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

Bruno Ferman† Cristine Pinto‡

Sao Paulo School of Economics - FGV

First Draft: October, 2015

This Draft: October, 2017

Please click here for the most recent version

**Abstract**

We show that existing inference methods used in Differences-in-Differences might not perform well with few treated groups and heteroskedastic errors. This is restrictive because variation in the number of observations per group inherently leads to heteroskedasticity in the group x time aggregate model. We provide theoretical justification and empirical evidence from placebo simulations with real datasets showing that this problem may remain relevant even in datasets with a large number of observations per group. We then derive an alternative inference method that works when there are few treated groups (or even just one) and many control groups in the presence of heteroskedasticity. Combined with feasible generalized least squares estimation, our test is uniformly most powerful under normality and a consistent estimator for the variance-covariance matrix, while it can still provide a test with correct size if the serial correlation is misspecified or errors are not normally distributed.

**Keywords:** differences-in-differences; inference; heteroskedasticity; clustering; bootstrap; permutation tests; Behrens-Fisher problem

**JEL Codes:** C12; C21; C33

*We would like to thank Josh Angrist, Aureo de Paula, Marcelo Fernandes, Sergio Firpo, Bernardo Guimaraes, Michael Leung, Lance Lochner, Ricardo Masini, Marcelo Moreira, Marcelo Medeiros, Whitney Newey, Vladimir Ponczek, Andre Portela, Vitor Possebom, Rodrigo Soares, Chris Taber, Gabriel Ulyssea and seminar participants at Boston University, MIT Econometrics lunch, Sao Paulo School of Economics - FGV, PUC-Rio, Insper, Latin American Workshop in Econometrics, EPGE-FGV, USP, 3rd conference of the IAAE, the Bristol Econometric Study Group Annual Conference, the European Meeting of the Econometric Society, and the Africa Meeting of the Econometric Society for comments and suggestions. We also thank Lucas Finamor and Deivis Angeli for excellent research assistance. Cristine Pinto gratefully acknowledges financial support from FAPESP.

†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 errors might exhibit intra-group and serial correlations.[1] Not taking these problems into account can lead to severe underestimation of the DID standard errors, as highlighted in Bertrand et al. (2004). Still, there is as yet no unified approach to dealing 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."*

With many treated and many control groups, one of the most common inference methods used in DID applications is the cluster-robust variance estimator (CRVE) at the group level, which allows for unrestricted intra-group correlation and is also heteroskedasticity robust.[2] With a small number of groups, it might still be possible to obtain tests with correct size, even with unrestricted heteroskedasticity (for example, Cameron et al. (2008), Brewer et al. (2013), Canay et al. (2014), Ibragimov and Müller (2010), Ibragimov and Müller (2016), and MacKinnon and Webb (2015a)). However, all of these inference methods do not perform well when the number of treated groups is very small. In particular, none of these methods perform well when there is only one treated group.[3] There are alternative inference methods that are valid with very few treated groups, such as Donald and Lang (2007), henceforth DL, Conley and Taber (2011), henceforth CT, and cluster residual bootstrap, analyzed by Cameron et al. (2008). However, all these methods rely on some sort of homoskedasticity assumption in the group x time aggregate model, which might be a very restrictive assumption in common DID applications. For example, if there is variation in the number of observations in each group x time cell, then the group x time DID aggregate model should be inherently heteroskedastic.[4] As a consequence, these methods would tend to (under-) over-reject the null hypothesis when the number of observations in the treated groups is (large) small relative to the number of observations in the control groups.[5]

In this paper, we first formalize the idea that variation in group sizes may lead to distortions in inference methods designed to work with very few treated groups, and we show that this problem may remain relevant even when the number of observations per group is large. More specifically, we show that there are plausible structures on the errors such that the group x time aggregate model remains heteroskedastic even when the number of observations per group goes to infinity. In placebo simulations with the American Community Survey (ACS) and the Current Population Survey (CPS), we also provide evidence that this problem can be relevant in datasets commonly used in empirical applications, even when we have a very large numbers of

[1] We refer to "group" as the unit level that is treated. In typical applications it stands for states, counties, or countries.

[2] The CRVE was developed by Liang and Zeger (1986), and we can think of this method as a generalization of the heterocedasticity-robust variance matrix due to White (1980). Bertrand et al. (2004) show that CRVE at the group level works well when the number of groups is large, while Wooldridge (2003) provides an overview of cluster-sample methods in linear models and shows that CRVE provides valid inference when the number of groups increases and groups sizes are fixed.

[3] MacKinnon and Webb (2015b) show that CRVE t-statistics and the wild bootstrap have important size distortions in this case. Canay et al. (2014) inference method for DID would have poor power when the number of treated groups is very small, as they point out in remark S.2.5, while Ibragimov and Müller (2010) and Ibragimov and Müller (2016) require at least two observations in each group. Finally, the method proposed in MacKinnon and Webb (2015a) can lead to important size distortions when there are very few treated groups, as they present in their paper.

[4] Even in case of individual-level data, all these methods require (implicitly or explicitly) some kind of aggregation at the group x time level. Therefore, this problem is relevant whether one considers DID regressions using individual-level or aggregate data.

[5] The problem of variation in group sizes leading to heteroskedasticity and, therefore, to distortions in methods that rely on homoskedasticity, was already acknowledged in CT. In parallel to our paper, MacKinnon and Webb (2015a) also provided evidence on this problem based on Monte Carlo simulations.

observations per group. For example, in placebo simulations with the ACS, rejection rates at 5% significance level (under the null) for tests that rely on homoskedasticity are close to zero when the number of observations in the treated group is above median, and close to 10% when it is below median. Therefore, while DL argue that a large number of observations per group would justify the homoskedasticity assumption and CT provide an extension of their method that would be valid with individual-level data when the number of observations per group grows at the same rate as the number of control groups, we provide a theoretical justification and empirical evidence based on real datasets showing that these results would not be valid under more complex (and more plausible) structures on the errors.[6]

We then derive an alternative method for inference when there are only few treated groups (or even just one) and errors are heteroskedastic. The main assumption is that we can model the heteroskedasticity of a *linear combination* of the errors.[7,8] Under this assumption, we can re-scale this linear combination of the residuals of the control groups using the (estimated) heteroskedasticity structure so that they become informative about the distribution of this linear combination of the errors of the treated groups. Importantly, the main advance of our method is that, by focusing on a linear combination of the errors, we circumvent the need to impose strong assumptions and to specify a structure for the intra-group x time and serial correlations. Moreover, we also circumvent the incidental parameter problem caused by the estimation of group fixed effects with a finite number of time periods. We show that a cluster residual bootstrap with this heteroskedasticity correction provides valid hypothesis testing asymptotically when the number of control groups goes to infinity, even when there is only one treated group. Our Monte Carlo simulations and simulations with real datasets (the ACS and the CPS) suggest that our method provides reliable hypothesis testing when there are around 25 groups in total (1 treated and 24 controls). It is important to note that no heteroskedasticity-robust inference method in DID performs well with one treated group. Therefore, although our method is not robust to any form of unknown heteroskedasticity, it provides an important improvement relative to existing methods.

CT present in their online appendix an example model that would allow for temporal dependence and heteroskedasticity depending on group sizes. However, the method they propose imposes strong assumptions on the structure of the errors. For example, they assume stationarity and a separability of the errors in the group x time aggregate model into two Gaussian processes, one capturing dependence and another one heteroskedasticity. This essentially implies that the serial correlation can only come from a common shock that affects all observations in a group x time in the same way, which should not be a plausible assumption for researchers using, for example, the CPS.[9] In contrast, our method is robust to a much wider variety of assumptions on the structure of the errors, allowing for more complex intra-group and serial correlations without the need to parametrically specifying them, even if the correlation between individuals within a group depends on variables that are unobserved by the econometrician. While it may be possible

[6] In both cases, their methods would work when the number of observations goes to infinity *if* there is a common group x time error, but would fail when there is no common group x time error. The intuition is that, when the number of observations goes to infinity, the average of the individual-level error will be $o_p(1)$, while the average of a common shock that equally affects everyone in the same group x time cell will remain $O_p(1)$. We show that there might be more complex structures on the errors in which the aggregate group x time errors are $o_p(1)$, but ignoring intra-cluster correlations would still underestimate the standard errors. In such cases, the aggregate model remains heteroskedastic even when the number of observations per cell goes to infinity.

[7] The crucial assumption for our method is that, conditional on a set of covariates, the distribution of a linear combination of the errors does not depend on treatment status. We consider a stronger assumption that the conditional distribution of this linear combination of the errors is i.i.d. up to a variance parameter in order to reduce the dimensionality of the problem.

[8] While our method is more general, this assumption would be satisfied in the particular example in which the heteroskedasticity is generated by variation in the number of observations per group.

[9] We show in Appendix A.6 that the method proposed in the online appendix of CT may lead to significant size distortions in placebo simulations with the CPS.

to apply the method suggested in CT in their online appendix under a different set of assumptions, this would require derivation of a different set of moment conditions, which is complicated by the fact that conventional estimators of the time series model's parameters based on the DID residuals would be biased due to the problem of incidental parameters (see e.g. Hansen (2007)). In contrast, by focusing on a linear combination of the errors, our method circumvents the incidental parameter problem and, as a consequence, it is straightforward to implement.[10],[11]

Our inference method can also be combined with feasible generalized least squares (FGLS) estimation. The use of FGLS to improve efficiency of the DID estimator has been proposed by Hausman and Kuersteiner (2008), Hansen (2007), and Brewer et al. (2013). One important challenge for implementing a FGLS estimator in the DID setting is that sample analogs for the variance/covariance matrix parameters will be inconsistent when the number of periods is fixed. While these papers provide bias-corrected estimators for the parameters of the variance/covariance matrix, they rely on strong assumptions on the structure of the errors, including homoskedasticity.[12] Following Wooldridge (2003), Hansen (2007) and Brewer et al. (2013) combine their FGLS estimators with cluster-robust inference. This way, their inference is robust to misspecification in either the serial correlation or the heteroskedasticity structures. However, with few treated groups the use of CRVE would not work. In this case, we show that it is possible to combine FGLS estimation with our inference method. If the FGLS estimator is asymptotically equivalent to the GLS estimator and errors are normally distributed, then we show that our test is asymptotically uniformly most powerful (UMP) when the number of control groups goes to infinity. If, however, we have misspecification of the serial correlation, the estimators of the serial correlation parameters are inconsistent, or we do not have normality, then a t-test based on the FGLS estimator would not be valid, while our test can still provide the correct size. Therefore, our method provides an important safeguard for the use of FGLS estimation in DID applications with few treated groups.

With only one treated group, we show that the assumption that we can model the heteroskedasticity of a linear combination of the errors can only be relaxed if we impose instead restrictions on the intra-group correlation. If we assume that, for each group, errors are strictly stationary and ergodic, then we show that it is possible to apply Andrews' (2003) end-of-sample instability test on a transformation of the DID model for the case with many pre-treatment and a fixed number of post-treatment periods. This approach works even when there is only one treated and one control group. We also consider the use of linear factor models for estimation of regional policies treatment effects, as suggested by Gobillon and Magnac (2013). This approach requires both many control groups and many pre-treatment periods, but it allows selection into treatment to freely depend on unobserved heterogeneity terms. We show that CT and our inference methods can be extended to linear factor models when there are only a few treated groups.

Another estimation method for the case with few treated groups when the number of pre-treatment periods is large is the synthetic control (SC) estimator (Abadie and Gardeazabal (2003) and Abadie et al. (2010)). Abadie et al. (2010) recommend a permutation test for inference with the SC method using as test statistic the ratio of post-/pre-treatment mean squared prediction error (MSPE). If the variance of transitory

[10] Stata do-files to implement our method are available at https://sites.google.com/site/brunoferman/.

[11] In an earlier version of their paper (Conley and Taber (2005)), CT also propose another alternative to this problem, where they consider a deconvolution problem to separately estimate the distributions of the common group x time error and of the individual-level error. This solution, however, would also require strong modeling assumptions on the structure of the errors. In particular, it heavily relies on an error structure that can be decomposed between a common shock that affects every observations equally in a group x time cell and an idiosyncratic shock that is independent across individuals. Moreover, this method relies on sieve estimators and, consequently, requires non-trivial bandwidth choices.

[12] Hausman and Kuersteiner (2008) assume that the variance/covariance matrix is block diagonal with the same block for each group, while Hansen (2007) and Brewer et al. (2013) assume homoskedasticity and constant AR coefficients.

shocks is the same in the pre- and post-treatment periods, then dividing the post-treatment MSPE by the pre-treatment MSPE helps adjust the variance of the test statistic in the presence of heteroskedasticity. However, Ferman and Pinto (2017) show that this permutation test can have important size distortions under heteroskedasticity if the number of pre-treatment periods is finite. In contrast, our main inference method works even when the number of pre-intervention periods is small, and it does not rely on any kind of stationarity assumption on the time series.

Our inference method is also related to the Randomization Inference (RI) approach proposed by Fisher (1935). The RI approach assumes that the assignment mechanism is known. In this case, it would be possible calculate the exact distribution of the test statistic under the null (Lehmann and Romano (2008)). We argue that the RI approach would not provide a satisfactory solution to our problem. First, a permutation test would not provide valid inference if the assignment mechanism is unknown.[13] Moreover, even under random assignment, a permutation test would only remain valid for *unconditional* tests (that is, before we know which groups were treated). However, unconditional tests have been recognized as inappropriate and potentially misleading conditional on a particular data at hand.[14] In our setting, once one knows that the treated groups are (large) small relative to the control groups, then one should know that a permutation test that does not take this information into account would (under-) over-reject the null when the null is true. Therefore, such test would not have the correct size conditional on the data at hand.[15]

Finally, our paper is also related to the Behrens-Fisher problem. They considered the problem of hypothesis testing concerning the difference between the means of two normally distributed populations when the variances of the two populations are not assumed to be equal.[16] In order to take intra-group and serial correlation into account, we consider a linear combination of the errors such that the DID estimator collapses into a simple difference between treated and control groups' averages. Therefore, our method would work in any situation in which the estimator can be rewritten as a comparison of means. For example, this would be the case for experiments with cluster-level treatment assignment. While there are several solutions to this problem with good properties even in very small samples, there is, to the extend of our knowledge, no solution for the case where there is only one observation in one of the groups.[17] Our assumption that, conditional on a set of observable variables, the distribution of the errors does not depend on treatment status guarantees that we can learn about the distribution of the treated groups based on the residuals of the control groups, while still allowing for some heteroskedasticity. We focus on the case of DID estimator because the scenario of very few treated groups and many control groups is more common in this case.

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

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

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

[15] This is essentially the same issue that we document for CT method. In fact, CT propose an alternative way to implement their method which is *heuristically* motivated by the literature on permutation tests and randomization inference.

[16] See Behrens (1929), Fisher (1939), Scheffe (1970), Wang (1971), and Lehmann and Romano (2008). Imbens and Kolesar (2016) show that some methods used for robust and cluster robust inference in linear regressions, such as Bell and McCaffrey (2002), can be considered as natural extensions of inference procedures designed to the Behrens-Fisher problem.

[17] For example, Ibragimov and Müller (2016) provide valid tests at conventional significance levels as long as there are at least two observations in each group.

an alternative inference method that corrects for heteroskedasticity even when there is only one treated group. In Section 3 we extend our inference method to FGLS estimation. In Section 4 we consider an alternative application of our method that relies on a different set of assumptions when the number of pre-treatment periods is large. In Section 5, we extend our inference method to linear factor models with few treated groups. We perform Monte Carlo simulations to examine the performance of existing inference methods and to compare that to the performance of our method with heteroskedasticity correction in Section 6, while we compare the different inference methods by simulating placebo laws in real datasets in Section 7. We conclude in Section 8.

# 2 Base Model

## 2.1 A Review of Existing Methods

Consider first 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 $\alpha$ is the main parameter of interest; $\theta_j$ is a time-invariant fixed effect for group $j$, while $\gamma_t$ is a time fixed-effect; $\eta_{jt}$ is a group x time error term that might be correlated over time, but uncorrelated across groups. Depending on the application, "groups" might stand for states, counties, countries, and so on.

We start considering a group x time DID aggregate model because it is well known that this way we take into account any possible individual-level within group x time cell correlation in the errors (DL and Moulton (1986)). Therefore, we can focus on the inference problems that are still unsettled in the literature, which is how to deal with serial correlation and heteroskedasticity when there are few treated groups. However, both the diagnosis of the inference problem with existing methods and the solutions we propose are valid whether we have aggregate or individual-level data.[18]

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

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

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

The variance of the DID estimator, under the assumption that $\eta_{jt}$ are independent across $j$, is given by:[19]

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

[18] Note that, without individual-level covariates, the DID estimator with individual-level data will be numerically the same as the estimator with aggregated data if we use the number of observations per group x time cell as sampling weights.

[19] This is a finite sample formula. So far, we assume that $d_{jt}$ is non-stochastic and that $var(W_j)$ can vary with $j$.

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

When there are many treated and control groups, Bertrand et al. (2004) suggest that CRVE at the group level works well, as this method allows for unrestricted intra-group and serial correlation in the residuals $\eta_{jt}$. One important point is that this method is not only cluster-robust, but also heteroskedasticity-robust. The CRVE has a very intuitive formula in the DID framework:[20]

$$\widehat{var(\hat{\alpha})}_{\text{Cluster}} = \left[\frac{1}{N_1}\right]^2 \sum_{j=1}^{N_1} \widehat{W}_j^2 + \left[\frac{1}{N_0}\right]^2 \sum_{j=N_1+1}^{N} \widehat{W}_j^2 \tag{3}$$

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 CRVE, we calculate each component of the variance of the DID estimator separately. In other words, we use the residuals of the treated groups to calculate the component related to the treated groups, and the residuals of the control groups to calculate the component related to the control groups. This way, CRVE allows for unrestricted heteroskedasticity. When both the number of treated and control groups goes to infinity, the DID estimator is asymptotically normal, and we can consistently estimate its asymptotic variance using CRVE. However, equation 3 makes it clear why CRVE becomes unappealing when there are few treated groups. In the extreme case when $N_1 = 1$, from the OLS normal equations we will have $\widehat{W}_1 = 0$ *by construction*. Therefore, the variance of the DID estimator would be severely underestimated (as noticed in MacKinnon and Webb (2015b)). The same problem applies to other clustered standard errors corrections such as BRL (Bell and McCaffrey (2002)). It is also problematic to implement heteroskedasticity-robust cluster bootstrap methods such as pairs-bootstrap and wild cluster bootstrap when there are few treated groups. In pairs-bootstrap, there is a high probability that the bootstrap sample will not include a treated unit. Wild cluster bootstrap generates variation in the residuals of each $j$ by randomizing whether its residual will be $\hat{\eta}_{jt}$ or $-\hat{\eta}_{jt}$. However, in the extreme case with only one treated, the wild cluster bootstrap would not generate variation in the treated group, since $\widehat{W}_1 = 0$. Another alternative presented by Bertrand et al. (2004) is to collapse the pre- and post-information. This approach would take care of the auto-correlation problem. However, in order to allow for heteroskedasticity, one would have to use heteroskedasticity-robust standard errors. In this case, this method would also fail when there are few treated groups.

It is clear, then, that the inference problem in DID models with few treated groups revolves around how to provide information on the errors related to the treated groups using the residuals $\hat{\eta}_{jt}$ of the treated groups. Alternative methods use information on the residuals of the control groups in order to provide information on the errors of the treated groups. These methods, however, rely on restrictive assumptions regarding the error terms. DL assume that the group x time errors are normal, homoskedastic, and serially uncorrelated. Under these assumptions, the test statistic based on the group x time aggregate model will have a student-t distribution. The assumption that errors are serially uncorrelated, however, might be unappealing in DID applications (Bertrand et al. (2004)).

CT provide an interesting alternative inference method that allows for unrestricted auto-correlation in the error terms and also relaxes the normality assumption. Their method uses the residuals of the control groups to estimate the distribution of the DID estimator under the null. One of the key differences relative to DL is that CT look at a linear combination of the residuals that takes into account any form of serial

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

correlation instead of using the group x time level residuals. In the simpler case with only one treated group, $\hat{\alpha} - \alpha$ would converge to $W_1$ when $N_0 \to \infty$. In this case, they use $\{\widehat{W}_j\}_{j=2}^{N_0+1}$ (a linear combination of the control group residuals) to estimate the distribution of $W_1$. While CT relax the assumptions of no auto-correlation and normality, it requires that errors are i.i.d. across groups, so that $\{\widehat{W}_j\}_{j=2}^{N_0+1}$ approximates the distribution of $W_1$ when $N_0 \to \infty$. Finally, cluster residual bootstrap methods resample the residuals while holding the regressors constant throughout the pseudo-samples. The residuals are resampled at the group level, so that the correlation structure is preserved. It is possible that a treated group receives the residuals of a control group. Therefore, a crucial assumption is again that errors are homoskedastic.

A potential problem with these methods, as originally explained in CT, is that variation in the number of observations per group might generate heteroskedasticity in the group x time aggregate model. DL argues that a large number of observations per cell would justify the homoskedasticity assumptions, while CT consider an extension of their method to individual-level data, and they show that their method remains valid if the number of observations per group grows at the same rate as the number of number of controls. However, we show in Section 2.2 that these methods may not be valid even when the number of observations per group is large under plausible assumptions on the structure of the errors. In their online appendix and in an earlier version of their paper (Conley and Taber (2005)), CT also suggest alternative strategies for the case with fixed sample sizes that vary across group x time cells. We show in Section 2.3 that the alternative method we propose relies on weaker assumptions on the structure of the errors and is more straightforward to implement.

## 2.2 Leading Example: Variation in Group Sizes

In this section, we formalize the idea that the 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. Moreover, we show that the aggregate group x time model may remain heteroskedastic even when the number of observations per cell is large. It is important to point out, however, that this is not the only case that might generate heteroskedasticity in the group x time aggregate DID model, and that our inference method derived in Section 2.3 is more general and can be applied in other settings.

We start with a simple individual-level DID model:

$$Y_{ijt} = \alpha d_{jt} + \theta_j + \gamma_t + \nu_{jt} + \epsilon_{ijt} \tag{4}$$

where $Y_{ijt}$ represents the outcome of individual $i$ in group $j$ at time $t$; $\nu_{jt}$ is a group x time error term (possibly correlated over time), and $\epsilon_{ijt}$ is an individual-level error term. The main features that define a "group" in this setting are that the treatment occurs at the group level and that errors $(\nu_{jt} + \epsilon_{ijt})$ of two individuals in the same group might be correlated, while errors of individuals in different groups are uncorrelated. For ease of exposition, we start assuming that $\epsilon_{ijt}$ are all uncorrelated, while allowing for unrestricted auto-correlation in $\nu_{jt}$, and then we consider more complex structures. Importantly, our correction will require much weaker assumptions on the error structure, as will be presented in Section 2.3.

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

$$Y_{jt} = \alpha d_{jt} + \theta_j + \gamma_t + \eta_{jt} \tag{5}$$

If we let $M(j,t)$ be the number of observations in group $j$ at time $t$, then:

$$\eta_{jt} = \nu_{jt} + \frac{1}{M(j,t)} \sum_{i=1}^{M(j,t)} \epsilon_{ijt} \tag{6}$$

where the errors in the group x time aggregate model ($\eta_{jt}$) are heteroskedastic across $j$, unless $M(j,t)$ is constant across $j$.[21]

Under the assumption that we have a panel of repeated cross-sections, so that $\epsilon_{ijt}$ are not correlated over time, and assuming for simplicity that $M(j,t) = M_j$ is constant across $t$, we have that the variance of $W_j$ conditional on $M_j$ is given by:

$$var(W_j|M_j) \quad = \quad var\left(\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\eta_{jt}|M_j\right) = A + \frac{B}{M_j}$$

for constants $A$ and $B$, regardless of the auto-correlation of $\nu_{jt}$.[22]

Importantly, for a much wider range of structures on the errors, the conditional variance of $W_j$ given $M_j$ will still have a parametric formula given in equation 7 that depends on only two parameters. For example, 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 $\epsilon_{ijt}$ auto-correlation parameters divided by the number of observations, so we would still end up with the same formula, $var(W_j|M_j) = A + \frac{B}{M_j}$. This formula may also remain valid even in situations where the correlation between two observations in the same subgroup (for example, the same municipality or the same school) is stronger than the correlation between two observations in the same group but in different subgroups (for example, observations in the same state but in different municipalities). More specifically, we can consider a model:

$$Y_{ikjt} = \alpha d_{jt} + \theta_j + \gamma_t + \nu_{jt} + \omega_{kjt} + \epsilon_{ikjt} \tag{7}$$

for individual $i$ in subgroup $k$, group $j$ and time $t$, where we allow for a common subgroup shock $\omega_{kjt}$ in addition to the group-level shock $\nu_{jt}$. If the number of subgroups for each group $j$ grows at the same rate as the total number of observations, then this model would also generate $var(W_j|M_j) = A + \frac{B}{M_j}$. Notice also that we do not need to assume that the individual-level model is homoskedastic to have the formula $var(W_j|M_j) = A + \frac{B}{M_j}$.

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

[21] Note that, if $M(j,t) = M_j$ is constant across $t$, then the information on group sizes is already incorporated in model 1 through the group fixed effects $\theta_j$, even though $M_j$ does not enter directly in model 1.

[22] In this simpler case in which $\epsilon_{ijt}$ is i.i.d., then $A = var\left(\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\nu_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\nu_{jt}\right)$ and $B = \left(\frac{1}{T-t^*} + \frac{1}{t^*}\right)var(\epsilon_{ijt})$. When the number of observations per group is not constant over time, the formula will be: $var(W_j) = \tilde{A} + \tilde{B}\left[\left(\frac{1}{T-t^*}\right)^2\sum_{t=t^*+1}^{T}\frac{1}{M(j,t)} + \left(\frac{1}{t^*}\right)^2\sum_{t=1}^{t^*}\frac{1}{M(j,t)}\right]$, for constants $\tilde{A}$ and $\tilde{B}$.

If $A > 0$, note that this would not be a problem when $M_j \to \infty$. In this case, $var(W_j|M_j) \to A$ for all $j$ when $M_j \to \infty$. In other words, when the number of observations in each group x cell is large, then a common shock that affects all observations in a group x time cell would dominate. In this case, if we assume that the group x time error $\nu_{jt}$ is i.i.d., then $\frac{var(W_j|M_j)}{var(W_{j'}|M_{j'})} \to 1$ when $M_j, M_{j'} \to \infty$, which implies that the residuals of the control groups would be a good approximation for the distribution of the treated groups' errors even when the number of observations in each group is different. This is one of the main rationales used in DL to justify the homoskedasticity assumption in the aggregate model, and this is the main reason why the extension of CT to individual-level data when the number of observations per cell is large is valid (proposition 4 in CT).

However, an interesting case occurs when $A = 0$. In this case, even though $var(W_j|M_j) \to 0$ for all $j$ when $M_j \to \infty$, the ratios $\frac{var(W_j|M_j)}{var(W_{j'}|M_{j'})}$ remain constant even if all $M_j$ grows at the same rate, which implies that the aggregate model remains heteroskedastic even asymptotically. Therefore, CT, DL, and cluster residual bootstrap would still 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. Note that we might have $A \approx 0$ under complex (and plausible) conditions on the structure of the errors in which standard inference using OLS regression on the individual-level data would be unreliable. For example, we would have $A \approx 0$ in model 4 if $\epsilon_{ijt}$ is serially correlated and $var(\nu_{jt}) \approx 0$. This may be the case if we have a panel of individual observations, as in the CPS. Alternatively, in model 7 we might have that most of the intra-group correlation comes from individuals in the same subgroup (that is, $var(\omega_{kjt}) > 0$ while $var(\nu_{jt}) \approx 0$), which would also imply that $A \approx 0$. In both cases, $var(W_j|M_j) \to 0$ when $M_j \to \infty$, but the aggregate model remains heteroskedastic even when $M_j$ is large.[23] In Section 7, we present results from placebo simulations with real datasets and provide evidence that this problem is relevant in large datasets commonly used in empirical applications. Taken together, these results suggest that one should be careful when applying methods such as those proposed in CT and DL even when there is a large number of observations in all group x time cells.

## 2.3 Inference with Heteroskedasticity Correction

We derive an inference method that uses information from the control groups to estimate the variance of the treated groups while still allowing for heteroskedasticity. Intuitively, our approach assumes that we know how the heteroskedasticity is generated, which is the case when, for example, heteroskedasticity is generated by variation in the number of observations per group. Under this assumption, we can re-scale the residuals of the control groups using the (estimated) structure of the heteroskedasticity in a way that allows us to use this information to estimate the distribution of the error for the treated groups. Importantly, our method only requires information on the heteroskedasticity structure for a linear combination of the errors, which implies that we do not have to impose strong assumptions on the structure of the serial correlation of the errors. While we motivate our method based on heteroskedasticity generated by variation in the number observations in each group, it is important to note that our method is more general, and we can consider any observable variable that may generate heteroskedasticity in the model, such as the standard textbook case in which the conditional variance is an exponential function of a subset of covariates.

More formally, we assume we have a total of $N$ groups where the first $j = 1, ..., N_1$ groups are treated.

[23] CRVE at the individual level (in the first example) and at the subgroup level (in the second example) should work well under these assumptions. However, the information on more complex intra-group correlations might not be available to the econometrician (for example, he/she might not be information on the relevant subgroups) and/or the econometrician might not want to impose assumptions on the structure of the errors.

For simplicity, we consider first the case where $d_{jt}$ changes to 1 for all treated groups starting after known date $t^*$. Let $X_j$ be a vector of observable covariates that do not necessarily enter in model 1 and $d_j$ be an indicator variable equal to 1 if group $j$ is treated.[24] We will define our assumptions directly on the linear combination of the errors $W_j = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\eta_{jt}$. The main assumptions for our method are:

1. $\{W_j, X_j\}$ is i.i.d. across $j \in \{1, ..., N_1\}$, i.i.d. across $j \in \{N_1+1, ..., N\}$ and independently distributed across $j \in \{1, ..., N\}$.
2. $W_j|X_j, d_j \overset{d}{=} W_j|\tilde{X}_j$, where $\tilde{X}_j$ is a subset of $X_j$.
3. $W_j|\tilde{X}_j$ has the same distribution across $\tilde{X}_j$ up to a scale parameter.
4. $E[W_j|X_j, d_j] = E[W_j|X_j] = 0$.
5. The conditional distribution of $W_j$ given $X_j$ is continuous.

Note that assumption 1 allows the distribution of $\{W_j, X_j\}$ for the treated groups to be different from the distribution for the control groups. Therefore, we might consider a case where treated states have different characteristics $X_j$ (including population sizes) than states in the control group.[25] Assumption 2 implies that, conditional on a subset of observable covariates, the distribution of $W_j$ will be the same independently of the treatment status. This is crucial for our method, as it guarantees that we can extrapolate information from the control groups' residuals to estimate the distribution of the treated groups' errors. This assumption would not be required with large $N_1$ and $N_0$ for inference with heteroskedasticity-robust methods. In this case, the DID would be asymptotically normal and it would be possible to allow for different distributions conditional on treatment status since there would be enough observations to estimate the variance component related to the treated groups using only information from the treated groups. In our setting, this would not be feasible since we assume that the number of treated groups is fixed and small. Assumption 3 implies that the distribution of $W_j|X_j$ only depends on $X_j$ through the variance parameter.[26] This assumption reduces the dimensionality of the problem. It might be possible to relax this assumption and estimate the conditional distribution of $W_j|\tilde{X}_j$ non-parametrically. However, this would require very large number of control groups. Without assumption 3, we can still guarantee that we can recover a distribution with the correct expected value and variance for the DID estimator. This should provide significant improvement relative to existing inference methods.[27] Finally, condition 4 is the standard identification assumption for DID.

Our method is an extension of the cluster residual bootstrap with $H_0$ imposed where we correct the residuals for heteroskedasticity. In cluster residual bootstrap with $H_0$ imposed, we estimate the DID regression imposing that $\alpha = 0$, generating the residuals $\{\widehat{W}_j^R\}_{i=1}^N$. If the errors are homoskedastic, then, under the null, $\widehat{W}_j^R$ converges in distribution to $W_j$ when $N_0 \to \infty$, which would have the same distribution across $j$. Therefore, we could resample with replacement $\mathcal{B}$ times from $\{\widehat{W}_j^R\}_{i=1}^N$, generating $\{\widehat{W}_{j,b}^R\}_{i=1}^N$,

[24] Note that we allow for covariates that vary with time, as we may consider the observations for each time period $t$ as one component in vector $X_j$.

[25] Note that assuming $\{W_j, X_j, d_j\}$ is i.i.d. would also allow for the distribution of $\{W_j, X_j\}$ conditional on $d_j = 1$ to be different from the distribution of $\{W_j, X_j\}$ conditional on $d_j = 0$. However, we do not state assumption 1 this way because we want to consider the asymptotic when $N_1$ is fixed and $N_0 \to \infty$.

[26] As noticed in Ibragimov and Müller (2016) and Canay et al. (2014), if both the number of pre- and post-treatment periods are large and we can apply a central limit theorem to $\frac{1}{t^*}\sum_{t=1}^{t^*}\eta_{jt}$ and $\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\eta_{jt}$, then $W_j$ will be approximately normal. In this case, assumption 3 would be guaranteed.

[27] For example, in our setting, CT method would recover a distribution with different variance relative to the distribution of the DID estimator. In Section 6.1, we provide evidence from Monte Carlo simulations that our inference method works well even in data generating processes that do not satisfy assumption 3, while in Section 7 we show that our inference method works well in simulations with real datasets, in which we do not have control over the data generating process.

and then calculate our bootstrap estimates as $\hat{\alpha}_b = \frac{1}{N_1}\sum_{j=1}^{N_1}\widehat{W}^R_{j,b} - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\widehat{W}^R_{j,b}$. Importantly, note that, in our setting, we do not need to work with the group x time residuals $\hat{\eta}_{jt}$ to construct our bootstrap estimates. Instead, we can work with a linear combination of the residuals that takes into account any form of auto-correlation in the residuals.

As explained in Section 2.1, the problem with cluster residual bootstrap is that it requires the residuals to be homoskedastic. In Theorem 2 in Appendix A.1, we show that, if we know the variance of $W_j$ conditional on $X_j$, then we can re-scale the residual $\widehat{W}^R_{j,b}$ so that it has (asymptotically) the same distribution as $W_j$. First, we normalize each observed $\widehat{W}^R_{j'}$ by $\widehat{W}^{norm}_{j'} = \widehat{W}^R_{j'}\frac{1}{\sqrt{var(W_{j'}|X_{j'})}}$. Then we generate a bootstrap sample with the re-scaled residuals $\widetilde{W}_{j,b} = \widehat{W}^{norm}_{j,b}\sqrt{var(W_j|X_j)}$. As a result, this procedure generates bootstrap estimators $\hat{\alpha}_b = \frac{1}{N_1}\sum_{j=1}^{N_1}\widetilde{W}_{j,b} - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\widetilde{W}_{j,b}$ that can be used to draw inferences about $\alpha$ with the correct size.[28] The main assumption we need is that $\{W_j\}_{j=1}^N$, which is a linear combination of the error terms $\eta_{jt}$, are independent across $j$ and have the same distribution up to the variance parameter. It is important to note that we only need to know the variance of a linear combination of the errors. This point is crucial for our method, because we do not need to specify the serial correlation structure of the errors $\eta_{jt}$. The main problem, however, is that $var(W_j|X_j)$ is generally unknown, so it needs to be estimated. In Theorem 3 in Appendix A.1, we show that this heteroskedasticity correction works asymptotically when $N_0 \to \infty$ if we have a consistent estimator for $var(W_j|X_j)$. That is, we can use $\widehat{var(W_j|X_j)}$ to generate $\widetilde{\widehat{W}}_{j,b} = \widehat{W}^R_{j,b}\sqrt{\frac{\widehat{var(W_j|X_j)}}{\widehat{var(W_{j,b}|X_{j,b})}}}$. Since we only need a consistent estimator for $var(W_j|X_j)$, in theory, one could estimate the conditional variance function non-parametrically. In practice, however, a non-parametric estimator would likely require a large number of control groups.

In our leading example where heteroskedasticity is generated by variation in group sizes, we show in Section 2.2 that we can derive a parsimonious function for the conditional variance without having to impose a strong structure on the error terms. More specifically, in this example, the conditional variance function would be given by $var(W_j|X_j, d_j) = var(W_j|M_j) = A + \frac{B}{M_j}$, for constants $A$ and $B$, where $X_j$ is the set of observable variables including $M_j$. We show in Lemma 4 in Appendix A.1 that we can get a consistent estimator for $var(W_j|M_j)$ by regressing $(\widehat{W}^R_j)^2$ on $\frac{1}{M_j}$ and a constant.[29] Note that we do not need individual-level data to apply this method, provided that we have information on the number of observations that were used to calculate the group x time averages. While we present our method for the group x time aggregate model, we show below that it is straightforward to extend our method to the case with individual-level data.

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

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

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

similar procedure for permutation tests.

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

1. Calculate the DID estimate:

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

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

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

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

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

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

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

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

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

We also show in Appendix A.3 that our method applies to DID models with both individual- and group-level covariates. With covariates at the group x time level, we estimate the OLS DID regressions in steps 1 and 2 of the bootstrap procedure with covariates. The other steps remain the same. If we have individual-level data, then we run the individual-level OLS regression with covariates in step 2 and then aggregate the residuals of this regression at the group x time level. The other steps in the bootstrap procedure remain the same. Finally, we extend our method to the case of individual-level data with sampling weights in Appendix A.4.

Considering our leading example, one of the main advantages of our method is that we do not require strong modeling assumptions on the structure of the errors, as there is a wide variety of assumptions on the errors that generate a conditional variance $var(W_j|M_j) = A + \frac{B}{M_j}$ (see Section 2.2). CT present in their online

appendix and in an earlier version of their paper (Conley and Taber (2005)) alternative methods for the case in which the number of observations per cell is finite and varies across $j$. However, these methods rely on stronger modeling assumptions on the structure of the errors. The method presented in the online appendix of CT assumes stationarity and a separability of the errors in the group x time aggregate model into two Gaussian processes, one capturing dependence and another one heteroskedasticity. This essentially excludes the possibility of serial correlation in the individual-level error, which should be relevant in panel datasets. By imposing a constant serial correlation parameter across groups, this method would underestimate the serial correlation of the error for smaller groups and overestimate the serial correlation of the error for larger groups. This would lead to over-rejection when the treated group is small and under-rejection when the treated group is large. In Appendix Section A.6, we show that this distortion is significant when we consider placebo simulations with the CPS. While it may be possible to apply the method suggested in CT in their online appendix under a different set of assumptions, this would require derivation of a different set of moment conditions, which may prevent applied researchers from using their method. Conley and Taber (2005) consider a deconvolution problem to separately estimate the distributions of $(\nu_{j1}, ..., \nu_{jT})$ and $(\epsilon_{ij1}, ..., \epsilon_{ijT})$. Importantly, they require an additive structure of the error in a common shock that affects all observations in a group x time and an individual-level shock, which should not be plausible in real applications. In contrast to these two alternative methods, our method is valid under a wider range of assumptions on the structure of the errors. In particular, we do not need to assume stationarity, we can allow for more complex within group correlations even if the researcher does not have information on the variables that determine whether two observations have correlated errors, and we can allow for serial correlation in the individual-level error.[30]

# 3 Improving Efficiency with a FGLS

One important feature of our inference method is that we do not need to specify the structure of the serial correlation. Moreover, since the linear combination $W_j$ does not depend on $\theta_j$, we circumvent the incidental parameter problem caused by the estimation of group fixed effects that complicates estimation of serial correlation parameters. We consider now the use of FGLS-DID estimator to improve efficiency. This strategy, however, presents some challenges. First, one needs to impose some structure on the entire variance/covariance matrix. Also, the residual $\hat{\eta}_{jt}$ depends on the group fixed effects estimator, which will not be consistent if $T$ is fixed. This complicates the estimation of the variance/covariance matrix even if parametric assumptions on the variance/covariance matrix are correct, as argued in Hansen (2007). Finally, with few treated groups, the FGLS estimator might not be normally distributed even when $N_0 \to \infty$. We show now that it is possible to combine FGLS estimation with our inference method. This will allow for robust inference in case the serial correlation is misspecified, estimators for the serial correlations parameters are biased, or errors are not normally distributed.

Since we assume that errors are uncorrelated across $j$, the variance/covariance matrix of $\eta_{jt}$ is block diagonal with $T \times T$ blocks given by $\Omega_j$. We assume that $\Omega_j = \Omega(\tilde{X}_j)$. Let $\widehat{\Omega}(\tilde{X}_j)$ be an estimator of $\Omega(\tilde{X}_j)$ that converges to $\bar{\Omega}(\tilde{X}_j)$ (we allow $\bar{\Omega}(\tilde{X}_j) \neq \Omega(\tilde{X}_j)$, so $\widehat{\Omega}(\tilde{X}_j)$ is inconsistent). The FGLS estimator using

[30] While it may be possible to apply the method suggested in CT in their online appendix under a different set of assumptions, this would require derivation of a different set of moment conditions, which is complicated by the fact that conventional estimators of the time series model's parameters based on the DID residuals would be biased due to the problem of incidental parameters (see e.g. Hansen (2007)). In contrast, by focusing on a linear combination of the errors, our method circumvents the incidental parameter problem and, as a consequence, it is straightforward to implement.

$\widehat{\Omega}(\tilde{X}_j)$ will be a linear estimator $\hat{\alpha}_{\text{FGLS}} = \sum_{t=1}^{T}\sum_{j=1}^{N}\hat{a}_{jt}Y_{jt}$. In Appendix A.5, we show that, in the case with only one treated group, $\hat{\alpha}_{\text{FGLS}} \xrightarrow{d} \sum_{t=1}^{T}\bar{a}_{1t}\eta_{1t}$ when $N_0 \to \infty$, where $\bar{\mathbf{a}}_1 = (\bar{a}_{11}, ..., \bar{a}_{1T})'$ is defined by:

$$\begin{aligned} \bar{\mathbf{a}}_1 \quad = \quad & \underset{\mathbf{a}_1}{\text{argmin}}\ \mathbf{a}_1'\bar{\Omega}(\tilde{X}_1)\mathbf{a}_1 \\ & \text{subject to: } \sum_{t=t^*+1}^{T} a_{1t} = 1 \text{ and } \sum_{t=1}^{T} a_{1t} = 0 \end{aligned} \tag{8}$$

Therefore, defining the linear combination $W_j^* = \sum_{t=1}^{T}\bar{a}_{1t}\eta_{jt}$, we show in Appendix A.5 that all results from Section 2.3 apply to the FGLS estimator. The only difference is that the assumptions should be based on the linear combination $W_j^*$ instead of on the linear combination $W_j$. Note that $\widehat{W}_j^{*R} \xrightarrow{d} W_j^*$ when $N_0 \to \infty$, so there would not be an incidental parameter problem by looking at the linear combination $W_j^*$.[31] For our leading example presented in Section 2.2, we would still have $var(W_j^*|M_j) = A + \frac{B}{M_j}$ for constants $A$ and $B$.

So far, we only assumed that $\widehat{\Omega}(\tilde{X}_j)$ converges to $\bar{\Omega}(\tilde{X}_j)$ when $N_0 \to \infty$. So our inference method is valid even if $\bar{\Omega}(\tilde{X}_j) \neq \Omega(\tilde{X}_j)$. If $\eta_{jt}$ is multivariate normal and $\bar{\Omega}(\tilde{X}_j) = \Omega(\tilde{X}_j)$, then we show in Appendix A.5 that our test has asymptotically the same power as a t-test based on the infeasible GLS estimator, which is the uniformly most powerful (UMP) test in this case.[32] By combining our method with FGLS estimation we provide a test that is asymptotically UMP if all these assumptions are satisfied. Importantly, if the serial correlation is misspecified, the estimators of the serial correlation parameters are inconsistent, or the error is not normally distributed, then our test would still have the correct size while a t-test based on the FGLS estimator would be biased. Therefore, our inference method provides an important safeguard for FGLS estimation in DID settings where there are few treated groups.[33] More specifically, we provide an alternative to cluster-robust inference in FGLS when it is not possible to estimate the CRVE.

# 4 Heteroskedasticity Correction with Large $t^*$

One of the main features of our inference method presented in Section 2.3 is that we collapse the time series structure when we consider the linear combination of the errors $W_j$, so that the inference problem becomes equivalent to a comparison of means between treated and control groups. This is why our inference method does not require any specification of the time series structure. However, in the case with only one treated group, this implies that we would have, in practice, only one observation for the treated group to estimate the distribution of the treated group error. This is why a crucial assumption of our method is that $W_j|X_j, d_j \overset{d}{=} W_j|\tilde{X}_j$. Under this assumption, the residuals of the control groups are informative about the distribution of the treated group errors. We can only relax this assumption if we impose some structure on the intra-group correlation.

We now show that, under strict stationarity and ergodicity of the time series, we can apply Andrews' (2003) end-of-sample instability test to a transformation of the DID model if we have a large number of pre-treatment periods and a small number of post-treatment periods. The main idea is that with large $t^*$ and

[31] Another difference relative to the OLS DID is that we have to estimate $\bar{a}_{jt}$. However, since we have a consistent estimator for $\bar{a}_{jt}$, this does not impose any problem to apply our method.

[32] The main intuition of the proof is that, even under the alternative hypothesis, we have that $\widehat{W}_j^{*R} \xrightarrow{d} W_j^*$ for all $j$ in the control group when $N_0 \to \infty$ but $N_1$ is small and fixed. Since the probability of resampling a treated group goes to zero, then the bootstrap distribution will approximate the distribution of $\hat{\alpha}_{FGLS}$ under the null even when the null is false.

[33] If we impose assumptions on the structure of the errors, then we could derive moment conditions based on variance/covariance matrix of the residuals and estimate the parameters of $\Omega(\tilde{X}_j)$ by GMM. However, such assumptions may be restrictive in some cases.

small $T-t^*$ the DID estimator would converge in distribution to a linear combination of the post-treatment errors. Therefore, under strict stationarity and ergodicity, we can use blocks of the pre-treatment periods to estimate the distribution of $\hat{\alpha}$. This is essentially the idea of the method suggested in CT, but exploiting the time instead of the cross-section variation.

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

$$\tilde{Y}_t = \begin{cases} \tilde{\theta} + \tilde{\eta}_t, \text{ for } t = 1, ..., t^* \\ \alpha + \tilde{\theta} + \tilde{\eta}_t, \text{ for } t = t^*+1, ..., T \end{cases} \tag{9}$$

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

Therefore, this is a particular case of Andrews' (2003) end-of-sample instability test in a model that includes only a constant.[34] We want to test whether the average of $\tilde{Y}_t$ is different after the treatment. With group-level covariates, we can estimate the OLS DID model and then construct $\tilde{Y}_t$ using $Y_{jt} - X'_{jt}\hat{\beta}$. Since $\hat{\beta}$ is consistent, this approach will work under strict stationarity and ergodicity of $\eta_{jt}$. The same approach works if we have individual-level covariates.[35]

This approach might be interesting because we do not need to assume the structure of the heteroskedasticity. Also, this approach works even if we have as few as one treated and one control group. However, this approach is unfeasible if there are few pre-treatment periods. Moreover, the stationarity assumption might be violated if, for example, there is variation in the number of observations per group across time. For example, if we divide the US states in the CPS by quartiles of number of observations for each year from 1979 to 2014, then 35 out of the 51 states belonged to 3 or 4 different quartiles depending on the survey year. In this scenario, our method using the function $var(W_j|\widehat{\{M(j,t)\}}_{t=1}^{T})$ would still provide a valid alternative, provided that we have a large number of control groups and we know how the heteroskedasticity was generated.

# 5 Linear Factor Model - Large $t^*$ and Large $N_0$

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

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

$$Y_{jt}(0) = x_{jt}\beta + f'_t\lambda_j + \eta_{jt}^{LFM} \tag{10}$$

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

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

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

$$Y_{jt}(1) = Y_{jt}(0) + \alpha_{jt} \tag{11}$$

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

As suggested in Gobillon and Magnac (2013), it is possible to estimate this model in two steps. In the first step, we estimate the linear factor model in equation 10 using the sample composed of non-treated observations over the whole period and treated observations in the pre-treatment ($t \leq t^*$). If $t^*$ and $N_0$ tend to $\infty$, then we get consistent estimators for $\beta$, $f_t$ and $\lambda_t$. In the second step, we estimate the counterfactual term imputing the estimated $\beta$, $f_t$ and $\lambda_t$. More specifically, we have that the average treatment on the treated effect in period $t$ is given by:

$$\alpha_t \equiv E[Y_{jt}(1) - Y_{jt}(0)| \text{ treated }] = E[\alpha_{jt}| \text{ treated }] = E[Y_{jt} - x_{jt}\beta - \lambda_i' f_t| \text{ treated }] \tag{12}$$

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

$$\begin{aligned}\hat{\alpha}_t &= \frac{1}{N_1}\sum_{j=1}^{N_1}\left[Y_{jt} - x_{jt}\hat{\beta} - \hat{\lambda}_i'\hat{f}_t\right] \xrightarrow{d} \frac{1}{N_1}\sum_{j=1}^{N_1}[Y_{jt} - x_{jt}\beta - \lambda_i' f_t] = \\ &= E[\alpha_{jt}| \text{ treated }] + \frac{1}{N_1}\sum_{j=1}^{N_1}\eta_{jt}^{LFM}\end{aligned} \tag{13}$$

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

This result is a natural extension of CT.[36] The key point is that common factors and factor loads are consistently estimated, so we can use the residuals from the linear factor model $\hat{\eta}_{jt}^{LFP}$ to estimate the distribution of $\eta_{jt}^{LFM}$. This works because as $t^*$ and $N_0$ tend to $\infty$, $\hat{\eta}_{jt}^{LFP} \xrightarrow{d} \eta_{jt}^{LFP}$. Since we have both $t^* \to \infty$ and $N_0 \to \infty$, we have two alternatives in this case. We can exploit the cross-section variation using the estimated residuals from the control groups, $\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\hat{\eta}_{jt}^{LFM}$ for $j > N_1$, to approximate the distribution of the errors of the treated groups, $\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\eta_{jt}^{LFM}$ for $j \leq N_1$. Under homoskedasticity across $j$, this is essentially the method presented in CT applied to linear factor models with few treated groups. If errors are heteroskedastic, then we can use our method, provided that we know how the heteroskedasticity was generated. Alternatively, we can exploit the time series variation as shown in Section 4 provided that $\eta_{jt}^{LFM}$ is strictly stationary and ergodic.

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

# 6 Monte Carlo Evidence

In this section, we provide Monte Carlo evidence of different hypothesis testing methods in DID. We assume that the underlying data generating process (DGP) is given by:

$$Y_{ijt} = \nu_{jt} + \epsilon_{ijt} \tag{14}$$

In our simulations, we estimate a DID model given by equation 4 where only $j = 1$ is treated and $T = 2$, and then we test the null hypothesis of $\alpha = 0$ using different hypothesis testing methods. We focus on the case with $j = 1$ as this is the case in which no method that allows for unrestricted heteroskedasticity provides reliable inference. We consider variations in the DGP along three dimensions:

1. The number of groups: $N_0 + 1 \in \{25, 100\}$.
2. The intra-group correlation: $\nu_{jt}$ and $\epsilon_{ijt}$ are drawn from normal random variables. We hold constant the total variance $var(\nu_{jt} + \epsilon_{ijt}) = 1$, while changing $\rho = \frac{\sigma^2_\nu}{\sigma^2_\nu + \sigma^2_\epsilon} \in \{.01\%, 1\%, 4\%\}$.
3. The number of observations within group: we draw, for each group $j$, $M_j$ from a discrete uniform random variable with range $[\underline{M}, \overline{M}] \in \{[50, 200], [200, 800], [50, 950]\}$.[37]

For each case, we simulated 100,000 estimates. We present rejection rate results for inference using robust standard errors in the individual-level OLS regression, and for the cluster residual bootstrap with and without our heteroskedasticity correction. All Results using DL and CT methods are similar to the results using cluster residual bootstrap without heteroskedasticity correction, as presented in the Appendix Tables. We do not include in the simulations methods that allow for unrestricted heteroskedasticity. As explained in Section 2.1, these methods do not work well when there is only one treated group. We also do not include the method suggested by MacKinnon and Webb (2015a) in the simulations because their method collapses to CT when there is only one treated group. We present in Appendix A.6 simulations based on the method proposed by CT in their online appendix. In both MC simulations and in simulations with the CPS, we show that their method leads to significant size distortions when $T > 2$ and there is serial correlation in the individual-level error.

## 6.1 Test Size

We present in Panel A of Table 1 results from simulations using 100 groups (one treated and 99 controls) for different values of the intra-group correlations. Column 1 shows that average rejection rates for a test with 5% significance using robust standard errors in the individual-level DID regression. The rejection rate is slightly higher than 5% when the intra-group correlation $\rho = 0.01\%$ (5.4%), but increases sharply for larger values of the intra-group correlation. The rejection rate is 19% when $\rho = 1\%$ and 42% when $\rho = 4\%$. With cluster residual bootstrap without correction, the average rejection rate is always around 5% (column 3 of Table 1). However, this average rejection rate hides an important variation with respect to the number of observations in the treated group ($M_1$).

In Figure 1.A, we show rejection rates for cluster residual bootstrap without correction conditional on the size of the treated group for the case with $\rho = 0.01\%$. The rejection rate is around 14% when the treated

[37] In the Monte Carlo simulations, we always consider the case $M(j,t) = M_j$. In each simulations, $\{M_j\}_{j=1}^N$ is redrawn according to the distribution of $M_j$ considered in the DGP. In the simulations with real datasets in Section 7, there is variation in $M(j,t)$ across $t$.

group is in the first decile of number of observations per group, while it is only 0.8% when the treated group is in the 10th decile. Note also that this distortion in rejection rates is not confined to the extremes of the distribution of group sizes. For example, the rejection rate is 3% when the treated group is in the 6th decile of number of observations per group. We summarize this variation in rejection rates by looking at the absolute difference in rejection rates for each decile of $M_1$ relative to the average rejection rate. Then we average these absolute differences across deciles. We call this measure "relative size distortion". We present these results in column 4 of Table 1 for the bootstrap without heteroskedasticity correction. Conditional on the number of observations of the treated group, these methods present a relative size distortion in the rejection rates of 3.4 percentage points for a 5% significance test when $\rho = 0.01\%$. We present rejection rates by decile of the treated group for cluster residual bootstrap without correction when $\rho = 1\%$ and when $\rho = 4\%$ in Figures 1.B and 1.C, respectively. As expected, this variation in rejection rates becomes less relevant when the intra-group correlation becomes stronger. This happens because the aggregation from individual to group x time averages induces less heteroskedasticity in the residuals when a larger share of the residual is correlated within group. Still, even when $\rho = 4\%$ the difference in rejection rates by number of observations in the treated group remains relevant. The rejection rate is around 6.5% when the treated group is in the first decile of number of observations per group, while it is 4.2% when the treated group is in the 10th decile. The relative size distortion in rejection rates for the bootstrap without correction is around 0.7 percentage points in this scenario (column 4 of Table 1). Inference using DL or CT methods present similar size distortions, as presented in Appendix Table A.1.

Given that inference using these methods is problematic when there is variation in the number of observations per group, we consider our residual bootstrap method with heteroskedasticity correction derived in Section 2.3. We present rejection rates by decile of the treated group when the intra-group correlation is 0.01%, 1%, and 4% in Figures 1.D to 1.F. Average rejection rates using our method are always around 5% and, more importantly, there is no variation with respect to the number of observations in the treated group. These results are also presented in columns 5 and 6 of Table 1. The relative size distortion in rejection rates is only around 0.2-0.3 percentage points, regardless of the value of the intra-group correlation.

We present in Appendix Table A.1 the simulation results with variations in the distribution of group sizes. We first change the range of the distribution of $M_j$ from $[50, 200]$ to $[200, 800]$. This way, we increase the number of observations per group while holding the ratio between the number of observations in different groups constant. Increasing the number of observations per group ameliorates the problem of (over-) under-rejecting the null when $M_1$ is (small) large relative to the number of observations in the control groups when $\rho = 1\%$ or $\rho = 4\%$. However, increasing the number of observations has no detectable effect when the intra-group correlation is 0.01%. In this case, 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 2.2. We also present in Appendix Table A.1 simulations when $M_j$ varies from 50 to 950. Therefore, the average number of observations remains constant, but we have more variation in $M_j$ relative to the $[200, 800]$ case. As expected, more variation in the number of observations per group worsens the inference problem we highlight with the bootstrap without correction. Importantly, our residual bootstrap with heteroskedasticity correction remains accurate irrespective of the variation in the number of observations per group.

As presented in Section 2.3, our method works asymptotically when $N_0 \to \infty$. This assumption is important for two reasons. First, as in any other cluster bootstrap method, a small number of groups implies a small number of possible distinct pseudo-samples. In this case, the bootstrap distribution will not be smooth even with many bootstrap replications (Cameron et al. (2008)). Additionally, our method

requires that we estimate $var(W_j|M_j)$ using the group x time aggregate data so that we can apply our heteroskedasticity correction. If there are only a few groups, then our estimator of $var(W_j|M_j)$ will be less precise. In particular, it might be the case that $\widehat{var(W_j|M_j)} < 0$ for some $j$, which implies that we would not be able to normalize the residual of observation $j$. When $\widehat{var(W_j|M_j)} < 0$ for some $j$, we use the following rule: if $\hat{A} < 0$, then we use $\widehat{var(W_j|M_j)} = \frac{1}{M_j}$, as $\hat{A} < 0$ would suggest that there is not a large intra-group correlation problem. If $\hat{B} < 0$, then we use $\widehat{var(W_j|M_j)} = 1$, as $\hat{B} < 0$ would suggest that there is not much heteroskedasticity. It is important to note that asymptotically this rule would not be relevant, since $var(W_j|M_j) > 0$ for all $M$. We had $\widehat{var(W_j|M_j)} > 0$ for all $j$ in more than 99% of our simulations with $N = 100$. However, when there are fewer control groups, the function $var(W_j|M_j)$ will be estimated with less precision.

We present in Panel B of Table 1 and in Figure 2 the simulation results when the total number of groups is 25. Average rejection rates are slightly higher for both bootstraps with and without correction, at 5.3-5.6%.[38] As shown in Figure 2, there is a minor distortion in rejection rates when the treated group is in the first decile of group size when using our bootstrap method with heteroskedasticity correction. Still, our method provides reasonably accurate hypothesis testing even with 25 groups. In particular, our method provides substantial improvement in relative size distortion when compared to the bootstrap without correction, especially when intra-group correlation is not too strong, as presented in column 6 of Table 1.

As a robustness check, we consider in Appendix Table A.3 alternative data generating processes for $\nu_{jt}$. In panel A we consider the case with a chi-squared distribution with one degree of freedom, in panel B we consider the case with a student-t distribution with 3 degrees of freedom, and in panel C we consider the case with a binary distribution. Note that our assumption 3 is not valid under these data generating processes. Our simulation results suggest that our method still provides reliable inference in these settings. The simulation results are very similar to the case where $\nu_{jt}$ is normally distributed. In particular, our method substantially improves relative to a bootstrap without correction when the intra-group correlation is not too strong. In Section 7 we provide evidence that our inference method also improves inference relative to alternative methods in simulations with real datasets, where we do not have control over the DGP.

## 6.2 Test Power

We have focused so far on Type I error. We saw in Section 6.1 that our method is efficient in providing tests that reject the null with the correct size when the null is true. We are interested now in whether our tests have power to detect effects when the null hypothesis is false. We run the same simulations as in Section 6.1, with the difference that we now add an effect of $\beta$ standard deviations for observation $\{ijt\}$ when $d_{jt} = 1$. Then we calculate rejection rates using our method. Given that we know the DGP in our Monte Carlo simulations, we can calculate the variance of $\hat{\alpha}$ given the parameters of the model and generate an (infeasible) t-statistic $t = \frac{\hat{\alpha}}{\sigma_{\hat{\alpha}}}$. Then we also calculate rejection rates based on this test statistic. Note that with two periods and one treated group, with $N_0 \to \infty$, the DID OLS estimator is asymptotically equivalent to the GLS estimator where the full structure of the variance/covariance matrix is known. Therefore, since the errors in our DGP are normally distributed, we also know that a test based on this t-statistic is the uniformly most powerful test (UMP) for this particular case. Given our results from Section 3, we know that our inference method has asymptotically the same power of the UMP test.

In Figures 3.A to 3.C, we present power results for different intra-group correlation parameters when

[38] Average rejection rates are approximately 8% for the original CT method, as presented in Appendix Table A.2.

there are 100 groups (1 treated and 99 control groups) separately when the treated group is above and below the median of number of observations per group. The most important feature in these graphs is that, for this particular DGP, the power of our method converges to the power of the UMP test when we have many control groups in all intra-group correlation and group size scenarios. It is also interesting to note that the power is higher when the treated group is larger. This is reasonable, since the main component of the variance of the DID estimator with few treated and many control groups comes from the variance of the treated groups. The difference in power for above- and below-median treated groups vanishes when the intra-group correlation increases. This happens because a higher intra-group correlation makes the model less heteroskedastic, so the size of the treated group would be less related to the precision of the estimator. Finally, the power of the test decreases with the intra-group correlation which reflects that, for a given number of observations per group, a higher intra-group correlation implies more volatility in the group x time regression.

When we have 25 groups (1 treated and 24 control), then the power of our method is slightly lower than the power of the UMP test (Figures 3.D to 3.F). This is partially explained by fact that we need to estimate the function $var(W_j|M_j)$ and, with a finite number of control groups, this function would not be precisely estimated. Still, the power of our method is relatively close to the power of the UMP test, especially when the intra-group correlation is not high.

# 7 Simulations with Real Datasets

The results presented in Section 6 suggest that heteroskedasticity generated by variation in group sizes invalidates inference methods that rely on homoskedasticity such as DL, CT, and cluster residual bootstrap, while our method performs well in correcting for heteroskedasticity when there are 25 or more groups. However, a natural question that arises is whether these results are "externally valid." In particular, we want to know (i) whether heteroskedasticity generated by variation in group sizes is a problem in real datasets with large number of observations, and (ii) whether our method works in real datasets, where we do not have control over the DGP. More specifically, our DGP in Section 6 implies that the *real* variance of $W_j$ would have exactly the relationship $var(W_j|M_j) = A + \frac{B}{M_j}$, which might not be the case in real datasets. To illustrate the magnitude of the heteroskedasticity problem and to test the accuracy of our method, we conduct simulations of placebo interventions using two different real datasets: the American Community Survey (ACS) and the Current Population Survey (CPS).[39]

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

We use the ACS dataset for the years 2000 to 2015, and the CPS Merged Outgoing Rotation Groups for the years 1979 to 2015.[40] We extract information on employment status and earnings for women between ages 25 and 50, following Bertrand et al. (2004). We present in Table 2 the distribution of number of

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

[40] For simulations using the ACS at the PUMA level, there is only information available from 2005 to 2015.

observations per group x cell for the PUMA-level ACS (column 1), for the state-level ACS (column 2) and for the state-level CPS (column 3). Considering the 2015 dataset, there are, on average, 505 observations in each PUMA x time cell in the ACS. This number, however, hides an important heterogeneity in cell sizes. The 10th percentile of PUMA x time cell sizes is 152, while the 90th percentile is 923. There is also substantial heterogeneity in state x time cell sizes in the ACS. While the average cell size is 9725, the 10th percentile is 1,290, while the 90th percentile is 18,913.[41] Finally, the state x time cells in the CPS have substantially fewer observations compared to the ACS. While the average cell size is 666, the 10th percentile is 376, and the 90th percentile is 857.

For the ACS simulations, we consider pairs of two consecutive years and estimate placebo DID regressions using one of the groups (PUMA or state) at a time as the treated group. Note that this differs from Bertrand et al. (2004) simulations, as they randomly selected half of the states to be treated. In each simulation, we test the null hypothesis that the "intervention" has no effect ($\alpha = 0$) using robust standard errors, and bootstrap with and without our heteroskedasticity correction. Since we are looking at placebo interventions, if the inference method is correct, then we would expect to reject the null roughly 5% of the time for a test with 5% significance level. For each pair of years, the number of PUMAs that appear in both years ranges from 427 to 982, leading to 7,152 regressions in total. For the state-level simulations, we have $51 \times 15 = 765$ regressions.[42] For the CPS simulations, we used 2, 4, 6, or 8 consecutive years, always using the first half of the years as pre-treatment and the second half as post-treatment. This leads to 1530 to 1836 regressions, depending on the number of years used in each regression.

## 7.1 American Community Survey (ACS) Results

In Panel A of Table 3, we present results from simulations using the PUMA-level treatments using the ACS. In column 1, we show rejection rates using OLS robust standard errors in the individual-level DID regression. Rejection rates for a 5% significance test are 6.8% when the outcome variable is employment, and 7.8% when it is log wages. This over-rejection suggests that there is some intra-group correlation that the robust individual-level standard error does not take into account. In column 3 of Table 3, we present results for the bootstrap without the heteroskedasticity correction (results for DL and CT are simular). As in the Monte Carlo simulations, average rejection rates without correction are very close to 5%. However, there is substantial variation when we look at rejection rates conditional on the size of the treated group. We present in column 4 of Table 3 the difference in rejection rates when the number of observations in the treated group is above and below the median.[43] For both outcome variables, the rejection rate is around 8 percentage points lower when the treated group has a group size above the median. This implies a rejection rate of around 9% when the treated group is below the median, and around 1% when the treated group is above the median. In columns 5 and 6 of Table 3, we present the rejection rates using bootstrap with our heteroskedasticity correction.[44] For both outcomes, average rejection rate has the correct size of 5% and, more importantly, there is virtually no difference between rejection rates when the treated group is above or below the median. Therefore, our method was successful in correcting for the heteroskedasticity problem

[41] The number of observations in the ACS increased substantially starting from the 2005 ACS. All results remain similar if we consider only the ACS data from 2005 to 2015.

[42] We include Washington, D.C.

[43] Given that we have a limited number of simulations, we do not calculate the relative size distortion in rejection rates across deciles, as we do in the Monte Carlo simulations. For the PUMA-level simulations, there are only approximately 700 simulations for each decile. For the state-level simulations there would be only around 70 simulations for each decile.

[44] In all simulations using real data, we use the version of our method that allows for samplings weights, as described in Appendix A.4.

even in a setting where we do not have control over the DGP.

We present in Panel B of Table 3 the results for state-level simulations. The most striking result in this table is that rejection rates using bootstrap without correction still depend on the size of the treated group. This happens in a dataset with, on average, around 10,000 observations per group x time cell. In particular, the rejection rate in the simulations with log wages as the outcome variable is zero when the treated group is below the median, and 10% when the treated group is above the median. We present rejection rates using bootstrap with our heteroskedasticity correction in columns 5 and 6. Average rejection rates are around 5%, and, more importantly, there is no significant difference in rejection rates depending on the size of the treated state.

## 7.2 Current Population Survey (CPS) Results

We present the simulation results using the CPS in Table 4. Panel A presents rejection rates of DID models using 2 years of data, while Panels B, C, and D present rejection rates using respectively 4, 6, and 8 years. Inference with OLS robust standard errors on the individual-level model becomes worse when we include more years of data in the model (column 1). This result is consistent with the findings in Bertrand et al. (2004). The key point is that the panel structure of the CPS Merged Outgoing Rotation Groups generates serial correlation in the errors. We present rejection rates for the residual bootstrap without correction in columns 3 and 4. The average rejection rates are close to 5% irrespective of the number of periods, which was expected given that this method takes serial correlation into account by looking at a linear combination of the residuals (as in CT). However, since this linear combination of the residuals is heteroskedastic, rejection rates based on this method vary significantly with the size of the treated group. We present rejection rates using bootstrap with our heteroskedasticity correction in columns 5 and 6. As in the ACS simulations, the results indicate that on average rejection rates have the correct size and that rejection rates do not depend on the size of the treated group in all simulations. Therefore, our method is efficient in correcting for heteroskedasticity in a scenario that serial correlation is important without the need to specify the structure of the serial correlation.

## 7.3 Power with Real Data Simulations

We saw in Sections 7.1 and 7.2 that our method provides tests with correct size in simulations with the ACS and the CPS. We now present power results from simulations with these datasets in Figure 4[45]. Figure 4.A shows power results using the ACS with state-level treatment. When the treated group is above the median, our method is able to detect an effect size of 0.07 log points with probability approximately equal to 80%. When the treated group is below the median, we are only able to attain this power for effects greater than 0.1 log points. This again reflects that the variance of $\hat{\alpha}$ is higher when the treated group is smaller. Figures 4.B to 4.E present results for simulations using the CPS with different numbers of time periods. The power in the CPS simulations is considerably lower than in the ACS simulations. The power to reject an effect of 0.07 log points when the treated group is above the median ranges from 32% to 49%, depending on the number of periods used in the simulations. This happens because the ACS has a much larger number of observations than the CPS. Even though we have only one treated group in all simulations, the larger

[45] As in the MC simulations, to calculate the power in these simulations with real data, we add an effect of $\beta$ for the unit that is randomly selected to be treated

number of observations in the ACS implies that the group x time variance of the error would be smaller.[46]

# 8 Conclusion

This paper shows that usual inference methods used in DID models might not perform well in the presence of heteroskedasticity when the number of treated groups is small. Then we derive an alternative inference method that corrects for heteroskedasticity when there are few treated groups (or even just one) and many control groups. With few pre-treatment periods, the main assumption is that we can model the heteroskedasticity of a *linear combination* of the errors. By focusing on this linear combination, we circumvent the incidental parameter problem and avoid imposing strong assumptions on the serial correlation. We focus on the example of variation in group sizes, in which it is possible to derive a parsimonious function for the conditional variance as a function of the number of observations per group under very mild assumptions on the errors. However, our model is more general and can be applied in any situation in which we are able to estimate (parametrically or non-parametrically) the conditional distribution of $W_1$ using the residuals of the control groups. It is important to note that there is no heteroskedasticity-robust inference method in DID when there is only one treated group. Therefore, although our method is not robust to any form of unknown heteroskedasticity, it provides an important improvement relative to existing methods that rely on homoskedasticity. Our method can also be combined with FGLS estimation, providing a safeguard in situations a where a t-test based on the FGLS estimator would be invalid.

Our method does not impose any restriction on the intra-group correlation. In particular, it does not require the specification of the serial correlation. With only one treated group, we show that it is only possible to relax the assumption that, conditional on a set of covariates, the distribution of $W_j$ does not depend on treatment status if we impose alternative restrictions on the intra-group correlation. With many pre-treatment periods, we provide an alternative inference method that relies on strict stationarity and ergodicity of the time series instead of the assumption on how the heteroskedasticity was generated. Finally, we extend our inference method to linear factor models with few treated groups, an estimation method that has been recently proposed as an alternative to DID when there are many pre-treatment periods and many control groups.

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

# References


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

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

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

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

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

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

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

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

**Behrens, W. U.**, "Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen," *Landwirtschaftliche Jahrbucher.*, 1929, *68*, 807–837.

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

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

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

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

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

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

**Conley, Timothy and Christopher Taber**, "Inference with "Difference in Differences" with a Small Number of Policy Changes," Working Paper 312, National Bureau of Economic Research July 2005.

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

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

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

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

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

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

**Ferman, Bruno and Cristine Pinto**, "Placebo Tests for Synthetic Controls," MPRA Paper 78079, University Library of Munich, Germany April 2017.

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

_ , *The design of experiments. 1935*, Edinburgh: Oliver and Boyd, 1935.

_ , "The comparison of sample with possibly unequal variances," *Annals of Eugenics*, 1939, *9* (2), 380–385.

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

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

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

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

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

**Ibragimov, Rustam and Ulrich K. Müller**, "t-Statistic Based Correlation and Heterogeneity Robust Inference," *Journal of Business & Economic Statistics*, 2010, *28* (4), 453–468.

_ **and** _ , "Inference with Few Heterogeneous Clusters," *The Review of Economics and Statistics*, March 2016, *98* (1), 83–96.

**Imbens, Guido W. and Michal Kolesar**, "Robust Standard Errors in Small Samples: Some Practical Advice," *The Review of Economics and Statistics*, 2016, *98* (4), 701–712.

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

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

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

_ **and** _ , "Wild Bootstrap Inference for Wildly Different Cluster Sizes," Working Papers 1314, Queen's University, Department of Economics February 2015.

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

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

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

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

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

**Scheffe, H.**, "Practical solutions of the Behrens-Fisher problem," *Journal of the American Statistical Association.*, 1970, *65*, 1501–1508.

**Wang, Y. Y.**, "Probabilities or the Type l errors of the Welch tests for the Behrens-Fisher problem," *Journal of the American Statistical Association.*, 1971, *66*, 605–608.

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

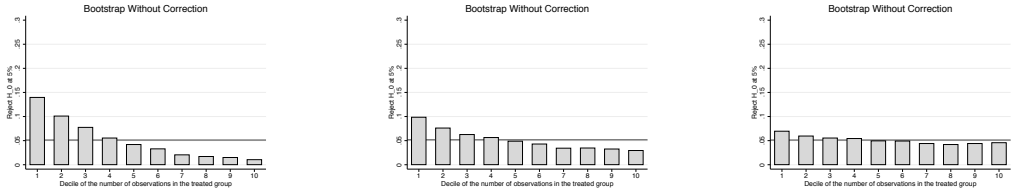
**White, Halbert**, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, May 1980, *48* (4), 817–838.

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

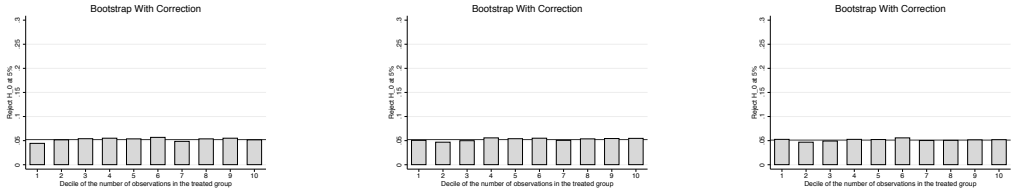
**Yates, F.**, "Tests of Significance for 2  2 Contingency Tables," *Journal of the Royal Statistical Society. Series A (General)*, 1984, *147* (3), 426–463.

Figure 1: **Rejection Rates in MC Simulations by Decile of $M_1$, $N = 100$**

1.A: w/o correction, $\rho = 0.01\%$ 1.B: w/o correction, $\rho = 1\%$ 1.C: w/o correction, $\rho = 4\%$



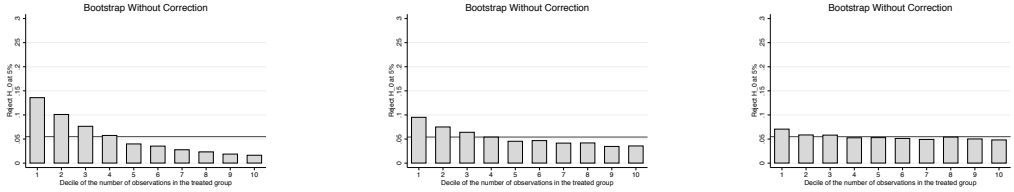

1.D: with correction, $\rho = 0.01\%$ 1.E: with correction, $\rho = 1\%$ 1.F: with correction, $\rho = 4\%$



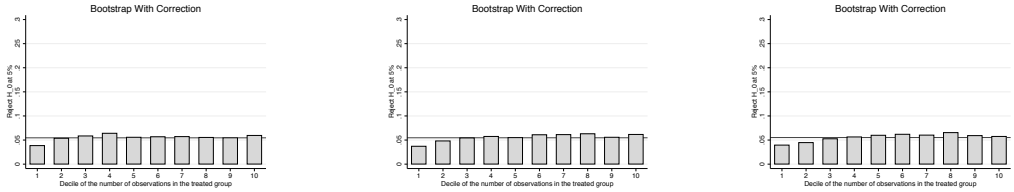

Notes: These figures present the rejection rates conditional on the decile of the number of observation of the treated group when $N = 100$ and $M \in [50, 200]$. These rejection rates are based on Monte Carlos simulations explained in Section 6. Figures 1.A to 1.C present results using the residual bootstrap without correction, while Figures 1.D to 1.F present results using the residual bootstrap method with our heteroskedasticity correction, as explained in Section 2.3.

Figure 2: **Rejection Rates in MC Simulations by Decile of $M_1$, $N = 25$**

2.A: w/o correction, $\rho = 0.01\%$ 2.B: w/o correction, $\rho = 1\%$ 2.C: w/o correction, $\rho = 4\%$

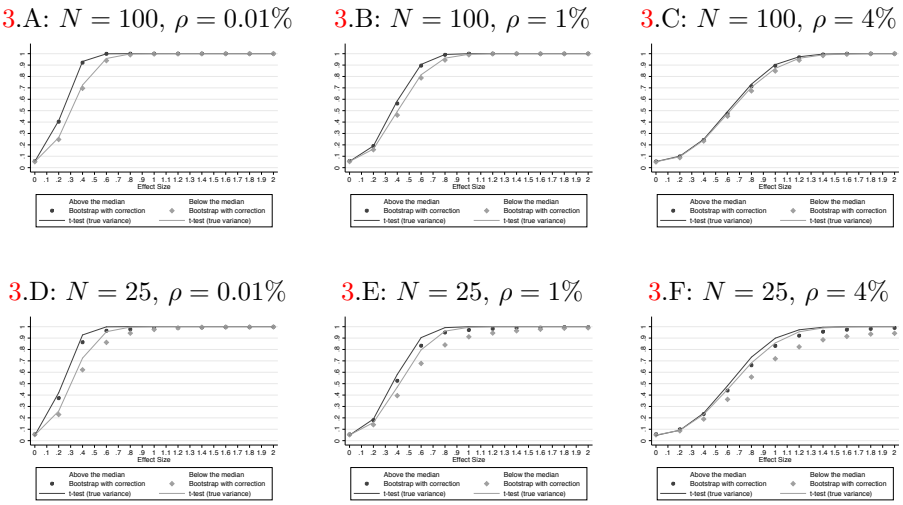



2.D: with correction, $\rho = 0.01\%$ 2.E: with correction, $\rho = 1\%$ 2.F: with correction, $\rho = 4\%$



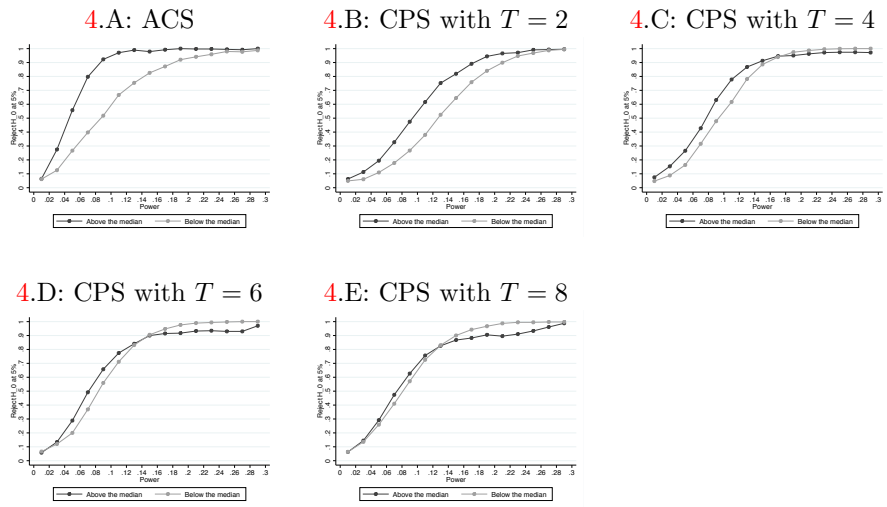

Notes: These figures present the rejection rates conditional on the decile of the number of observation of the treated group when $N = 25$ and $M \in [50, 200]$. These rejection rates are based on Monte Carlos simulations explained in Section 6. Figures 2.A to 2.C present results using the residual bootstrap without correction, while Figures 2.D to 2.F present results using the residual bootstrap method with our heteroskedasticity correction, as explained in Section 2.3.

Figure 3: **Test Power - Monte Carlo Simulations**




Notes: These figures present the power of the bootstrap with heteroskedasticity correction as a function of the effect size separately when the treated group is above and below the median of group size. The standard deviation of the individual-level observation is equal to one across the different scenarios. Therefore, the effect size is in standard deviation terms. In all simulations $M_j \in [50, 200]$.

Figure 4: **Test Power by Treated Group Size - Simulations with Real Dataset**




Notes: These figures present the power of the bootstrap with heteroskedasticity correction for simulations using real datasets. Results are presented separately when the treated group is above and below the median of group size. The outcome variable is log wages, and effect sizes are measured in log points. Figure 4.A presents results using the ACS, while Figures 4.B to 4.E present results using the CPS with varying number of periods.

Table 1: **Rejection Rates in MC Simulations**

| | Robust OLS | | Bootstrap w/o correction | | Bootstrap with correction | |
|---|---|---|---|---|---|---|
| $\rho$ | Mean | Relative size distortion | Mean | Relative size distortion | Mean | Relative size distortion |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Panel A: $N = 100$ | | | | | | |
| 0.01% | 0.054 | 0.003 | 0.051 | 0.034 | 0.052 | 0.003 |
| 1% | 0.193 | 0.032 | 0.052 | 0.017 | 0.052 | 0.002 |
| 4% | 0.418 | 0.062 | 0.051 | 0.007 | 0.051 | 0.002 |
| Panel B: $N = 25$ | | | | | | |
| 0.01% | 0.052 | 0.002 | 0.053 | 0.032 | 0.055 | 0.004 |
| 1% | 0.193 | 0.032 | 0.053 | 0.015 | 0.055 | 0.005 |
| 4% | 0.424 | 0.055 | 0.054 | 0.005 | 0.056 | 0.006 |

Notes: This table presents results from Monte Carlo simulations with 100 groups, as explained in Section 6. In all simulations, only one group is treated. Each line presents simulation for different values of intra-group correlation. All simulations consider $M \in [50, 200]$. We consider 3 inference methods: hypothesis testing using robust standard errors from the individual level regression, cluster residual bootstrap without correction, and cluster residual bootstrap with our heteroskedasticity correction. For each inference method, we report the average rejection rate for a 5% significance level test. We also report a measure of how rejection rates depend on the number of observations in the treated group, which we call "relative size distortion". To construct this measure, we calculate the absolute difference in rejection rates for each decile of $M_1$ relative to the average rejection rate, and then we average these absolute differences across deciles. We run 100,000 simulations for each scenario. The standard error for the average rejection rates is around 0.07 percentage points, while the standard error for the absolute difference is around 0.04-0.07 percentage points.

Table 2: **Number of Observations per Group x Time cell**

| | ACS | | CPS |
|---|---|---|---|
| | PUMA | State | State |
| | (1) | (2) | (3) |
| Average | 505.10 | 9,725.59 | 666.45 |
| 1% | 116 | 868 | 269 |
| 5% | 140 | 1,018 | 337 |
| 10% | 152 | 1,290 | 376 |
| 25% | 186 | 2,306 | 435 |
| 50% | 253 | 6,713 | 551 |
| 75% | 451 | 11,598 | 725 |
| 90% | 923 | 18,913 | 857 |
| 95% | 1,494 | 32,185 | 1,485 |
| 99% | 5,105 | 63,360 | 3,104 |

Notes: This table presents the distribution of number of observations per groups for the datasets used in our simulations. We present information for the 2015 ACS and for the 2015 CPS. Column 1 presents information for PUMA-level ACS simulations, column 2 presents information for state-level ACS simulations, while column 3 presents information for state-level CPS simulations.

Table 3: **Simulations with the ACS Survey**

| | Inference Method | | | | | |
|---|---|---|---|---|---|---|
| | Robust OLS | | Bootstrap w/o correction | | Bootstrap with correction | |
| Outcome Variable | Mean (1) | Diff (2) | Mean (3) | Diff (4) | Mean (5) | Diff (6) |
| Panel A: ACS with PUMA level interventions | | | | | | |
| Employment | 0.068*** | 0.012* | 0.050 | -0.076*** | 0.050 | -0.003 |
| | (0.003) | (0.006) | (0.003) | (0.005) | (0.003) | (0.005) |
| Log(wages) | 0.078*** | 0.004 | 0.049 | -0.082*** | 0.051 | -0.001 |
| | (0.003) | (0.007) | (0.003) | (0.005) | (0.003) | (0.006) |
| Panel B: ACS with state level interventions | | | | | | |
| Employment | 0.052 | 0.007 | 0.047 | -0.092*** | 0.056 | 0.010 |
| | (0.008) | (0.016) | (0.014) | (0.023) | (0.009) | (0.017) |
| Log(wages) | 0.071* | -0.002 | 0.052 | -0.103*** | 0.055 | -0.014 |
| | (0.011) | (0.021) | (0.017) | (0.029) | (0.010) | (0.018) |

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

Table 4: **Simulations with the CPS Survey**

| | Inference Method | | | | | |
|---|---|---|---|---|---|---|
| | Robust OLS | | Bootstrap w/o correction | | Bootstrap with correction | |
| Outcome Variable | Mean (1) | Diff (2) | Mean (3) | Diff (4) | Mean (5) | Diff (6) |
| | | | Panel A: 2 years | | | |
| Employment | 0.046 | -0.007 | 0.046 | -0.042*** | 0.051 | 0.008 |
| | (0.007) | (0.011) | (0.007) | (0.011) | (0.007) | (0.012) |
| Log(wages) | 0.068*** | -0.001 | 0.045 | -0.033*** | 0.053 | 0.011 |
| | (0.006) | (0.013) | (0.006) | (0.012) | (0.006) | (0.012) |
| | | | Panel B: 4 years | | | |
| Employment | 0.063** | 0.012 | 0.043 | -0.038*** | 0.051 | -0.006 |
| | (0.007) | (0.012) | (0.007) | (0.013) | (0.006) | (0.013) |
| Log(wages) | 0.100*** | 0.033 | 0.050 | -0.035** | 0.052 | 0.013 |
| | (0.011) | (0.021) | (0.007) | (0.014) | (0.007) | (0.015) |
| | | | Panel C: 6 years | | | |
| Employment | 0.087*** | -0.006 | 0.052 | -0.045*** | 0.053 | -0.015 |
| | (0.008) | (0.017) | (0.006) | (0.012) | (0.006) | (0.013) |
| Log(wages) | 0.141*** | 0.053** | 0.050 | -0.038** | 0.053 | 0.002 |
| | (0.014) | (0.027) | (0.008) | (0.015) | (0.009) | (0.016) |
| | | | Panel D: 8 years | | | |
| Employment | 0.132*** | 0.022 | 0.048 | -0.045*** | 0.048 | -0.016 |
| | (0.013) | (0.023) | (0.008) | (0.014) | (0.008) | (0.015) |
| Log(wages) | 0.207*** | 0.022 | 0.048 | -0.028 | 0.053 | 0.004 |
| | (0.015) | (0.033) | (0.010) | (0.019) | (0.010) | (0.019) |

Notes: This table presents rejection rates for the simulations using CPS data. In each simulation, we run a DID regression using one group as treated and the other groups as a control. The outcome variable is employment status or log(wages) for women aged between 25 and 40. Then we test the hypothesis that the effect of the "intervention" is equal to zero using different inference methods: hypothesis testing using robust standard errors from individual level DID model, bootstrap without and bootstrap with our heteroskedasticity correction. Panel A reports results of DID models using 2 consecutive years of data, while Panels B, C, and D report results of DID models using respectively 4, 6, and 8 consecutive years of data. We report average rejection rate (columns 1, 3, and 5) and the difference in rejection rates when the size of the treated group is above or below the median (columns 2, 4, and 6). Given that we have a limited number of simulations, we do not calculate the relative size distortion in rejection rates across deciles, as we do in the Monte Carlo simulations. We present in brackets standard errors for the rejection rates. Standard errors are clustered at the treated group level. For average rejection rates (columns 1, 3, and 5), * means that we reject at 10% that the average rejection rate is equal to 5%, while for the differences in rejection rates (columns 2, 4, and 6) * means that we reject at 10% that rejection rate for $M_1$ above and below the median are equal. ** means that we reject at 5%, while *** means that we reject at 1%.

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

## A.1 Proof of the Main Results

This supplemental appendix contains the main theorems and proofs of the paper "Inference in Differences-in-Differences with Different Group Sizes". We use the same notation as in the main paper.

The aggregated model is:

$$Y_{jt} = \alpha d_{jt} + \theta_j + \beta X_{jt} + \gamma_t + \eta_{jt} \tag{15}$$

where $X_{jt}$ is a $k$x1 vector of covariates. For simplicity, we start with the case that $\beta = 0$ and then extend to the case with covariates.

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

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

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

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

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

$$\widehat{\alpha} = \frac{1}{N_1} \sum_{j=1}^{N_1} \bigtriangledown Y_j - \frac{1}{N_0} \sum_{j=N_1+1}^{N} \bigtriangledown Y_j = \frac{1}{N_1} \sum_{j=1}^{N_1} (\bigtriangledown \widehat{Y}_j^R + \widehat{W}_j^R) - \frac{1}{N_0} \sum_{j=N_1+1}^{N} (\bigtriangledown \widehat{Y}_j^R + \widehat{W}_j^R)$$

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

Therefore:

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

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

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

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

Define $F_{j|X_j}$ as the conditional distribution function of $W_j$ on $X_j$.

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

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

**Assumption 3 (Heteroskedasticity):** $W_j|\tilde{X}_j$ has the same distribution across $\tilde{X}_j$ up to a scale parameter.

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

**Assumption 5 (Continuity):** $F_{j|X_j}$ is continuous.

**Lemma 1** *Under assumptions 1-5, under the null when $N_0 \to \infty$,*

$$\sup_{w\in\Theta}\left|F_{\widehat{W_j}^R\left|\widetilde{X}_j\right.}(w) - F_{j|\widetilde{X}_j}(w)\right| \to 0 \tag{16}$$

**Proof.** Note that $\widehat{\eta}_{jt}^R$ is the residual of a OLS regression under the null hypothesis, and it can be written as

$$\widehat{\eta}_{jt}^R = Y_{jt} - \alpha_0 d_{jt} - \widehat{\theta}_j - \widehat{\gamma}_t = \widetilde{\widetilde{\eta}}_{jt}$$

where $\widetilde{\widetilde{\eta}}_{jt} = \eta_{jt} - \overline{\eta}_j - \overline{\eta}_t + \overline{\eta}$, with $\overline{\eta}_j = \frac{1}{T}\sum_{t=1}^{T}\eta_{jt}$, $\overline{\eta}_t = \frac{1}{N}\sum_{i=1}^{N}\eta_{jt}$ and $\overline{\eta} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\eta_{jt}$

Therefore:

$$\begin{aligned}\widehat{W}_j^R &= \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\left(\eta_{jt} - \overline{\eta}_j - \overline{\eta}_t + \overline{\eta}\right) - \frac{1}{t^*}\sum_{t=1}^{t^*}\left(\eta_{jt} - \overline{\eta}_j - \overline{\eta}_t + \overline{\eta}\right)\\ &= W_j - \frac{1}{N}\sum_{j'=1}^{N}W_{j'} = W_j + o_p(1) \text{ as } N_0 \to \infty, \text{ by assumptions 1 and 4.}\end{aligned}$$

Hence, $\widehat{W}_j^R$ weakly converge to $W_j$ as $N_0 \to \infty$.

Under assumption 5, lemma 2.11 of Van der Vaart (1998),

$$\sup_{w\in\Theta}\left|F_{\widehat{W_j}^R\left|\widetilde{X}_j\right.}(w) - F_{j|\widetilde{X}_j}(w)\right| = o(1) \tag{17}$$

■

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

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

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

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

**Theorem 2** *Define $d^*_{1-\frac{\alpha}{2}}$ and $d^*_{\frac{\alpha}{2}}$ as the $(1-\frac{\alpha}{2})$th and $\frac{a}{2}$th quantile of the empirical distribution of $(\widehat{\alpha}_b - \alpha_0)$ given $\tilde{X}$, for $b = 1, ..., \mathcal{B}$. Assuming that we know the variance of $W_j|\tilde{X}_j$, under assumptions 1, 2, 3, 4 and 5,* $\Pr\left[d^*_{1-\frac{\alpha}{2}} \le \widehat{\alpha} - \alpha_0 \le d^*_{\frac{\alpha}{2}}\middle|\, \alpha_0, \tilde{X}\right] \to_p 1-\alpha$ *as $N_0 \to \infty$ and $\mathcal{B} \to \infty$.*

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

Under assumptions 1, 2 and 3:

$$\begin{aligned}\Gamma(w) &= \Pr\left[\sum_{j=1}^{N_1} W_j < w|\tilde{X}\right] \\ &= \int ...\int 1\left\{\sum_{j=1}^{N_1} W_j < w\right\} dF_{1|\tilde{X}_1}(W_1)\cdot dF_{2|\tilde{X}_2}(W_2)\cdot ...\cdot dF_{N_1|\tilde{X}_{N_1}}(W_{N_1})\end{aligned}$$

and

$$\begin{aligned}\widehat{\Gamma}_b(w) &= \Pr\left[\sum_{j=1}^{N_1} \widetilde{W}_{j,b} < w|\tilde{X}, b\right] \\ &= \int ...\int 1\left\{\sum_{j=1}^{N_1} W_j < w\right\} d\widehat{F}_{1|\tilde{X}_1}(W_1)\cdot d\widehat{F}_{2|\tilde{X}_2}(W_2)\cdot ...\cdot d\widehat{F}_{N_1|\tilde{X}_{N_1}}(W_{N_1})\end{aligned}$$

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

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

Let $c_{jb} \equiv \sqrt{\frac{Var\left[W_j|\tilde{X}_j\right]}{Var\left[W_{j,b}|\tilde{X}_{j,b}\right]}}$, and define $W'_{j,b} = W_{j,b}\cdot c_{jb}$. Since we known the variance, we can treat $c_{jb}$ a constant. Define $F_{W'_{j,b}|\tilde{X}_j}(w) = Pr\left[W'_{j,b} < w|\tilde{X}_j\right]$ and $F_{\widetilde{W}_{j,b}|\tilde{X}_j}(w) = Pr\left[\widetilde{W}_{j,b} < w|\tilde{X}_j\right]$ . Note that:

$$\begin{aligned}\sup_{w\in\Theta}\left|\widehat{F}_{\widetilde{W}_j|\tilde{X}_j}(w) - F_{j|\tilde{X}_j}(w)\right| &= \sup_{w\in\Theta}\left|\widehat{F}_{\widetilde{W}_j|\tilde{X}_j}(w) - F_{\widetilde{W}_{j,b}|\widetilde{X}_j}(w) + F_{\widetilde{W}_{j,b}|\widetilde{X}_j}(w) - F_{W'_{j,b}|\widetilde{X}_j}(w) + F_{W'_{j,b}|\widetilde{X}_j}(w) - F_{j|\tilde{X}_j}(w)\right| \\ &\leq \sup_{w\in\Theta}\left|\widehat{F}_{\widetilde{W}_j|\tilde{X}_j}(w) - F_{\widetilde{W}_{j,b}|\widetilde{X}_j}(w)\right| + \sup_{w\in\Theta}\left|F_{\widetilde{W}_{j,b}|\widetilde{X}_j}(w) - F_{W'_{j,b}|\widetilde{X}_j}(w)\right| \\ &\quad + \sup_{w\in\Theta}\left|F_{W'_{j,b}|\widetilde{X}_j}(w) - F_{j|\tilde{X}_j}(w)\right|\end{aligned}$$

Since the set of functions $\{W_{j,b}\cdot c_{j,b} - w|\,(y,w)\in\mathcal{Y}\text{ x }\Theta\}$ is contained in a finite-dimensional vector space, it is a VC-class. By the Theorem 2.64 of van der Vaart and Wellner (2000),the class $\mathcal{H} = \left\{1\left\{W'_{j,b} < w\right\}\middle| w\in\Theta\right\}$ has bracketing numbers $N_{[]}(\sqrt{\varepsilon},\mathcal{W},L_2(P)) \leq \frac{2}{\varepsilon}$. By the Glivenko-Cantelli Theorem, this class is P-Glivenko-Cantelli. Using the definition of P-Glivenko-Cantelli class, when $\mathcal{B}\to\infty$.,

$$\sup_{w\in\Theta}\left|\frac{1}{B}\sum_{b=1}^{B}\left\{W'_{j,b} < w\right\} - F_{W'_{j,b}|\widetilde{X}_j}(w)\right| = o_p(1)$$

Note that $c_{j.b}$ is a constant and every sequence of $1\left\{\widetilde{W}_{j,b} < w\right\}$ in $b$ are contained in $\mathcal{H}$, consequently as $\mathcal{B}\to\infty$.

$$\sup_{w\in\Theta}\left|\frac{1}{B}\sum_{b=1}^{B}\left\{\widetilde{W}_{j,b} < w\right\} - F_{\widetilde{W}_{j,b}|\widetilde{X}_j}(w)\right| = o_p(1)$$

Hence,the first term is $o_p(1)$.

Using the results from lemma 1, as $N_0 \to \infty$, $\widetilde{W}_{j,b} \to_d W_{j,b} \cdot c_{j,b}$. By assumption 5, $F_{W_{j.b}|\tilde{X}_j}$ is a continuous distribution, and by a trivial transformation of variable $F_{W'_{j.b}|\tilde{X}_j}$ is also continuous, using lemma 2.11 of van der Vaart (2011),

$$\sup_{w\in\Theta}\left|F_{\widetilde{W}_{j,b}|\tilde{X}_j}(w) - F_{W'_{j,b}|\tilde{X}_j}(w)\right| \to 0.$$

By assumptions 2 and 3, $W'_{j,b} \overset{d}{=} W_j$, and consequently

$$\sup_{w\in\Theta}\left|F_{W'_{j,b}|\tilde{X}_j}(w) - F_{j|\tilde{X}_j}(w)\right| = o_p(1)$$

By the results above,

$$\sup_{w\in\Theta}\left|\Gamma(w) - \widehat{\Gamma}_b(w)\right| = o_p(1)$$

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

$$\widehat{\alpha} - \alpha_0 \overset{d}{\to} \frac{1}{N_1}\sum_{j=1}^{N_1} W_j \text{ and } \widehat{\alpha}_b - \alpha_0 = \frac{1}{N_1}\sum_{j=1}^{N_1}\widetilde{W}_{j,b} \tag{18}$$

Using the results above, we can show that as $N_0 \to \infty$ and $\mathcal{B} \to \infty$.

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

■

The approach proposed to estimate $\widetilde{W}_{j,b}$ is unfeasible since we do know not the variances of $W_j$'s. Theorem 3 shows that if we have a consistent estimator of $\sqrt{\frac{Var\left[W_j|\tilde{X}_j\right]}{Var\left[W_{j,b}|\tilde{X}_{j,b}\right]}}$, we can construct $\widehat{\widetilde{W}}_{j,b} = \widehat{W}^R_{jb} \cdot \sqrt{\frac{\widehat{Var\left[W_j|\tilde{X}_j\right]}}{\widehat{Var\left[W_{j,b}|\tilde{X}_{j,b}\right]}}}$, and use the approach proposed above.

**Theorem 3** *Define $d^*_{1-\frac{\alpha}{2}}$ and $d^*_{\frac{\alpha}{2}}$ as the $(1-\frac{\alpha}{2})$th and $\frac{a}{2}$th quantile of the empirical distribution of $(\widehat{\alpha}_b - \alpha_0)$ given $\tilde{X}$, for $b = 1, ..., \mathcal{B}$. If for each $j$ $\sqrt{\frac{\widehat{Var\left[W_j|\tilde{X}_j\right]}}{\widehat{Var\left[W_{j,b}|\tilde{X}_{j,b}\right]}}}$ is a consistent estimator for $\sqrt{\frac{Var\left[W_j|\tilde{X}_j\right]}{Var\left[W_{j,b}|\tilde{X}_{j,b}\right]}}$, under assumptions 1, 2, 3, 4 and 5, as $N_0 \to \infty$ and $\mathcal{B} \to \infty$.,*

$$\Pr\left[d^*_{1-\frac{\alpha}{2}} \le \widehat{\alpha} - \alpha_0 \le d^*_{\frac{\alpha}{2}} \middle| \alpha_0, \tilde{X}\right] \to_p 1-\alpha \tag{19}$$

**Proof.** Let $\widehat{\widetilde{W}}_{j,b} = \widehat{W}^R_{j,b} \cdot \widehat{c}_{j,b}$, where $\widehat{c}_{j,b}$ is a consistent estimator for $c_{j,b}$ with $c_{j,b} > 0$. In this case, we need to define $\widehat{F}_{j|\tilde{X}}(w) = \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} 1\left\{\widehat{\widetilde{W}}_{j,b} < w\right\}$ and $F_{\widehat{\widetilde{W}}_{j,b}\middle|\tilde{X}}(w) = \Pr\left[\widehat{\widetilde{W}}_{j,b} < w \middle| \tilde{X}_j\right]$. Now, we need to work with,

$$\begin{aligned}\sup_{w\in\Theta}\left|\widehat{F}_{j|\widetilde{X_j}}(w)-F_{j|\widetilde{X_j}}(w)\right| &= \sup_{w\in\Theta}\left|\widehat{F}_{j|\widetilde{X_j}}(w)-F_{\widehat{\widetilde{W}}_{j,b}\left|\widetilde{X_j}\right.}(w)+F_{\widehat{\widetilde{W}}_{j,b}\left|\widetilde{X_j}\right.}(w)-F_{W'_{j.b}|\widetilde{X}}(w)+F_{W'_{j.b}|\widetilde{X}}(w)-F_{j|\widetilde{X_j}}(w)\right| \\ &\leq \sup_{w\in\Theta}\left|\widehat{F}_{j|\widetilde{X_j}}(w)-F_{\widehat{\widetilde{W}}_{j,b}\left|\widetilde{X_j}\right.}(w)\right|+\sup_{w\in\Theta}\left|F_{\widehat{\widetilde{W}}_{j,b}\left|\widetilde{X_j}\right.}(w)-F_{W'_{j.b}|\widetilde{X_j}}\right| \\ &+\sup_{w\in\Theta}\left|F_{W'_{j.b}|\widetilde{X_j}}(w)-F_{j|\widetilde{X_j}}(w)\right|\end{aligned}$$

From the proof of the previous theorem, $\sup_{w\in\Theta}\left|F_{W'_{j.b}|\widetilde{X}}-F_{j|\widetilde{X}}(w)\right|=o(1)$. Using lemma 1 and the fact that $\widehat{c}_{j,b}\to_p c_{j,b}$, we have $\widehat{W}^R_{j,b}\cdot\widehat{c}_{j,b}\to_d W_{j,b}\cdot c_{j,b}$. as $\mathcal{B}\to\infty$. By assumption 5, $F_{W_{j.b}|\widetilde{X}_j}$ is a continuous distribution, and by a trivial transformation of variable $F_{W'_{j.b}|\widetilde{X}_j}$ is also continuous, using lemma 2.11 of van der Vaart (2011), as $N_0\to\infty$, $\sup_{w\in\Theta}\left|F_{\widehat{\widetilde{W}}_{j,b}\left|\widetilde{X}_j\right.}(w)-F_{W'_{j.b}|\widetilde{X}_j}(w)\right|\to 0$.

Note $\widehat{c}_{j,b}$ is an estimator for the square-root of a ratio of variances, and $\widehat{c}_{j,b}\in\Lambda\subset(0,\infty)$. In addition, the set of functions $\left\{\widehat{W}_{j,b}\cdot\widehat{c}_{j,b}-w\middle|(y,w)\in\mathcal{Y}\text{ x }\Theta\right\}$ is contained in a finite-dimensional vector space, and the sequence of functions $1\left\{\widehat{W}^R_{j,b}\cdot\widehat{c}_{j,b}-w<0\right\}$ in $b$ is contained in the class $\mathcal{H}$ defined in the previous theorem. By the definition of a P-Glivenko-Cantelli class of functions, as $\mathcal{B}\to\infty$.,

$$\begin{aligned}&\sup_{w\in\Theta}\left|\widehat{F}_{j|\widetilde{X}_j}(w)-F_{\widehat{\widetilde{W}}_{j,b}\left|\widetilde{X}_j\right.}(w)\right| \\ &=\sup_{w\in\Theta}\left|\frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}}1\left\{\widehat{W}^R_{j,b}\cdot\widehat{c}_{j,b}<w\right\}-\Pr\left[\widehat{W}^R_{j,b}\cdot\widehat{c}_{j,b}<w\right]\right| \\ &=o_p(1)\end{aligned}$$

By the same argument as in the previous theorem, we can show that $N_0\to\infty$ and $\mathcal{B}\to\infty$.,

$$\Pr\left[d^*_{1-\frac{\alpha}{2}}\leq\widehat{\alpha}-\alpha_0\leq d^*_{\frac{\alpha}{2}}\middle|\alpha_0,\tilde{X}\right]\to_p 1-\alpha \tag{20}$$

■

For the leading example in the paper, the conditional variance of $W_j$ on $X_j$ only depends on $M(j,t)$ and it is given by:

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

where $A$ and $\widetilde{B}$ are constants, and $h(M(j,t))\equiv\frac{1}{(T-t^*)^2}\sum_{t=t^*+1}^{T}\frac{1}{M(j,t)}+\frac{1}{(t^*)^2}\sum_{t=1}^{t^*}\frac{1}{M(j,t)}$. For simplicity, we work with the case in which $M(j,t)=M_j$. In this case, the variance simplifies to $Var\left[W_j\middle|M_j\right]=A+\frac{B}{M_j}$ for a constant $B$. We assume that the distribution of $M_j$ does not vary with $N_0$.

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

$B$ to construct a consistent estimator for the $Var[W_j|M_j]$,

$$\widehat{Var[W_j|M_j]} = \widehat{A} + \frac{\widehat{B}}{M_j} \tag{21}$$

We use these two estimators to estimate the ratio $\widehat{c}_{jb} \equiv \sqrt{\frac{\widehat{Var[W_j|M_j]}}{\widehat{Var[W_{j,b}|M_j]}}}$. Lemma 4 shows that is $\widehat{c}_{jb}$ is a consistent estimator for $\sqrt{\frac{Var[W_1|M_1]}{Var[W_j|M_j]}}$.

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

**Proof.** Under these assumptions, we have that:

$$\mathbb{E}\left[(W_j)^2 \middle| M_j\right] = A + \frac{B}{M_j} \tag{22}$$

From Lemma 1, $(\widehat{W}_j^R)^2 - (W_j)^2 = o_p(1)$, which implies that OLS of $(\widehat{W}_j^R)^2$ on a constant and $\frac{1}{M_j}$ yield consistent estimators for $A$ and $B$. ■

## A.2 Extension: Two or more treated periods

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

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

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

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

By the Frisch-Waugh-Lovell theorem we know that

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

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

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

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

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

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

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

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

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

When $N_0 \to \infty$:

$$\begin{aligned} \widetilde{d}_{jt}^2 &= \left( d_{jt} - \frac{1}{T}(T - t_k^*) \right)^2 \\ &= d_{jt}^2 + \frac{1}{T^2}(T - t_k^*)^2 - 2\frac{d_{jt}}{T}(T - t_k^*) \end{aligned} \tag{23}$$

Therefore:

$$\begin{aligned} \sum_{t=1}^{T} \widetilde{d}_{jt}^2 &= t_k^* \left( \frac{1}{T^2}(T - t_k^*)^2 \right) + (T - t_k^*) \left( 1 + \frac{1}{T^2}(T - t_k^*)^2 - 2\frac{1}{T}(T - t_k^*) \right) \\ &= \frac{1}{T^2} \left[ t_k^*(T - t_k^*)^2 + (T - t_k^*)(t_k^*)^2 \right] = \frac{t_k^*(T - t_k^*)}{T} \end{aligned} \tag{24}$$

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

$$\sum_{j=1}^{N} \sum_{t=1}^{T} \widetilde{d}_{jt}^2 = \frac{1}{T} \sum_{k=1}^{K} N_k \left[ t_k^*(T - t_k^*) \right] \tag{25}$$

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

$$\begin{aligned} \widetilde{d}_{jt} &= \frac{1}{NT} \sum_{k=1}^{K} (T - t_k^*) N_k - \frac{1}{N} \sum_{k=1}^{K} 1[t > t_k^*] \times N_k \\ &= \frac{1}{N} \sum_{k=1}^{K} N_k \left[ \frac{(T - t_k^*)}{T} - 1[t > t_k^*] \right] \end{aligned} \tag{26}$$

$$\sum_t \tilde{d}_{jt}Y_{jt} = \frac{1}{N}\left(N_1\frac{(T-t_1^*)}{T} + \ldots + N_K\frac{(T-t_K^*)}{T}\right)\sum_{t\leq t_1^*} Y_{jt} +$$
$$\frac{1}{N}\left(N_1\frac{(-t_1^*)}{T} + N_2\frac{(T-t_2^*)}{T} + \ldots + N_K\frac{(T-t_K^*)}{T}\right)\sum_{t_1^*<t\leq t_2^*} Y_{jt} +$$
$$\frac{1}{N}\left(N_1\frac{(-t_1^*)}{T} + N_2\frac{(-t_2^*)}{T} + N_3\frac{(T-t_3^*)}{T} + \ldots + N_K\frac{(T-t_K^*)}{T}\right)\sum_{t_2^*<t\leq t_3^*} Y_{jt} +$$
$$\vdots$$
$$\frac{1}{N}\left(N_1\frac{(-t_1^*)}{T} + N_2\frac{(-t_2^*)}{T} + \ldots + N_K\frac{(-t_K^*)}{T}\right)\sum_{t>t_K^*} Y_{jt}$$
$$= -\frac{N_1}{N}\left(\frac{t_1^*}{T}\sum_{t>t_1^*} Y_{jt} - \frac{(T-t_1^*)}{T}\sum_{t\leq t_1^*} Y_{jt}\right) - \ldots - \frac{N_K}{N}\left(\frac{t_K^*}{T}\sum_{t>t_K^*} Y_{jt} - \frac{(T-t_K^*)}{T}\sum_{t\leq t_K^*} Y_{jt}\right) \quad (27)$$

As $N_0 \to \infty$:

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

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

$$\widetilde{d}_{jt} = d_{jt} - \frac{1}{T}(T - t_k^*) \quad (28)$$

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

$$\sum_t \tilde{d}_{jt}Y_{jt} = \frac{t_k^*}{T}\sum_{t>t_k^*} Y_{jt} - \frac{(T-t_k^*)}{T}\sum_{t\leq t_k^*} Y_{jt}$$

Therefore,

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

Note that

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

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

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

and use the same hypothesis test stated in the previous section. If the variance of $W_j$ is unknown, then we can estimate this variance using theorem 4.

## A.3 Extension: Model with Covariates

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

$$Y_{jt} = \alpha d_{jt} + \theta_j + \beta Z_{jt} + \gamma_t + \eta_{jt} \tag{29}$$

where $Z_{jt}$ is a $k$ x1 vector of covariates.

First, we can eliminate the fixed effects by a transformation of the model,

$$\widetilde{Y}_{jt} = \alpha \widetilde{d}_{jt} + \beta \widetilde{Z}_{jt} + \widetilde{\eta}_{jt} \tag{30}$$

where for a generic variable $H_{jt}$, $\widetilde{H}_{jt} = H_{jt} - \overline{H}_j - \overline{H}_t + \overline{H}$, with $\overline{H}_j = \frac{1}{T}\sum_{t=1}^{T} H_{jt}$, $\overline{H}_t = \frac{1}{N}\sum_{j=1}^{N} H_{jt}$ and $\overline{H} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{j=1}^{N} H_{jt}$. Recall that $N = N_0 + N_1$.

From the normal equations of OLS, we can show that:

$$\widehat{\alpha} = \alpha_0 + \frac{\sum_j \sum_t \widetilde{d}_{jt}\widetilde{\eta}_{jt}}{\sum_j \sum_t \widetilde{d}_{jt}^2} + \left(\beta - \widehat{\beta}\right)\left[\frac{\sum_j \sum_t \widetilde{d}_{jt}\widetilde{Z}_{jt}}{\sum_j \sum_t \widetilde{d}_{jt}^2}\right] \tag{31}$$

and,

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

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

$$\begin{aligned}
\nabla^k Y_j &= \frac{1}{T-t_k^*}\sum_{t>t_k^*}\left(\widetilde{Y}_{jt} - \widehat{\beta}\widetilde{X}_{jt}\right) - \frac{1}{t_k^*}\sum_{t\le t_k^*}\left(\widetilde{Y}_{jt} - \widehat{\beta}\widetilde{X}_{jt}\right) \\
&= \alpha_0 + \left(\widehat{\beta} - \beta\right)\left(\frac{1}{T-t_k^*}\sum_{t>t_k^*}\widetilde{X}_{jt} - \frac{1}{t_k^*}\sum_{t\le t_k^*}\widetilde{X}_{jt}\right) \\
&+ \frac{1}{T-t_k^*}\sum_{t>t_k^*}\widetilde{\eta}_{jt} - \frac{1}{t_k^*}\sum_{t\le t_k^*}\widetilde{\eta}_{jt}
\end{aligned}$$

Under the assumption of conditional exogeneity of $\eta_{jt}$ ($E[\eta_{jt}|X_{j1},...,X_{jT}] = 0$), by proposition 1 in Conley and Taber (2011), $\widehat{\beta}\xrightarrow{p}\beta$.[47] Therefore:

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

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

## A.4 Sampling weights

We consider here the extension of our method to the case with individual-level data with sampling weights. Consider the model:

$$Y_{ijt} = \alpha d_{jt} + \theta_j + \gamma_t + \nu_{jt} + \epsilon_{ijt} \tag{32}$$

where each observation has a sampling weight $\omega_{ijt}$. There are $M(j,t)$ individual observations for each $j \times t$ cell. We assume that $j = 1$ is treated and $j = 2,...,N$ is control.

When $N_0 \to \infty$, the DID estimator using the individual-level data with sampling weights converges in

[47] Note that this assumptions implies assumption 4 in the main text.

distribution to:

$$\hat{\alpha} \rightarrow_d \sum_{t=t^*}^{T} \frac{P_{1t}}{\sum_{t'=t^*}^{T} P_{1t'}} \tilde{\eta}_{1t} - \sum_{t=1}^{t^*} \frac{P_{1t}}{\sum_{t'=1}^{*} P_{1t'}} \tilde{\eta}_{1t} \tag{33}$$

where $P_{jt} = \sum_{i=1}^{M(j,t)} \omega_{ijt}$ and $\tilde{\eta}_{jt} = \sum_{i=1}^{M(j,t)} \frac{\omega_{ijt}}{P_{jt}} \eta_{ijt}$.

Now define the linear combination of the $\eta_{jt}$ errors with sampling weights:

$$W_j^s = \sum_{t=t^*}^{T} \frac{P_{1t}}{\sum_{t'=t^*}^{T} P_{1t'}} \tilde{\eta}_{jt} - \sum_{t=1}^{t^*} \frac{P_{1t}}{\sum_{t'=1}^{*} P_{1t'}} \tilde{\eta}_{jt} \tag{34}$$

It is important to note that $W_j^s$ is a linear combination of the errors of group $j$ with coefficients based on the number of observations and weights of the treated group. Assuming, for simplicity, that $\epsilon_{ijt}$ is i.i.d. while allowing for unrestricted serial correlation in $\nu_{jt}$, we have that:

$$\begin{aligned} var(W_j^s | M(j,t), \{\omega_{ijt}\}) &= var\left( \sum_{t=t^*}^{T} \frac{P_{1t}}{\sum_{t'=t^*}^{T} P_{1t'}} \nu_{jt} - \sum_{t=1}^{t^*} \frac{P_{1t}}{\sum_{t'=1}^{*} P_{1t'}} \nu_{jt} \right) + \\ &+ \left[ \sum_{t=t^*}^{T} \left( \frac{P_{1t}}{\sum_{t'=t^*}^{T} P_{1t'}} \right)^2 \left( \sum_{i=1}^{M(j,t)} \left( \frac{\omega_{ijt}}{P_{jt}} \right)^2 \right) + \sum_{t=1}^{t^*} \left( \frac{P_{1t}}{\sum_{t'=1}^{t^*} P_{1t'}} \right)^2 \left( \sum_{i=1}^{M(j,t)} \left( \frac{\omega_{ijt}}{P_{jt}} \right)^2 \right) \right] \sigma_\epsilon^2 \\ &= A + B q_j \end{aligned} \tag{35}$$

where $q_j$ is a function of the number of observations in all periods and the set of sampling weights for all individuals from group $j$.

This formula is very intuitive. It shows that the variance of the linear combination $W_j^s$ will be lower when the contribution of each observation $i$ to the weighted average $\tilde{\eta}_{jt}$ is lower. This will happen when $M(j,t)$ is large and weights are not concentrated on a few individuals. This formula will still apply under alternative assumptions on the errors, for example, with a balanced panel of individuals $i$ with serially correlated $\epsilon_{ijt}$, or with more complex intra-group correlation structures such as described in footnote 23.

## A.5 FGLS Estimation

In this section, we prove the main results of section 3. Under the assumption that the errors are uncorrelated across $j$, the variance/covariance matrix of f $\eta_{jt}$ is block diagonal with $T \times T$. We define $\Omega_j$ as one of the blocks. Under assumption 2, $\Omega_j \equiv \Omega(\tilde{X}_j)$. Note that $\Omega\left(\widetilde{X}_j\right)$ is a symmetric matrix that allows for heteroskedasticity and serial correlation. Suppose that we know $\Omega\left(\widetilde{X}_j\right)$. In this case, we can appply the standard GLS theory to the problem. The linear model is:

$$y_j = \alpha d_j + \theta_j + \gamma_t + \eta_j$$

where $y_j$ is a $T$x1 vector with $\{y_{j,t}\}_{t=1}^T$, $d_j$ is a $T$x1 vector with $\{d_{j,t}\}_{t=1}^T$ and $\eta_j$ is a $T$x1 vector with $\{\eta_{j,t}\}_{t=1}^T$. The GLS estimator will be based on the following model:

$$\Omega\left(\widetilde{X}_j\right)^{-\frac{1}{2}} y_j = \alpha \Omega\left(\widetilde{X}_j\right)^{-\frac{1}{2}} d_j + \Omega\left(\widetilde{X}_j\right)^{-\frac{1}{2}} \theta_j + \Omega\left(\widetilde{X}_j\right)^{-\frac{1}{2}} \gamma_t + \Omega\left(\widetilde{X}_j\right)^{-\frac{1}{2}} \eta_j$$

Note that the GLS estimator is the Gauss-Markov estimator in this problem. In other words, it has the lowest variance among all the linear unbiased estimator. The GLS is in the class of linear unbiased estimator that can be represented by:[48]

[48] We show that the restrictions imposed on the $a_{jt}$ are necessary and sufficient for unbiasedness of the linear estimators.

$$\Omega(\widetilde{\alpha}) = \left\{ \widetilde{\alpha}(a) = \sum_{t=1}^{T}\sum_{j=1}^{N} a_{tj}y_{jt} : \sum_{t=t^*+1}^{T} a_{jt} = 1, \sum_{t=1}^{T} a_{jt} = 0 \text{ for all } j \text{ and } \sum_{j=1}^{N} a_{jt} = 0 \text{ for all } t \right\}$$

Note that the variance of any estimator in this class is:

$$Var\left[\widetilde{\alpha}(a) | D, X\right] = \sum_{j=1}^{N} a_j' \Omega\left(\widetilde{X}_j\right) a_j$$

where $a_j = \begin{bmatrix} a_{1j} \\ \vdots \\ a_{Tj} \end{bmatrix}$. The GLS estimator is the estimator in the class $\Omega(\widetilde{\alpha})$ that has the minimum variance. In other words, it is the estimator that solves the following optimization problem:

$$\min_{a_1,..,a_N} \sum_{j=1}^{N} a_j' \Omega\left(\widetilde{X}_j\right) a_j \text{ s.t. } \sum_{t=t^*+1}^{T} a_{1t} = 1, \sum_{t=1}^{T} a_{jt} = 0 \text{ for all } j, \sum_{j=1}^{N} a_{jt} = 0 \text{ for all } t \tag{36}$$

Let $\widetilde{a}(N_0) = (\widetilde{a}_1(N_0), ..., \widetilde{a}_N(N_0))$ be the unique solution of the minimization problem above.

**Lemma 5** *Define $a_1^*$ as the unique solution for following the minimization problem:*

$$\min_{a_1} a_1' \Omega\left(\widetilde{X}_1\right) a_1 \ \ s.t. \sum_{t=t^*+1}^{T} a_{t1} = 1, \sum_{t=1}^{t^*} a_{t1} = -1$$

*As $N_0 \to \infty$,*

$$\widetilde{a}_1(N_0) \xrightarrow{p} a_1^* \text{ and } \widetilde{a}_j(N_0) \xrightarrow{p} 0 \text{ for } j = 2, ..., N$$

**Proof.** By contradiction, suppose that $\widetilde{a}_1(N_0)$ does not converge in probability to $a_1^*$. If this is the case, $\exists$ $N_0$ sufficient large such that $|\widetilde{a}_1(N_0) - a_1^*| > \varepsilon$ for $\varepsilon > 0$, and for a sufficient large $N_0$ we have that

$$\widetilde{a}_1(N_0)' \Omega\left(\widetilde{X}_1\right) \widetilde{a}_1(N_0) - a_1^{*\prime} \Omega\left(\widetilde{X}_1\right) a_1^* \geq c_2$$

for some positive constant $c_2$. One possible solution for the minimization problem 36 is:

$$a_1 = a_1^* \text{ and } a_j = -\frac{a_1^*}{N-1}$$

This solution satisfies all the constraints in the minimization problem, including the restriction that $a_{1t} =$

---

Consider any linear estimator $\widetilde{\alpha}(a) = \sum_{t=1}^{T}\sum_{j=1}^{N} a_{jt}y_{jt}$:

$$\mathbb{E}\left[\widetilde{\alpha}(a) | D, X\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{N} a_{jt}y_{jt} \middle| D, X\right] = \alpha\left(\sum_{t=1}^{T}\sum_{j=1}^{N} a_{jt}d_{jt}\right) + \sum_{j=1}^{N}\theta_j\sum_{t=1}^{T} a_{jt} + \sum_{t=1}^{T}\gamma_t\sum_{j=1}^{N} a_{jt} + \sum_{t=1}^{T}\sum_{j=1}^{N} a_{jt}\mathbb{E}\left[\eta_{jt} | X, D\right]$$

Assuming that $\mathbb{E}\left[\eta_{jt} | X, D\right] = 0$, in order to have an unbiased estimator, we need:

$$\sum_{t=1}^{T}\sum_{j=1}^{N} a_{jt}d_{jt} = \sum_{t=t^*+1}^{T} a_{1t} = 1, \ \sum_{t=1}^{T} a_{jt} = 0 \text{ for all } j, \ \sum_{j=1}^{N} a_{jt} = 0 \text{ for all } t$$

$-\sum_{j=2}^{N} a_{jt}$ for all $t$. Therefore, we have an upper bound for the optimization function when $N_0 \to \infty$,

$$\widetilde{a}_1(N_0)' \Omega\left(\widetilde{X}_1\right) \widetilde{a}_1(N_0) \leq$$

$$\widetilde{a}_1(N_0)' \Omega\left(\widetilde{X}_1\right) \widetilde{a}_1(N_0) + \sum_{j=2}^{N} a_j'(N_0) \Omega\left(\widetilde{X}_j\right) a_j(N_0)$$

$$\leq a_1^{*\prime} \Omega\left(\widetilde{X}_1\right) a_1^* + \frac{1}{(N-1)^2} \sum_{j=2}^{N} a_1^{*\prime} \Omega\left(\widetilde{X}_j\right) a_1^*$$

So by the expression above,

$$\widetilde{a}_1(N_0)' \Omega\left(\widetilde{X}_1\right) \widetilde{a}_1(N_0) - a_1^{*\prime} \Omega\left(\widetilde{X}_1\right) a_1^* \leq \frac{1}{(N-1)^2} \sum_{j=2}^{N} a_1^{*\prime} \Omega\left(\widetilde{X}_j\right) a_1^* = o_p(1)$$

However, as we showed before if $\widetilde{a}_1(N_0)$ does not converge in probability to $a_1^*$, $\widetilde{a}_1(N_0)' \Omega\left(\widetilde{X}_1\right) \widetilde{a}_1(N_0) - a_1^{*\prime} \Omega\left(\widetilde{X}_1\right) a_1^* \geq c_2$ for some positive constant $c_2$. So, by contradiction,

$$\widetilde{a}_1 \to_p a_1^*$$

Using this result, the value of the objective function that we are minimizing when $N_0 \to \infty$ is:

$$a_1^{*\prime} \Omega\left(\widetilde{X}_1\right) a_1^* + \sum_{j=2}^{N} a_j'(N_0) \Omega\left(\widetilde{X}_j\right) a_j(N_0) \leq a_1^{*\prime} \Omega\left(\widetilde{X}_1\right) a_1^* + o_p(1)$$

so the best choice of $a_j(N_0)$ is one that converges in probability to zero as $N_0 \to \infty$. ■

By Lemma 5, we have that

$$\widehat{\alpha}_{GLS} - \alpha_0 \to_d \sum_{t=1}^{T} a_{1t}^* \eta_{1t}$$

**Assumption 6(Consistent Variance Matrix):** There is an estimator $\widehat{\Omega}\left(\widetilde{X}_j\right)$ that converges in probability to a positive semidefinite matrix, called $\overline{\Omega}\left(\widetilde{X}_j\right)$.

Now, we work a FGLS estimator using $\widehat{\Omega}\left(\widetilde{X}_j\right)$ as an estimator for $\Omega(\tilde{X}_j)$. The FGLS estimator will converge to

$$\widehat{\alpha}_{FGLS} - \alpha_0 \to_d \sum_{t=1}^{T} \overline{a}_{1t}^* \eta_{1t}$$

where $\overline{a}_{1t}^*$ depends on the elements of the matrix $\overline{\Omega}\left(\widetilde{X}_1\right)$.

We combine our method with FGLS by using $W_j^* = \sum_{t=1}^{T} \overline{a}_{1t}^* \eta_{jt}$. In order to apply our method, we need to impose assumptions 1-5 on the behavior of $W_j^*$:

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

**2*** $W_j^* \left| X_j, d_j \right. \stackrel{d}{=} W_j^* \left| \widetilde{X}_j \right.$.

**3*** $W_j^* \left| \widetilde{X}_j \right.$ has the same distribution across $\widetilde{X}_j$ up to a scale parameter.

**4*** $\mathbb{E}\left[W_j^* \middle| X_j, d_j\right] = \mathbb{E}\left[W_j^* \middle| X_j\right] = 0$.

**5*** $F_{W_j^*|\widetilde{X}_j}$ is continuous.

These assumptions can be implied by primitive conditions on $\eta_{jt}$,for $j = 1, ..., N$ and $t = 1, ..., T$.
In our leading example,

$$Var\left[W_j^* \middle| X_j, d_j\right] = Var\left[W_j^* \middle| M_j\right] = A^* + \frac{B^*}{M_j}$$

where $A^*$ and $B^*$ will depend on $\{\overline{a}_{1t}^*\}_{t=1}^T$ and the variance-covariance matrix of $\eta_{jt}$.

We apply our method to $W_j^* = \sum_{t=1}^{T} \overline{a}_{1t}^* \eta_{jt}$. For each control unit, we construct $\widehat{W}_j^{*R} = \sum_{t=1}^{T} \widehat{a}_{1t}^* \widehat{\eta}_{jt}$ using the estimator of the variance-covariance matrix, $\widehat{\Omega}\left(\widetilde{X}_j\right)$. Then, we generate a bootstrap sample of $\left\{\widehat{W}_{j,b}^{*R}\right\}$. As in the main procedure, we generate the normalized $\widehat{W}_{j,b}^{*R} = \frac{\widehat{W}_{j,b}^{*R}}{\sqrt{Var\left[W_j^* \middle| \widetilde{X}_j\right]}}$, and then we generate a bootstrap sample using the re-scaled residuals, $\widehat{\widehat{W}}_{j,b}^{*} = \widehat{W}_{j,b}^{*R} \cdot \sqrt{Var\left[W_1^* \middle| \widetilde{X}_j\right]}$. This procedure generates bootstrap estimators, $\widehat{\alpha}_{FGLS,b} - \alpha_0 = \widehat{\widehat{W}}_{j,b}^{*}$.

We reject $H_0$ at level $\alpha$ if and only if $\widehat{\alpha}_{FGLS} - \alpha_0 < (\widehat{\alpha}_{FGLS,b} - \alpha_0)\left[\frac{\alpha}{2}\right]$ or $\widehat{\alpha}_{FGLS} - \alpha_0 > (\widehat{\alpha}_{FGLS,b} - \alpha_0)\left[1 - \frac{\alpha}{2}\right]$, where $(\widehat{\alpha}_{FGLS,b} - \alpha_0)[q]$ denotes the $q$th quantile of the distribution of $\{(\widehat{\alpha}_{FGLS,1} - \alpha_0), ..., (\widehat{\alpha}_{FGLS,B} - \alpha_0)\}$.

**Theorem 6** *Define $c_{1-\frac{\alpha}{2}}$ and $c_{\frac{\alpha}{2}}$ as the $\left(1 - \frac{\alpha}{2}\right)$th and $\frac{\alpha}{2}$th quantile of the empirical distribution of $\widehat{\alpha}_{FGLS,b} - \alpha_0$ given $X$, for $b = 1, ..., \mathcal{B}$. Assuming that we know the variance of $W_j^* | \widetilde{X}_j$, under assumptions 1*-5*, as as $N_0 \to \infty$ and $\mathcal{B} \to \infty$.,*

$$\Pr\left[c_{1-\frac{\alpha}{2}} \leq \widehat{\alpha}_{FGLS,b} - \alpha_0 \leq c_{\frac{\alpha}{2}}\right] \to_p 1 - \alpha$$

**Proof.** The proof is very similar to the proof of Theorem 2, so we skip some steps. First, we need to show that $\Gamma_b^*(w) = \Pr\left[\widetilde{W}_{j,b}^* < w \middle| \widetilde{X}_j, b\right]$ converges in probability to $\Gamma^*(w) = \Pr\left[W_1^* < w | \widetilde{X}_1\right]$. Under assumption 1*,

$$\Gamma_b^*(w) = \int 1\{W_1^* < w\} \cdot d\widehat{F}_{j|\widetilde{X}_j}^*\left(W_j^*\right)$$

$$\Gamma^*(w) = \int 1\{W_1^* < w\} \cdot dF_{1|\widetilde{X}_1}^*\left(W_1^*\right)$$

In this case, we estimate the empirical CDF $\widehat{F}_{j|\widetilde{X}_j}^*(.)$ using the re-scaled residuals, $\widetilde{W}_{j,b}^* = \widehat{W}_{j,b}^{*R} \cdot \frac{\sqrt{Var\left[W_1^* | \widetilde{X}_1\right]}}{\sqrt{Var\left[W_j^* | \widetilde{X}_j\right]}}$,

$$\widehat{F}_{j|\widetilde{X}_j}^*(w) = \frac{1}{\mathcal{B}} \sum_{b=1}^{\mathcal{B}} 1\left\{\widetilde{W}_{j,b}^* < w\right\}$$

Define $F_{\widetilde{W}_{j,b}^*|\widetilde{X}_j}(w) = \Pr\left[\widetilde{W}_{j,b}^* < w\right]$ and $F_{W_{j,b}^*|\widetilde{X}_j}(w) = \Pr\left[W_{j,b}^* < w\right]$, with $W_{j,b}^* = W_{j,b}^* \cdot \frac{\sqrt{Var\left[W_1^* | \widetilde{X}_1\right]}}{\sqrt{Var\left[W_j^* | \widetilde{X}_j\right]}}$, and note that:

$$\begin{aligned}\sup_{w \in \Theta} \left|\widehat{F}_{j|\widetilde{X}_j}^*(w) - F_{1^*|\widetilde{X}_1}(w)\right| \leq & \sup_{w \in \Theta} \left|\widehat{F}_{j|\widetilde{X}_j}^*(w) - F_{\widetilde{W}_{j,b}^*|\widetilde{X}_j}(w)\right| + \sup_{w \in \Theta} \left|F_{\widetilde{W}_{j,b}^*|\widetilde{X}_j}(w) - F_{W_{j,b}^*|\widetilde{X}_j}(w)\right| \\ & + \sup_{w \in \Theta} \left|F_{W_{j,b}^*|\widetilde{X}_j}(w) - F_{1|\widetilde{X}_1}^*(w)\right|\end{aligned}$$

Note that th sequence of functions $1\left\{\widetilde{W}^*_{j,b} < w\right\}$ in $b$ belong to the class of functions $\mathcal{H}$ defined in Theorem 2. By the definition of P-Glivenko-Cantelli class of functions,

$$\sup_{w\in\Theta}\left|\widehat{F}^*_{j|\widetilde{X}_j}(w) - F_{\widetilde{W}^*_{j,b}|\widetilde{X}_j}(w)\right| = o_p(1)$$

Based on the results of lemma 1, we know that:

$$\widehat{\eta}_{jt} \to_p \left(\eta_{jt} - \overline{\eta}_j\right)$$

when $N_0 \to \infty$.

By the Slutsly Theorem and assumption 5, we can write:

$$\begin{aligned}\sum_{t=1}^{T}\widehat{a}^*_{1t}\widehat{\eta}_{jt} &= \sum_{t=1}^{T}\overline{a}^*_{1t}\left(\eta_{jt} - \overline{\eta}_j\right) + o_p(1)\\ &= \sum_{t=1}^{T}\overline{a}^*_{1t}\eta_{jt} - \overline{\eta}_j\sum_{t=1}^{T}\overline{a}^*_{1t} + o_p(1)\\ &= \sum_{t=1}^{T}\overline{a}^*_{1t}\eta_{jt} + o_p(1)\end{aligned}$$

when $N_0 \to \infty$ and by the restriction that $\sum_{t=1}^{T}\overline{a}^*_{1t} = 0$.

So at the end, $\sum_{t=1}^{T}\widehat{a}^*_{1t}\widehat{\eta}_{jt} \Rightarrow \sum_{t=1}^{T}\overline{a}^*_{1t}\eta_{jt}$. Under assumption 5 and by lemma 2.11 of Vaan deer Vaart (1998),

$$\sup_{w\in\Theta}\left|F_{\widetilde{W}^*_{j,b}|\widetilde{X}_j}(w) - F_{W^*_{j,b}|\widetilde{X}_j}(w)\right|$$

By assumption 2 and 3, $\sup_{w\in\Theta}\left|F_{W^*_{j,b}|\widetilde{X}_j}(w) - F^*_{1|\widetilde{X}_1}(w)\right| = o_p(1)$.

At the end, $\Pr\left[c_{1-\frac{\alpha}{2}} \leq \widehat{\alpha}_{FGLS} - \alpha_0 \leq c_{\frac{\alpha}{2}}\right] \to o_p 1-\alpha$. ■

It is important to note that in this section we are dealing with the case of one treated group and many control groups. However, we can extend all the results to the case of few treated groups (more than one).

Now, we compare the power of of GLS test with the inference procedure that combines our main methodology with FGLS. First, notice that in our case, the inference procedure is respect to one dimensional parameter, $\alpha$, and it well know from the literature that uniformly most powerful test (UMP) will exist for one sided alternative. In addition, since the GLS is the Gauss-Markov estimator, under normality the t-test based on the GLS estimator with known variance-covariance matrix is the UMP (or UMPU) test (Hausman and Kuersteiner, 2004). In order to show the power result, we need to add the normality assumption.

**Assumption 7 (Normality):** $\eta_j|\, d_j$ follows a normal distribution with mean zero and variance $\Omega\left(\widetilde{X}_j\right)$.

**Theorem 7** *Under assumptions 1-7, if $\bar{\Omega}\left(\widetilde{X}_j\right) = \Omega\left(\widetilde{X}_j\right)$, then the power of the inference procedure based on the bootstrap of the FGLS converge to the power of UMP test against one sided fixed alternative.*

**Proof.** Without lost of generality, we are going to look to the following one-sided test:

$$H_0 : \alpha \leq \alpha_0 \text{ and } H_1 : \alpha > \alpha_0$$

Since we are considering, for simplicity, the case with only one treated group, then, under assumptions 1 and 6:

$$\widehat{\alpha}_{GLS} - \alpha_0 \to_d N\left(0, (\sigma^*_1)^2\right)$$

where $\sigma^*_1 = \sqrt{var(W^*_1|X_1)}$.

For a fixed alternative, the power of a t-test based on the unfeasible GLS is:

$$\beta^{*}(\alpha) = \Pr\left[\left(\frac{\widehat{\alpha}_{GLS}-\alpha_0}{\sigma_1^{*}}\right) > z_{1-\alpha}\right] = \Pr\left[\left(\frac{\widehat{\alpha}_{GLS}-\alpha}{\sigma_1^{*}}\right) > z_{1-\alpha} - \frac{\alpha-\alpha_0}{\sigma_1^{*}}\right]$$

Notice that, under assumptions 1-7, if $\bar{\Omega}\left(\widetilde{X}_j\right) = \Omega\left(\widetilde{X}_j\right)$, then:

$$\widehat{\alpha}_{FGLS} - \alpha_0 \rightarrow_d N\left(0, (\sigma_1^{*})^2\right)$$

Let $c_{1-\alpha} = (\widehat{\alpha}_{FGLS,b} - \alpha_0)[1-\alpha]$. Then, under fixed alternatives, the power of the procedure that combines FGLS with bootstrap is,

$$\begin{aligned}\beta(\alpha) &= \Pr\left[\widehat{\alpha}_{FGLS} - \alpha > c_{1-\alpha} - (\alpha-\alpha_0)\right] \\ &= \Pr\left[\frac{\widehat{\alpha}_{FGLS}-\alpha}{\sigma_1^{*}} > \frac{c_{1-\alpha}-(\alpha-\alpha_0)}{\sigma_1^{*}}\right]\end{aligned}$$

Since $\widehat{\alpha}_{FGLS} - \widehat{\alpha}_{GLS} = o_p(1)$, then it suffices to show that $\frac{c_{1-\alpha}}{\sigma_1^{*}} \rightarrow_p z_{1-\alpha}$ even under the alternative to show that $\beta(\alpha) \rightarrow_p \beta^{*}(\alpha)$.

We argue first that, under the alternative, $\sup_{w_j\in\Theta}\left|\widehat{F}^{*}_{j|\widetilde{X}_j}(w_j) - F^{*}_{1|\widetilde{X}_1}(w_1)\right| = o_p(1)$. Notice that $\widehat{W}_j^{*R} \rightarrow_d W_j^{*}$ for all $j$ in the control group, even if the null is false. In addition, by assumption 5, $F_{W_j^{*}}$ is continuous. Now notice that, with a small and fixed number of treated groups, the probability of resampling a treated group converges to zero as $N_0 \rightarrow \infty$. Therefore, our Theorem 6 would apply even under the alternative. Finally, since, under normality, the quantiles of $\widehat{\alpha}_{FGLS}$ are well defined, then we can apply the continuity theorem to show that $\frac{c_{1-\alpha}}{\sigma_1^{*}}$ converges in probability to $z_{1-\alpha}$.

■

## A.6 Simulations with the method proposed in the Appendix of CT

In their online appendix, CT propose a method to deal with heteroskedasticity generated by variation in group sizes. They assume a model in which the group x time error, $\eta_{jt}$, is the sum of two independent Gaussian processes, $\mu_{jt}$ and $v_{jt}$, where the first captures dependence and the second captures heteroskedasticity. A simple parametrization mentioned in CT is given by $cov(\mu_{jt}, \mu_{ks}) = \theta_1 e^{(-\theta_2 dist_{jk} - \theta_3|t-s|)}$, which allows for both correlation across $t$ and across $j$. Since we focus on the case with no dependence across $j$, we assume $\theta_2 = 0$. They assume the component $v_{jt}$ independent across group and time with a variance that is a function $g(M(j,t), \theta_v)$ of the number of observations in cell $(j,t)$ that depends on the parameter $\theta_v$. They do not suggest a functional form for $g()$ in the paper, so we consider a parametrization similar to the one we use, given by $g(M(j,t)) = A + \frac{B}{M(j,t)}$.

As explained in CT, the residuals of the DID regression converge in probability to $\eta_{jt} - \bar{\eta}_j$, and we can use the covariances and variances of the residuals as moment conditions. With the proposed parametrization, we have that:

$$\begin{aligned}var(\eta_{jt}-\bar{\eta}_j) &= \theta_1 + \frac{1}{T^2}\sum_{t'}\sum_{s'}\theta_1 e^{-\theta_3|t'-s'|} - \frac{2}{T}\sum_{s'}\theta_1 e^{-\theta_3|t-s'|} + \frac{T-2}{T}\left(A + \frac{B}{M(j,t)}\right) + \\ &\quad + \frac{1}{T^2}\sum_{s'}\left(A + \frac{B}{M(j,s')}\right) \\ cov(\eta_{jt}-\bar{\eta}_j, \eta_{js}-\bar{\eta}_j) &= \theta_1 e^{-\theta_3|t-s|} + \frac{1}{T^2}\sum_{t'}\sum_{s'}\theta_1 e^{-\theta_3|t'-s'|} - \frac{1}{T}\sum_{s'}\theta_1 e^{-\theta_3|t-s'|} - \frac{1}{T}\sum_{s'}\theta_1 e^{-\theta_3|s-s'|} \\ &\quad - \frac{1}{T}\left(A + \frac{B}{M(j,t)}\right) - \frac{1}{T}\left(A + \frac{B}{M(j,s)}\right) + \frac{1}{T^2}\sum_{s'}\left(A + \frac{B}{M(j,s')}\right)\end{aligned}$$

They suggest estimating the parameters by GMM and then use that to directly provide an estimator for the distribution of $W_1$. We consider first a MC simulation with a correctly specified model with $\theta_1 = 1$,

$\theta_3 = 1$, $A = 0$, and $B = 100$. We set $T = 8$, $M \in [50, 200]$. We set $N = 400$, with $j = 1$ treated, so we have a large number of control groups to estimate the parameters using GMM. Based on 5,000 simulations, we find an average rejection rate of 5.2% with virtually no difference when we consider large versus small groups as treated. Therefore, as expected, the method suggested in the online appendix of CT works well when the model is correctly specified when the number of groups is large.

A potential problem with the specification suggested in CT is that it only allows for serial correlation in the common shock $\mu_{jt}$. This assumption may be implausible if the dataset has some painel structure at the individual level, as in the CPS. We consider then an alternative DGP in which the individual-level shocks exhibit serial correlation. We assume that the covariance of the error for an individuals at times $t$ and $s$ is given by $e^{-|t-s|}$, which implies that, in the group x time aggregate model, we have that $cov(\eta_{jt} - \bar{\eta}_j, \eta_{js} - \bar{\eta}_j) = \frac{1}{M_j} e^{-|t-s|}$. In this case, the covariances across time will diminish with the number of observations, which is not allowed in the example model presented in CT. Based again on 5,000 MC simulations, the test over-rejects when the treated group is small (9.5% when it is in the lowest decile of the distribution of $M$) and under-rejects when the treated group is large (2% when it is in the largest decile of the distribution of $M$). This happens because the estimated distribution used in the test proposed in CT would underestimate the serial correlation when $M$ is small and overestimate it when $M$ is large. It should be possible to derive an alternative inference method based on CT under a different set of assumptions. However, this would require derivation of a new set moment equations for the GMM, which is not straightforward given the incidental parameter problem (see Hansen (2007)), which may prevent applied researchers from using this method.

In order to show that misspecification of the GMM model in the CT method may be relevant in real applications, we also consider simulations using the CPS, as we do in Section 7.2. We consider again the case with $T = 8$. In Section 7.2, we show that inference based on robust OLS standard errors over-rejects under the null due to serial correlation induced by the panel structure of the CPS. Using the method proposed in the appendix of CT with log wages as outcome variable, we find an average rejection rate of 5.2%. However, rejection rates are 0.034 percentage points lower when the number of observations of the treated group is above median relative to when it is below median (the p-value of a test that rejection rates are the same for these two groups is 0.078). When we consider employment as outcome variable, then rejection rate is 3.8% lower when the number of observations of the treated group is above median (the p-value of a test that rejection rates are the same for these two groups is 0.003). These results suggest that the structure of the errors in applications with real datasets may exhibit more complex structures than assumed in the example model suggested in the appendix of CT.

## A.7 Appendix Tables

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

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

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

Table A.2: **Rejection Rates in MC Simulations with $N_0 + 1 = 25$**

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

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

Table A.3: **Rejection Rates in MC Simulations with alternative distributions for $\nu_{jt}$**

| | $N = 100$ | | | | $N = 25$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Bootstrap w/o correction | | Bootstrap with correction | | Bootstrap w/o correction | | Bootstrap with correction | |
| $\rho$ | Mean (1) | Relative size distortion (2) | Mean (3) | Relative size distortion (4) | Mean (5) | Relative size distortion (6) | Mean (7) | Relative size distortion (8) |
| | | | | Panel A: $\nu_{jt} \sim \chi^2(1)/\sqrt{2}$ | | | | |
| 0.01% | 0.051 | 0.035 | 0.053 | 0.002 | 0.051 | 0.029 | 0.055 | 0.005 |
| 1% | 0.052 | 0.013 | 0.052 | 0.003 | 0.052 | 0.012 | 0.056 | 0.007 |
| 4% | 0.051 | 0.004 | 0.051 | 0.003 | 0.055 | 0.003 | 0.056 | 0.007 |
| | | | | Panel B: $\nu_{jt} \sim t(3)/\sqrt{3}$ | | | | |
| 0.01% | 0.051 | 0.034 | 0.051 | 0.002 | 0.052 | 0.029 | 0.055 | 0.005 |
| 1% | 0.051 | 0.015 | 0.052 | 0.001 | 0.053 | 0.014 | 0.055 | 0.006 |
| 4% | 0.051 | 0.005 | 0.051 | 0.002 | 0.053 | 0.005 | 0.055 | 0.005 |
| | | | | Panel C: $\nu_{jt} = 2$ with prob. 0.2 and $= -0.5$ with prob. 0.8 | | | | |
| 0.01% | 0.050 | 0.033 | 0.052 | 0.003 | 0.051 | 0.030 | 0.056 | 0.005 |
| 1% | 0.050 | 0.021 | 0.051 | 0.002 | 0.052 | 0.017 | 0.055 | 0.005 |
| 4% | 0.050 | 0.017 | 0.050 | 0.005 | 0.050 | 0.010 | 0.051 | 0.003 |

Notes: This table replicates the Monte Carlo simulation results presented in Table 1 with alternative distributions for $\nu_{jt}$. All simulations consider $M \in [50, 200]$.