

Note that $\widetilde{\eta}_{jt} = \eta_{it} - \overline{\eta}_j - \overline{\eta}_t + \overline{\eta} \to_p \eta_{jt} - \overline{\eta}_j$ as $N_0 \to \infty$. Using this definition,

$$\begin{aligned}\widehat{F}_j^*(w_1) &= \frac{1}{N_0} \sum_{m=1}^{N_0} 1\left\{\left(\widetilde{Y}_{m1} - \widetilde{Y}_{m0}\right) \sqrt{\frac{Var[W_1]}{Var[W_j]}} < w_m \sqrt{\frac{Var[W_1]}{Var[W_j]}}\right\} \\ &= \frac{1}{N_0} \sum_{i=1}^{N} 1\left\{W_j^* < w_1\right\}\end{aligned}$$

Note that W_j^* are now i.i.d across j.

Define

$$\phi(w_1) = \Pr\left[(\eta_{j1} - \eta_{j0}) \cdot \sqrt{\frac{Var[W_1]}{Var[W_j]}} < w_1\right]$$

As in C&T, we first need to show that $\widehat{F}_j^*(w_j)$ converges uniformly to $\phi(w_j)$ over w_j. Note that

$$\widetilde{Y}_{m1} - \widetilde{Y}_{m0} = \eta_{j1} - \eta_{j0}$$

and $\sqrt{\frac{Var[W_1]}{Var[W_j]}} = c_j$ that is a constant for each j.

We need to show that

$$\sup_{w_j \in \Theta} \left|\widehat{F}_j^*(w_j) - \phi(w_j)\right| \to_p 0$$

where Θ is the support of W_1. This is satisfied by the Glivenko–Cantelli theorem. ■

The approach proposed to estimate $\widetilde{W}_j$ is unfeasible since we do not know the variances of W_j and W_1. Theorem 2 shows that if we have a consistent estimator of $\sqrt{\frac{Var[W_1]}{Var[W_j]}}$, we can construct $\widehat{\widetilde{W}}_j = (\widehat{\eta}_{j1} - \widehat{\eta}_{j0}) \cdot \sqrt{\frac{\widehat{Var[W]}}{\widehat{Var[W_j]}}}$, and use the approach proposed above. Define $\widehat{\widetilde{\Gamma}}(w_j) = 1\left\{\widehat{\widetilde{W}}_j < w\right\}$.

Theorem 2 *If for each j $\sqrt{\frac{\widehat{Var[W]}}{\widehat{Var[W_j]}}}$ is a consistent estimator for $\sqrt{\frac{Var[W_1]}{Var[W_j]}}$, under assumptions 1 and 2, $\widehat{\widetilde{\Gamma}}(w_j)$ converges in probability to $\Gamma(w_1)$ uniformly on any compact subset of the support of W, as $N_0 \to \infty$.*

Proof. Note that

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

By Theorem 2, $\sup_{w_j \in \Theta} \left|\widehat{F}_j^*(w_j) - \phi(w_j)\right| \to_p 0$ We only need to work with the first term,

$$\begin{aligned}\sup_{w_j \in \Theta} \left|\widehat{F}_j^*(\widehat{w}_j) - \widehat{F}_j^*(w_j)\right| &= \sup_{w_j \in \Theta} \left|\frac{1}{N_0}\sum_{m=1}^{N_0} 1\{W_m \cdot \widehat{c}_j < w_m \cdot \widehat{c}_j\} - \frac{1}{N_0}\sum_{i=1}^{N} 1\{W_m \cdot c_j < w_m \cdot c_j\}\right| \\ &= \sup_{w_j \in \Theta} |(1\{W_m \cdot \widehat{c}_j < w_m \cdot \widehat{c}_j\} - 1\{W_m \cdot c_j < w_m \cdot c_j\})| \\ &\leq \sum_{m=1}^{N_0} \sup_{w_j \in \Theta} |(1\{W_m \cdot \widehat{c}_j < w_m \cdot \widehat{c}_j\} - 1\{W_m \cdot c_j < w_m \cdot c_j\})| \\ &\to_p 0 \text{ since } \widehat{c}_j \to_p c_j.\end{aligned}$$

■

We proposed a consistent estimator of $\sqrt{\frac{Var[W_1]}{Var[W_j]}}$ based on an ordinary least squares estimator. We estimate a linear regression that relates squares of $\widehat{W}_j^2$ and $\frac{1}{M(j,1)+M(j,0)}$ and constant. We obtain $\widehat{A}$ as the least squares coefficient associated with the constant, and $\widehat{B}$ as the coefficient associated with $\frac{1}{M(j,1)+M(j,0)}$. We use A and B to construct a consistent estimator for the $Var[W_j]$,

$$\widehat{Var[W_j]} = \widehat{A} + \frac{\widehat{B}}{M(j,1) + M(j,0)}$$

and

$$\widehat{Var[W_1]} = \widehat{A} + \frac{\widehat{B}}{M(1,1) + M(1,0)}$$

We use these two estimator to estimate the ratio $\widehat{c}_j \equiv \sqrt{\frac{\widehat{Var[W_1]}}{\widehat{Var[W_j]}}}$. Theorem 3 shows that $\widehat{c}_j$ is a consistent estimator for $\sqrt{\frac{Var[W_1]}{Var[W_j]}}$.

Theorem 3 *Under assumptions 1 and 2, $\widehat{c}_j$ is a consistent estimator for $\sqrt{\frac{Var[W_1]}{Var[W_j]}}$.*

Proof. *Under assumptions 1 and 2,*

$$Var[W_{jt}] = A + \frac{B}{M(j,t)} \text{ and } \mathbb{E}[W_{jt}] = 0$$

So we can write

$$\mathbb{E}\left[W_{jt}^2\right] = A + \frac{B}{M\left(j,t\right)}$$

or

$$W_{jt}^2 = A + \frac{B}{M\left(j,t\right)} + \omega$$

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

The method proposed above provides consistent results if we have a large number of controls. In the last part of the article, we compare the method proposed above with a bootstrap-based method that works better with a not so large N_0. We propose to work with the following test statistics:

$$s = \frac{\widehat{\alpha}}{\sqrt{Var\left[\widehat{\alpha}\right]}}$$

where $Var\left[\widehat{\alpha}\right]$ is given by equation 15.

If you know A and B, under the null hypothesis, s will converge in distribution to a normal with mean 0 and variance 1. Note that s is a pivotal test statistics. Cameron, Gelbach and Miller (2008) shows that is asymptotically better to bootstrap an asymptotically pivotal statistics.

However, we do not know A and B, and we estimate A and B using the OLS estimators of a regression of W_j^2 on a constant and $\left(\frac{1}{M(j,1)} + \frac{1}{M(j,0)}\right)$ as explained above. When we use $\widehat{Var\left[\widehat{\alpha}\right]}$ in the place of $Var\left[\widehat{\alpha}\right]$, the test statistics does not have known distribution in small sample. In large sample, we can show that the distribution of the test statistics approximately a normal with mean 0 and variance 1.

$$\begin{aligned}\widehat{s} &= \frac{\widehat{\alpha}}{\sqrt{\widehat{Var\left[\widehat{\alpha}\right]}}} \\ &= \frac{\widehat{\alpha}}{\sqrt{Var\left[\widehat{\alpha}\right]}} \cdot \sqrt{\frac{Var\left[\widehat{\alpha}\right]}{\widehat{Var\left[\widehat{\alpha}\right]}}}\end{aligned}$$

Under assumptions 1 and 2, $\sqrt{\frac{Var[\widehat{\alpha}]}{\widehat{Var[\widehat{\alpha}]}}} \to_p 1$ and $\widehat{s} \to_d N\left(0,1\right)$.

Since the distribution of $\widehat{s}$ is unknown in not so large samples, we use bootstrap to approximate the conditional distribution function $\widehat{s}$. By bootstrap, we obtain B resamples of size N of the original sample Z_N. In each bth sample (Z_{Nb}^*), we calculate $\widehat{\alpha}_{Nb}$ and $\widehat{Var\left[\widehat{\alpha}_{Nb}\right]}$, and compute the following statistics,

$$\widehat{s}_{Nb} = \frac{\widehat{\alpha}_{Nb} - \widehat{\alpha}}{\sqrt{\widehat{Var\left[\widehat{\alpha}_{Nb}\right]}}}$$

The empirical distribution of $\widehat{s}_{Nb}$, $b = 1, ..., B$ is used to compute the test critical values and p-values.

Theorem 4 *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* $\widehat{s}_{Nb}$, $b = 1, ..., B$. *Under assumptions 1 and 2,*

$$\Pr\left[d_{1-\frac{\alpha}{2}} \leq \widehat{s} \leq d_{\frac{\alpha}{2}} \middle| \alpha_0\right] \to_p 1 - \alpha$$

Proof. *This proof is divided in two parts. In the first part, we show that given a sample* Z, $\sqrt{N}\widehat{s}_{Nb}$ *converges conditionally in distribution to the same limit as* $\sqrt{N}\widehat{s}$. *Then we show that* $\Pr\left[d_{1-\frac{\alpha}{2}} \leq \widehat{s} \leq d_{\frac{\alpha}{2}} \middle| \alpha_0\right] \to_p 1 - \alpha$.

After estimating the model by Difference in Difference, we generate normalized $\widehat{\widetilde{W}}_j = \widehat{W}_j \cdot \sqrt{\frac{1}{\widehat{Var[W_j]}}}$, *with* $\widehat{Var[W_j]} = \widehat{A} + \widehat{B}\left(\frac{1}{M(j,1)} + \frac{1}{M(j,0)}\right)$.

In each bootstrap replication, we generate a sample with replacement of size N, $\left(\widehat{\widetilde{W}}_1\right), ..., \left(\widehat{\widetilde{W}}_N\right)$ *and generate*

$$\widehat{\widetilde{W}}^*_{j,b} = \widehat{\widetilde{W}}_{j,b} \cdot \sqrt{\widehat{Var[W_{j,b}]}}$$

where $\widehat{Var[W_{j,b}]}$ *is the variance of* W_j *the corresponding group b.*

$$\begin{aligned}\widehat{\alpha}_{Nb} &= \left(\widehat{Y}_{11} - \widehat{Y}_{10}\right) + \widehat{\widetilde{W}}^*_{1,b} - \frac{1}{N_0}\sum_{j=2}^{N_0+1}\left(\widehat{Y}_{j1} - \widehat{Y}_{j0} + \widehat{\widetilde{W}}^*_{j,b}\right) \\ &= \left(\widehat{Y}_{11} - \widehat{Y}_{10}\right) - \frac{1}{N_0}\sum_{j=2}^{N_0+1}\left(\widehat{Y}_{j1} - \widehat{Y}_{j0}\right) + \left(\widehat{\widetilde{W}}^*_{1,b} - \frac{1}{N_0}\sum_{j=2}^{N_0+1}\widehat{\widetilde{W}}^*_{j,b}\right)\end{aligned}$$

$$\widehat{Y}_{jt} = \widehat{\alpha} d_{jt} + \widehat{\theta}_j + \widehat{\gamma}_t$$

Using the formulas of the traditional Difference in Difference,

$$\widehat{\alpha}_{Nb} - \widehat{\alpha} = \widehat{\widetilde{W}}^*_{1,b} - \frac{1}{N_0}\sum_{j=2}^{N_0+1}\widehat{\widetilde{W}}^*_{j,b}$$

and we do a regression of $\widehat{\widetilde{W}}^*_{j,b}$ *on a constant and* $\left(\frac{1}{M(j,1)} + \frac{1}{M(j,0)}\right)$, *and construct*

$$\widehat{Var\left[\widehat{\widetilde{W}}^*_{j,b}\right]} = \widehat{A}_b + \widehat{B}_b\left(\frac{1}{M(j,1)} + \frac{1}{M(j,0)}\right)$$

and

$$\widehat{Var[\widehat{\alpha}_{Nb}]} = \widehat{Var\left[\widehat{\widetilde{W}}^*_{1,b}\right]} + \frac{1}{N_0^2}\sum_{j=2}^{N_0+1}\widehat{Var\left[\widehat{\widetilde{W}}^*_{j,b}\right]}$$

Note that

$$\widehat{\widetilde{W}}^*_{j,b} = \widehat{W}_j \cdot \sqrt{\frac{\widehat{Var[W_b]}}{\widehat{Var[W_j]}}} = \widehat{W}_j \cdot c_{jb}$$

Under assumptions 1 and 2,

$$E\left[\left\|\widehat{\widetilde{W}}^*_b\right\|^2\right] \cdot 1\left\{\left\|\widehat{\widetilde{W}}^*_b\right\| > \varepsilon\sqrt{n}\right\} = \frac{1}{n}\sum_{j=1}^{n} E\left[\left\|\widehat{W}_{j,b}\right\|^2 c_{jb}^2\right] \cdot 1\left\{\left\|\widehat{W}_{j,b}\right\| > c_{jb}\varepsilon\sqrt{n}\right\} \to_p 0$$

$$Var\left[\widehat{\widetilde{W}}^*_b\right] \to_p \Sigma$$

By the Lindeberg-Feller Central Limit Theorem,

$$\widehat{s}_{Nb} = \frac{\left(\widehat{\widetilde{W}}^*_{1,b} - \frac{1}{N_0}\sum_{j=2}^{N_0+1}\widehat{\widetilde{W}}^*_{j,b}\right)}{\sqrt{\widehat{Var\left[\widehat{\widetilde{W}}^*_{1,b}\right]} + \frac{1}{N_0^2}\sum_{j=2}^{N_0+1}\widehat{Var\left[\widehat{\widetilde{W}}^*_{j,b}\right]}}} \to_d N(0,1)$$

and by theorem 23.3 of Vaart (1998),

$$\Pr\left[d_{1-\frac{\alpha}{2}} \leq \widehat{s} \leq d_{\frac{\alpha}{2}} \middle| \alpha_0\right] \rightarrow_p 1-\alpha$$

■