Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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

Bruno Ferman[†] Cristine Pinto[‡]

Sao Paulo School of Economics - FGV

Abstract

We derive an inference method that works in Differences-in-Differences settings with few treated and many control groups in the presence of heteroskedasticity. As a leading example, we provide theoretical justification and empirical evidence that heteroskedasticity generated by variation in group sizes can invalidate existing inference methods, even in datasets with a large number of observations per group. In contrast, our inference method remains valid in this case. Our test can also be combined with feasible generalized least squares, providing a safeguard against misspecification of the serial correlation.

Keywords: clustering; bootstrap; permutation tests; Behrens-Fisher problem

JEL Codes: C12; C21; C33

[*] We thank Josh Angrist, Colin Cameron, 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 numerous seminar participants 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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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 by Bertrand et al. (2004). Still, there is as yet no unified approach to dealing with this problem. As stated by 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. (2017), Canay et al. (2017), Ibragimov and Müller (2010), Ibragimov and Müller (2016), and MacKinnon and Webb (2015)). However, all 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. There are alternative inference methods

[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). Bertrand et al. (2004) show that CRVE at the group level works well when there are many treated and many control groups, while Wooldridge (2003) shows that CRVE provides asymptotically valid inference when the number of groups increases. See Abadie et al. (2017) for a recent discussion on the use of CRVE.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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. 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.[3]

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. 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 structures on the errors.

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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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 the pre-post difference in average errors.[4] 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. Under this assumption, we can re-scale the pre-post difference in average residuals of the control groups using the (estimated) heteroskedasticity structure, so that they become informative about the distribution of the pre-post difference in average errors of the treated groups. By focusing on this 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. We show that a cluster residual bootstrap with this heteroskedasticity correction provides asymptotically valid hypothesis testing when the number of control groups goes to infinity, even when there is only one treated group. Our Monte Carlo (MC) 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). 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.[5]

[4] The crucial assumption for our method is that, conditional on a set of covariates, the distribution of this 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.

[5] In their online appendix and in an earlier version of their paper (Conley and Taber (2005)), CT present alternative methods that allow for heteroskedasticity depending on group sizes. However, these methods impose strong assumptions on the structure of the errors. See details in Section 2.3.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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. (2017). While all these papers rely on strong assumptions on the structure of the errors — including homoskedasticity — to derive the FGLS estimator, Hansen (2007), and Brewer et al. (2017) follow a recommendation by Wooldridge (2003), and combine FGLS estimation with CRVE. 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, the serial correlation is misspecified, 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 the pre-post difference in average 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.

An alternative estimation method for the case with few treated groups is the synthetic control (SC) estimator, proposed by 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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

test statistic the ratio of post-/pre-treatment mean squared prediction error (MSPE). If the variance of transitory 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, when the number of pre-treatment periods is large, as shown by Ferman and Pinto (2017). However, Ferman and Pinto (2017) also 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 with a single pre-intervention period, 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 to calculate the exact distribution of the test statistic under a sharp null hypothesis (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.[6] 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.[7] 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

[6] 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.

[7] 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), Efron and Hinkley (1978)and Yates (1984)

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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.

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.[8] 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. 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. 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.

The remainder of this paper proceeds as follows. In Section 2 we present our base model. We briefly explain the necessary assumptions in the existing inference methods, and explain why heteroskedasticity usually invalidates inference methods designed to deal with the case of few treated groups. Then we derive an alternative inference method that corrects for heteroskedasticity even when there is only one treated group. In Section 3 we 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. We perform MC simulations to examine the performance of existing inference methods and to compare that to the performance of our method with

[8] 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 can be considered as natural extensions of inference procedures designed to the Behrens-Fisher problem.

heteroskedasticity correction in Section 5, while we compare the different inference methods by simulating placebo laws in real datasets in Section 6. We conclude in Section 7.

2 Base Model

2.1 A Review of Existing Methods

Consider first a group x time DID aggregate model given by

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

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

There are N_1 treated groups and N_0 control groups. We start assuming that d_{jt} changes to 1 for all treated groups starting after date t^*, and we consider the pre-post difference in average errors for each group j, which is given by

$$W_j = \frac{1}{T - t^*} \sum_{t=t^*+1}^{T} \eta_{jt} - \frac{1}{t^*} \sum_{t=1}^{t^*} \eta_{jt}. \tag{2}$$

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

In this simpler case, the DID estimator is given by

$$\begin{aligned}
\hat{\alpha} &= \frac{1}{N_1}\sum_{j=1}^{N_1}\left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} Y_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} Y_{jt}\right] - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} Y_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} Y_{jt}\right] \qquad (3)\\
&= \alpha + \frac{1}{N_1}\sum_{j=1}^{N_1} W_j - \frac{1}{N_0}\sum_{j=N_1+1}^{N} W_j .
\end{aligned}$$

The variance of the DID estimator, under the assumption that η_{jt} are independent across j, is given by[9]

$$var(\hat{\alpha}) = \left[\frac{1}{N_1}\right]^2 \sum_{j=1}^{N_1} var(W_j) + \left[\frac{1}{N_0}\right]^2 \sum_{j=N_1+1}^{N} var(W_j). \tag{4}$$

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

When there are many treated and many control groups, Bertrand et al. (2004) suggest that CRVE at the group level works well, as this method allows for unrestricted intra-group and serial correlation in the residuals η_{jt}. The CRVE has a very intuitive formula in the DID framework which, up to a degrees-of-freedom correction, is given by

$$\widehat{var(\hat{\alpha})}_{\text{Cluster}} = \left[\frac{1}{N_1}\right]^2 \sum_{j=1}^{N_1} \widehat{W}_j^2 + \left[\frac{1}{N_0}\right]^2 \sum_{j=N_1+1}^{N} \widehat{W}_j^2, \tag{5}$$

where $\widehat{W}_j = (T-t^*)^{-1}\sum_{t=t^*+1}^{T}\hat{\eta}_{jt} - (t^*)^{-1}\sum_{t=1}^{t^*}\hat{\eta}_{jt}$, and $\hat{\eta}_{jt}$ are the residuals from the DID regression.

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

[9] 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.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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 5 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 have $\widehat{W}_1 = 0$ *by construction*. Therefore, the variance of the DID estimator would be severely underestimated (as noticed by Mackinnon and Webb (2016)). 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 does 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. However, in order to allow for heteroskedasticity, one would have to use heteroskedasticity-robust standard errors, so this 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, as noticed by 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 also 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 the pre-post difference in average residuals, which takes into account any form of serial correlation, instead of using the group x time level residuals. In the simpler case with only one treated group, $\hat{\alpha} - \alpha$ would converge to W_1 when $N_0 \rightarrow \infty$. In this case, they recommend using $\{\widehat{W}_j\}_{j=2}^{N_0+1}$ 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 $N_0 \rightarrow \infty$. Finally, cluster residual bootstrap methods resample the residuals while holding the regressors constant throughout the pseudo-samples. The residuals are resampled at the group level, so that the correlation structure is preserved. It is possible, however, that a treated group receives the residuals of a control group, so a crucial assumption is again that errors are homoskedastic.

A potential problem with these methods, as originally explained by 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 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. 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.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

2.2 Leading Example: Variation in Group Sizes

In this section, we formalize the idea that variation in the number of observations per group inherently leads to heteroskedasticity in the group x time DID aggregate models, and we derive the implications of this heteroskedasticity for inference methods that rely on homoskedasticity. We start with a simple individual-level DID model,

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

where Y_{ijt} represents the outcome of individual i in group j at time t; ν_{jt} is a group x time error term (possibly correlated over time), and ϵ_{ijt} is an individual-level error term. The main 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 ϵ_{ijt} are all uncorrelated, while allowing for unrestricted auto-correlation in ν_{jt}. Later we consider more complex structures. Let $M(j,t)$ be the number of observations in group j at time t.

When we aggregate by group x time, this model becomes the same as the one in equation 1, that is,

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

where $\eta_{jt} = \nu_{jt} + (M(j,t))^{-1} \sum_{i=1}^{M(j,t)} \epsilon_{ijt}$.

Under the simplifying assumption that ϵ_{ijt} is i.i.d. across i, j, and t, and assuming for simplicity that $M(j,t) = M_j$ is constant across t, we have that

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

for constants A and B, regardless of the auto-correlation of ν_{jt}, where, in this simplified case,

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

$A = var((T-t^*)^{-1}\sum_{t=t^*+1}^{T}\nu_{jt} - (t^*)^{-1}\sum_{t=1}^{t^*}\nu_{jt})$ and $B = ((T-t^*)^{-1} + (t^*)^{-1})var(\epsilon_{ijt})$. Therefore, the group x time aggregate model is heteroskedastic, unless M_j is constant across j. See Suplemental Appendix A1 for the case in which the $M(j,t)$ is not constant over time.

For a much wider range of structures on the errors, the conditional variance of W_j given M_j will have the same parametric formula given in equation 8, which depends on only two parameters. For example, if we had a panel and allow for the individual-level errors to be correlated across time, then we would have another term that would depend on the ϵ_{ijt} auto-correlation parameters divided by the number of observations, so we would still end up with the same formula, $var(W_j|M_j) = A + B/M_j$. This formula may also remain valid 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

$$Y_{ikjt} = \alpha d_{jt} + \theta_j + \gamma_t + \nu_{jt} + \omega_{kjt} + \epsilon_{ikjt}, \tag{8}$$

for individual i in subgroup k, group j and time t, where we allow for a common subgroup shock ω_{kjt} in addition to the group-level shock ν_{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 $var(W_j|M_j) = A + B/M_j$.

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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

aggregate data.

If $A > 0$, 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 ν_{jt} is i.i.d. across j, then $var(W_j|M_j)/var(W_{j'}|M_{j'}) \to 1$ 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 by 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 (CT, Proposition 4).

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 $var(W_j|M_j)/var(W_{j'}|M_{j'})$ remain constant if all M_j grow at the same rate, which implies that the aggregate model remains heteroskedastic even asymptotically. Therefore, the distortions in rejection rates we described above would remain even when there is a large number of individual observations per group. We might have $A \approx 0$ under complex (and plausible) conditions on the structure of the errors. For example, we would have $A \approx 0$ in model 6 if ϵ_{ijt} is serially correlated and $var(\nu_{jt}) \approx 0$. Alternatively, in model 8 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 $A \approx 0$. Therefore, one should be careful when applying methods such as those proposed by 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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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. However, we do not restrict to this particular example. Under this assumption, we can re-scale the residuals of the control groups using the (estimated) structure of the heteroskedasticity, so that we can use this information to estimate the distribution of the error for the treated groups.

More formally, assume we have a total of N groups where the first $j = 1, ..., N_1$ groups are treated. 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, and d_j be an indicator variable equal to 1 if group j is treated.[10] There is no restriction on whether covariates X_j enter or not in model 1. We define our assumptions directly on $W_j = (T - t^*)^{-1} \sum_{t=t^*+1}^{T} \eta_{jt} - (t^*)^{-1} \sum_{t=1}^{t^*} \eta_{jt}$. 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. $W_j | X_j, d_j \overset{d}{=} W_j | X_j$,
3. $W_j | X_j \sim \sigma(X_j) \times Z_j$, where Z_j is i.i.d. across j, and $\sigma(X_j)$ is 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.

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. Assumption 2 implies that, conditional on a subset of observable covariates,

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

the distribution of W_j will be the same independently of 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. Assumption 3 implies that the distribution of $W_j|X_j$ only depends on X_j through the variance parameter. The random variable Z_j is not necessarily normally distributed. This assumption reduces the dimensionality of the problem. It might be possible to relax this assumption and estimate the conditional distribution of $W_j|X_j$ 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, which should provide significant improvement relative to existing inference methods. Finally, Assumption 4 is the standard identification assumption for DID, while Assumption 5 is a necessary condition for consistency of the bootstrap method.

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 $N_0 \rightarrow \infty$, 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 $\{\widehat{W}_{j,b}^R\}_{i=1}^N$, and then calculate our bootstrap estimates as $\hat{\alpha}_b = {N_1}^{-1}\sum_{j=1}^{N_1}\widehat{W}_{j,b}^R - {N_0}^{-1}\sum_{j=N_1+1}^{N}\widehat{W}_{j,b}^R$. We do not need to work with the group x time residuals $\hat{\eta}_{jt}$ to construct our bootstrap estimates, because the DID estimator can be constructed only with information on $\widehat{W}_j^R$.

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 Supplemental Appendix A1, we show that, if we know the variance of W_j conditional on X_j, then we can, for each b, re-scale the bootstrap draw $\{\widehat{W}_{j,b}^R\}_{j=1}^N$ using instead $\{\widetilde{W}_{j,b}^R\}_{j=1}^N$, where $\widetilde{W}_{j,b}^R = \widehat{W}_{j,b}^R\sqrt{var(W_j|X_j)/var(W_{j,b}|X_{j,b})}$. In this case, under Assumptions 2 and 3, $\widetilde{W}_{j,b}^R$ has (asymptotically) the same distribution as W_j. As a result, this procedure generates bootstrap estimators $\hat{\alpha}_b = {N_1}^{-1}\sum_{j=1}^{N_1}\widetilde{W}_{j,b} -$

${N_0}^{-1}\sum_{j=N_1+1}^{N}\widetilde{W}_{j,b}$ that can be used to draw inferences about α with the correct size. We only need to know the variance of W_j, so we do not need to specify the serial correlation structure of the errors η_{jt}.

The main problem, however, is that $var(W_j|X_j)$ is generally unknown, so it needs to be estimated. In Theorem 3 in Supplemental Appendix A1, 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 $\widetilde{\widehat{W}}_{j,b} = \widehat{W}^R_{j,b}\sqrt{\widehat{var(W_j|X_j)}/\widehat{var(W_{j,b}|X_{j,b})}}$. 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 $var(W_j|X_j, d_j) = var(W_j|M_j) = A + B/M_j$, for constants A and B, where X_j is the set of observable variables including M_j. We show in Lemma 4 in Supplemental Appendix A1 that we can get a consistent estimator for $var(W_j|M_j)$ by regressing $(\widehat{W}^R_j)^2$ on $1/M_j$ and a constant.[11] 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.

Summarizing, our bootstrap procedure, for this specific case in which heteroskedasticity is generated by variation in group sizes, consists of

[11] See Appendix A1 for the case in which the number of observations per group is not constant over time.

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 $1/M_j$, obtaining $\widehat{var(W_j|M_j)}$.

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

   (a) Resample with replacement N times from $\{\widehat{W}_j^R\}_{i=1}^N$ to obtain $\left\{\widehat{W}_{j,b}^R\right\}_{i=1}^N$.

   (b) Re-scale the bootstrap draw to obtain $\left\{\widetilde{\widehat{W}}_{j,b}\right\}_{i=1}^N$, where

   $$\widetilde{\widehat{W}}_{j,b} = \widehat{W}_{j,b}^R\sqrt{\widehat{var(W_j|X_j)}/\widehat{var(W_{j,b}|X_{j,b})}}.$$

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

5. 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 $\hat{\alpha}_1, ..., \hat{\alpha}_{\mathcal{B}}$.

Once we consider W_j, the DID estimator collapses to a comparison of means. Therefore, an alternative solution to this problem would be to consider a weighted least squares estimator for this comparison of means, where the weights are given by $[\widehat{var(W_j|X_j)}]^{-\frac{1}{2}}$. This is similar to a generalized least squares estimator, with the difference that we do not take into account the intra-group correlation structure (see Section 3 for a complete analysis of FGLS in this setting). An important caveat with this approach, however, is that we cannot guarantee that this weighted least squares estimator will be asymptotically normal if $N_0 \to \infty$ but N_1 is fixed. More specifically, this will only be the case if $W_j|X_j$ is normally distributed. Therefore, we prefer our bootstrap approach that does not rely on normality.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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 Supplemental Appendix A2 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 $N_k(T - t^*_k)t^*_k / (\sum_{k=1}^{K} N_k(T - t^*_k)t^*_k)$. 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 $\widehat{W}^{R,k}_j = (T - t^*_k)^{-1} \sum_{t=t^*_k+1}^{T} \hat{\eta}^R_{jt} - (t^*_k)^{-1} \sum_{t=1}^{t^*_k} \hat{\eta}^R_{jt}$. We generalize our method to this case by estimating K functions $\widehat{var(W^k_j|M_j)}$ by regressing $(\widehat{W}^{R,k}_j)^2$ on a constant and $1/M_j$. Each function $\widehat{var(W^k_j|M_j)}$ 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 Supplemental Appendix A3 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 aggregate the residuals of this regression at the group x time level. The other steps in the bootstrap procedure remain the same. Finally, we consider the case of individual-level data with sampling weights in Suplemental Appendix A4.

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 than we do in our method. 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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

one heteroskedasticity. This essentially excludes the possibility of serial correlation in the individual-level error, which should be relevant in panel datasets. In Supplemental Appendix A6, we show significant distortions using this method when we consider placebo simulations with the CPS. 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. In contrast to these two alternative methods, our method is valid under a wider range of assumptions on the structure of the errors.

3 Improving Efficiency with a FGLS

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 by 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 η_{jt} is block diagonal with $T \times T$ blocks given by Ω_j. We assume that $\Omega_j = \Omega(X_j)$. Let $\widehat{\Omega}(X_j)$ be an estimator of $\Omega(X_j)$ that converges to $\bar{\Omega}(X_j)$ (we allow $\bar{\Omega}(X_j) \neq \Omega(X_j)$, so $\widehat{\Omega}(X_j)$ may be inconsistent). The FGLS estimator using $\widehat{\Omega}(X_j)$ will be a linear estimator $\hat{\alpha}_{\text{FGLS}} = \sum_{t=1}^{T} \sum_{j=1}^{N} \hat{a}_{jt} Y_{jt}$. In Supplemental Appendix A5, we show that, when $N_1 = 1$,

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

$\hat{\alpha}_{\text{FGLS}} \xrightarrow{d} \sum_{t=1}^{T} \bar{a}_{1t}\eta_{1t}$ when $N_0 \to \infty$, where $\bar{\mathbf{a}}_1 = (\bar{a}_{11}, ..., \bar{a}_{1T})'$ is defined by

$$\begin{aligned} \bar{\mathbf{a}}_1 \quad = \quad & \underset{\mathbf{a}_1 \in \, \mathbb{R}^T}{\text{argmin}} \; \mathbf{a}_1' \bar{\Omega}(\tilde{X}_1)\mathbf{a}_1 \\ & \text{subject to: } \sum_{t=t^*+1}^{T} a_{1t} = 1 \text{ and } \sum_{t=1}^{T} a_{1t} = 0. \end{aligned} \tag{9}$$

Therefore, defining the linear combination $W_j^* = \sum_{t=1}^{T} \bar{a}_{1t}\eta_{jt}$, we show in Supplemental Appendix A5 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^*. For our leading example presented in Section 2.2, we would still have $var(W_j^*|M_j) = A + B/M_j$ for constants A and B.

So far, we only assumed that $\widehat{\Omega}(X_j)$ converges to $\bar{\Omega}(X_j)$ when $N_0 \to \infty$. So our inference method is valid even if $\bar{\Omega}(X_j) \neq \Omega(X_j)$. If η_{jt} is multivariate normal and $\bar{\Omega}(X_j) = \Omega(X_j)$, then we show in Supplemental Appendix A5 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. Therefore, 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.

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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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 $W_j|X_j, d_j \overset{d}{=} W_j|X_j$. Therefore, we can only relax this assumption if we impose some structure on the intra-group correlation.

We 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 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 $\tilde{Y}_t = {N_1}^{-1} \sum_{j=1}^{N_1} Y_{jt} - {N_0}^{-1} \sum_{j=N_1+1}^{N} Y_{jt}$, then

$$\tilde{Y}_t = \begin{cases} \tilde{\theta} + \tilde{\eta}_t, \text{ for } t = 1, ..., t^* \\ \alpha + \tilde{\theta} + \tilde{\eta}_t, \text{ for } t = t^* + 1, ..., T, \end{cases} \tag{10}$$

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

Therefore, this is a particular case of Andrews' (2003) end-of-sample instability test in a model that includes only a constant, where the idea is to test whether this constant remains stable before and after t^*. Since the difference in the constant before and after t^* is exactly α, we are essentially testing the null $\alpha = 0$.

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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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 Monte Carlo Evidence

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

$$Y_{ijt} = \nu_{jt} + \epsilon_{ijt}. \tag{11}$$

In our simulations, we estimate a DID model given by equation 6 where only $j = 1$ is treated and $T = 2$, and then we test the null hypothesis of $\alpha = 0$ using different hypothesis testing methods. We 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: ν_{jt} and ϵ_{ijt} are drawn from normal random variables. We hold constant the total variance $var(\nu_{jt} + \epsilon_{ijt}) = 1$, while changing $\rho = \sigma_\nu^2/(\sigma_\nu^2 + \sigma_\epsilon^2) \in \{.01\%, 1\%, 4\%\}$;

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

For each case, we simulated 100,000 estimates. We present rejection rates for inference using robust standard errors in the individual-level OLS regression, and for the cluster residual bootstrap with and without our heteroskedasticity correction. Rejection rates using DL and CT methods are similar to those using cluster residual bootstrap without heteroskedasticity correction (see Supplemental Appendix). 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 (2015) because their method collapses to CT when there is only one treated group.

5.1 Test Size

Panel A of Table 1 presents 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).

Figure 1.A presents 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 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. 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". These results are presented 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 $\rho = 0.01\%$. Rejection rates by decile of the treated group for cluster residual bootstrap without correction when $\rho = 1\%$ and when $\rho = 4\%$ are presented 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).

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. Figures 1.D to 1.F present rejection rates by decile of the treated group when the intra-group correlation is 0.01%, 1%, and 4%. 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.

Simulations with variations in the distribution of group sizes are presented in Supplemental Appendix Table A1. 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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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 A1 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 $N_0 \rightarrow \infty$. This assumption is important for two reasons. First, as in any other cluster bootstrap method, a small number of groups implies a small number of possible distinct pseudo-samples. In this case, the bootstrap distribution will not be smooth even with many bootstrap replications (Cameron et al. (2008)). 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 $\widehat{var(W_j|M_j)} = 1/M_j$, 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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

estimated with less precision.

Panel B of Table 1 and Figure 2 present 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%. 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.

5.2 Test Power

We have focused so far on Type I error. We saw in Section 5.1 that our method is effective 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 5.1, with the difference that we now add an effect of β 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 MC simulations, we can calculate the variance of $\hat{\alpha}$ given the parameters of the model and generate an (infeasible) t-statistic $t = \hat{\alpha}/\sigma_{\hat{\alpha}}$. Then we also calculate rejection rates based on this test statistic. With two periods and one treated group, with $N_0 \rightarrow \infty$, the DID OLS estimator is asymptotically equivalent to the GLS estimator where the full structure of the variance/covariance matrix is known. Therefore, since the errors in our DGP are normally distributed, we also know that a test based on this t-statistic is the uniformly most powerful test (UMP) for this particular case.

Figures 3.A to 3.C present rejection rates for different effect sizes and intra-group correlation parameters when 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), 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.

6 Simulations with Real Datasets

The results presented in Section 5 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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

observations, and (ii) whether our method works in real datasets. More specifically, our DGP in Section 5 implies that the *real* variance of W_j would have exactly the relationship $var(W_j|M_j) = A + B/M_j$, 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).[12]

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 $N_0 \rightarrow \infty$. 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 effective 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.[13] We extract information on employment status and earnings for women between ages 25 and 50, following Bertrand et al. (2004). There is substantial variation in the number of observations per group in these datasets. Considering the 2015 dataset, there are, on average, 505 observations in each PUMA in the ACS. This number, however, hides an important heterogeneity in cell sizes. The 10th percentile of PUMA cell sizes is 152, while the 90th percentile is 923. There is also substantial heterogeneity in state sizes in the ACS. While the average cell size is 9725, the 10th percentile

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

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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

is 1,290, while the 90th percentile is 18,913.[14] Finally, the state 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. See details in Appendix Table A3.

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. 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.[15] 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.

6.1 American Community Survey (ACS) Results

Panel A of Table 2 presents results from simulations using the PUMA-level treatments using the ACS. Column 1 shows 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. Column 3 of Table 2 presents results for the bootstrap without the

[14] 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.

[15] We include Washington, D.C.

heteroskedasticity correction. As in the MC 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. Column 4 of Table 2 presents the difference in rejection rates when the number of observations in the treated group is above and below the median.[16] 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. Columns 5 and 6 of Table 2 presents the rejection rates using bootstrap with our heteroskedasticity correction.[17] 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.

Panel B of Table 2 presents 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. Rejection rates using bootstrap with our heteroskedasticity correction are presented 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.

[16] 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 MC simulations.

[17] In all simulations using real data, we use the version of our method that allows for samplings weights, as described in Supplemental Appendix A4.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

6.2 Current Population Survey (CPS) Results

Simulation using the CPS are presented in Table 3. 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 is consistent with the findings in Bertrand et al. (2004). Rejection rates for the residual bootstrap without correction are presented 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. Rejection rates using bootstrap with our heteroskedasticity correction are presented 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 effective in correcting for heteroskedasticity in a scenario that serial correlation is important without the need to specify the structure of the serial correlation.

6.3 Power with Real Data Simulations

We saw in Sections 6.1 and 6.2 that our method provides tests with correct size in simulations with the ACS and the CPS. Figure 4 presents power results from simulations with these datasets.[18] 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

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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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

7 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. 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. 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.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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.

_ **, Susan Athey, Guido Imbens, and Jeffrey Wooldridge**, "When Should You Adjust Standard Errors for Clustering?," 2017.

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

Angrist, Joshua and Jorn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2009.

Behrens, Walter, "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," *Journal of Econometric Methods*, oct 2017, *7* (1).

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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

Canay, Ivan A., Joseph P. Romano, and Azeem M. Shaikh, “Randomization Tests Under an Approximate Symmetry Assumption,” *Econometrica*, 2017, *85* (3), 1013–1030.

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, David R., “Some Problems Connected with Statistical Inference,” *Ann. Math. Statist.*, 06 1958, *29* (2), 357–372.

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

_ **and David 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, Bradley and David V. Hinkley, “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,” April 2017.

Fisher, Ronald 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.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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

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

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.

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, Rich L. and Joseph 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.

Mackinnon, James G. and Matthew D. Webb, "Wild Bootstrap Inference for Wildly Different Cluster Sizes," *Journal of Applied Econometrics*, feb 2016, *32* (2), 233–254.

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

Pitman, Edwin. 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].," Technical Report, Minnesota Population Center, Minneapolis, MN 2015.

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

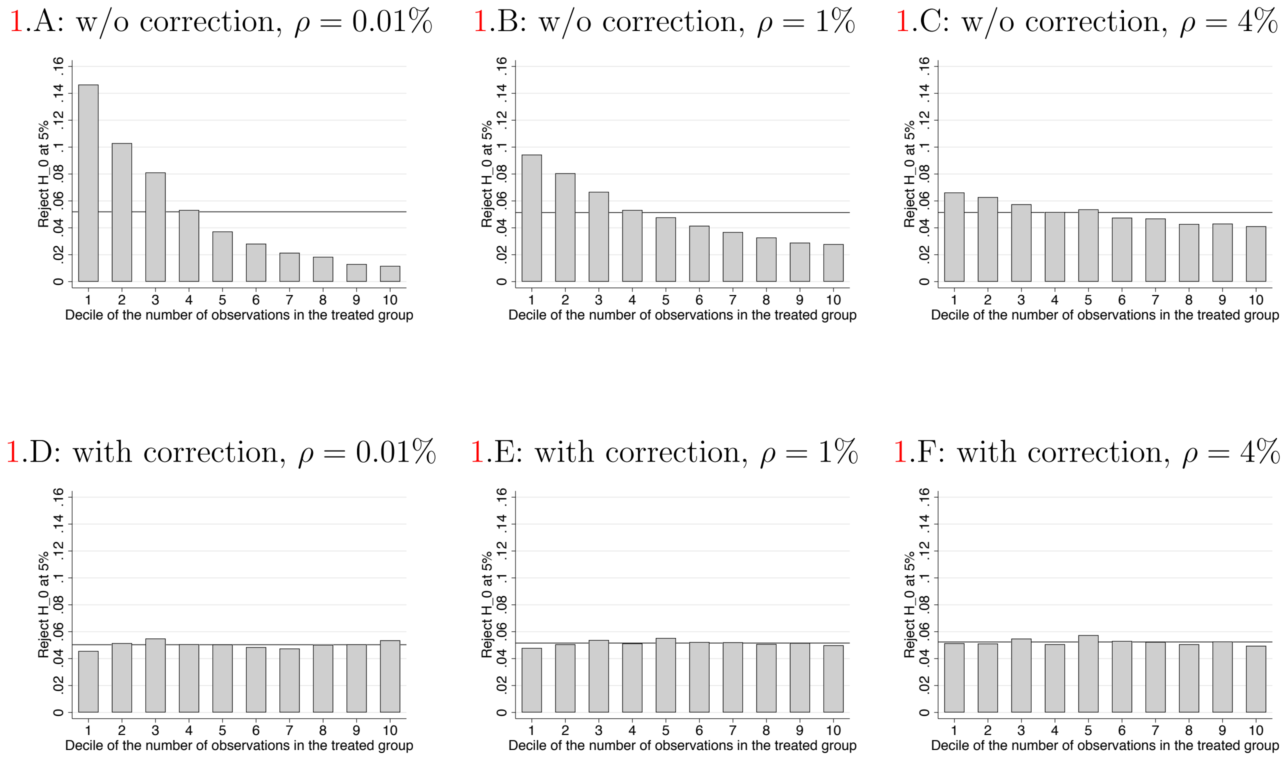
Wang, Ying Y., "Probabilities or the Type 1 errors of the Welch tests for the Behrens-Fisher problem," *Journal of the American Statistical Association.*, 1971, *66*, 605–608.

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

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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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

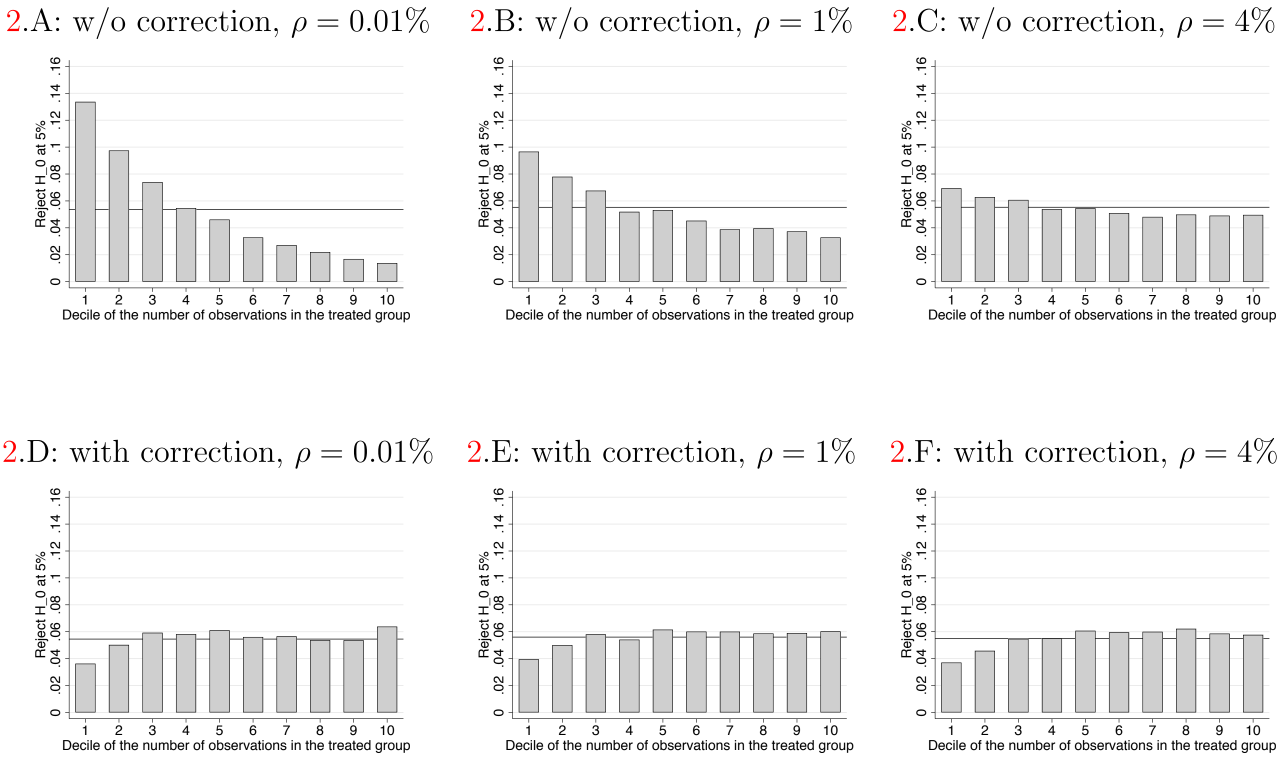

1.A: w/o correction, $\rho = 0.01\%$ 1.B: w/o correction, $\rho = 1\%$ 1.C: w/o correction, $\rho = 4\%$

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 based on the MC simulations explained in Section 5, for $N = 100$ and $M \in [50, 200]$. 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.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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



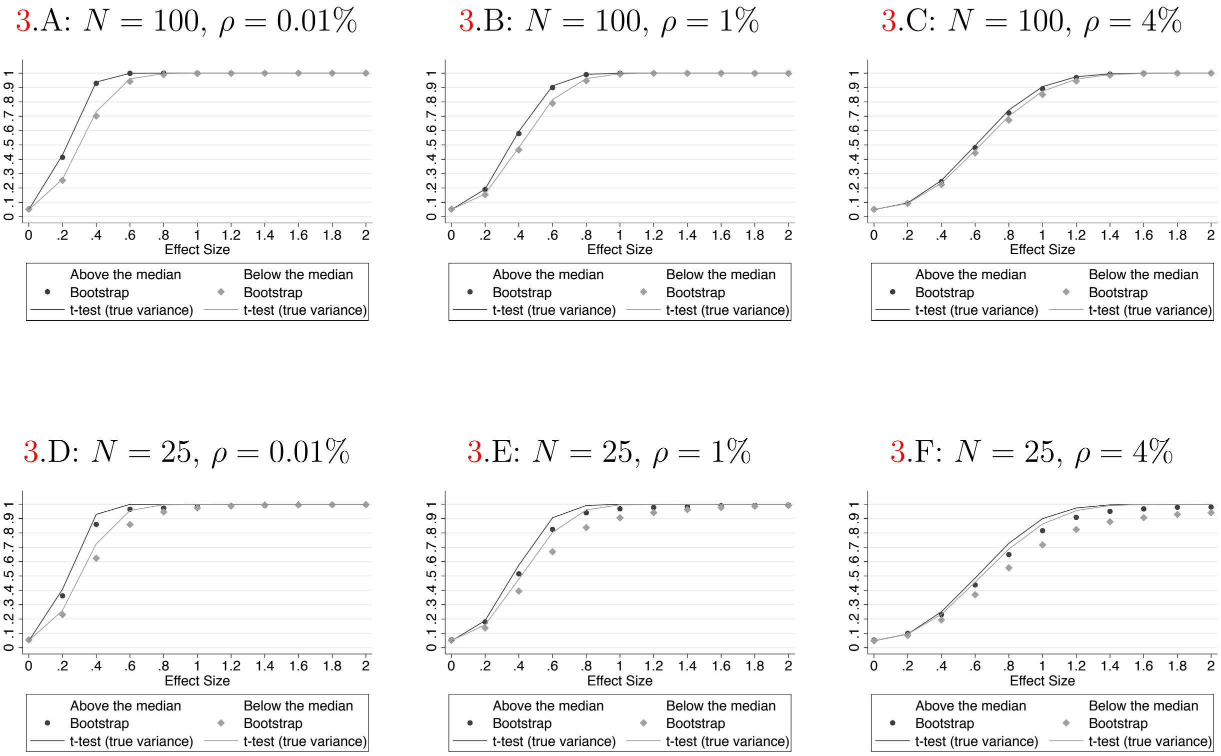
2.A: w/o correction, $\rho = 0.01\%$ 2.B: w/o correction, $\rho = 1\%$ 2.C: w/o correction, $\rho = 4\%$

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 based on the MC simulations explained in Section 5, for $N = 25$ and $M \in [50, 200]$. 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.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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


3.A: $N = 100,\ \rho = 0.01\%$

3.B: $N = 100,\ \rho = 1\%$

3.C: $N = 100,\ \rho = 4\%$

3.D: $N = 25,\ \rho = 0.01\%$

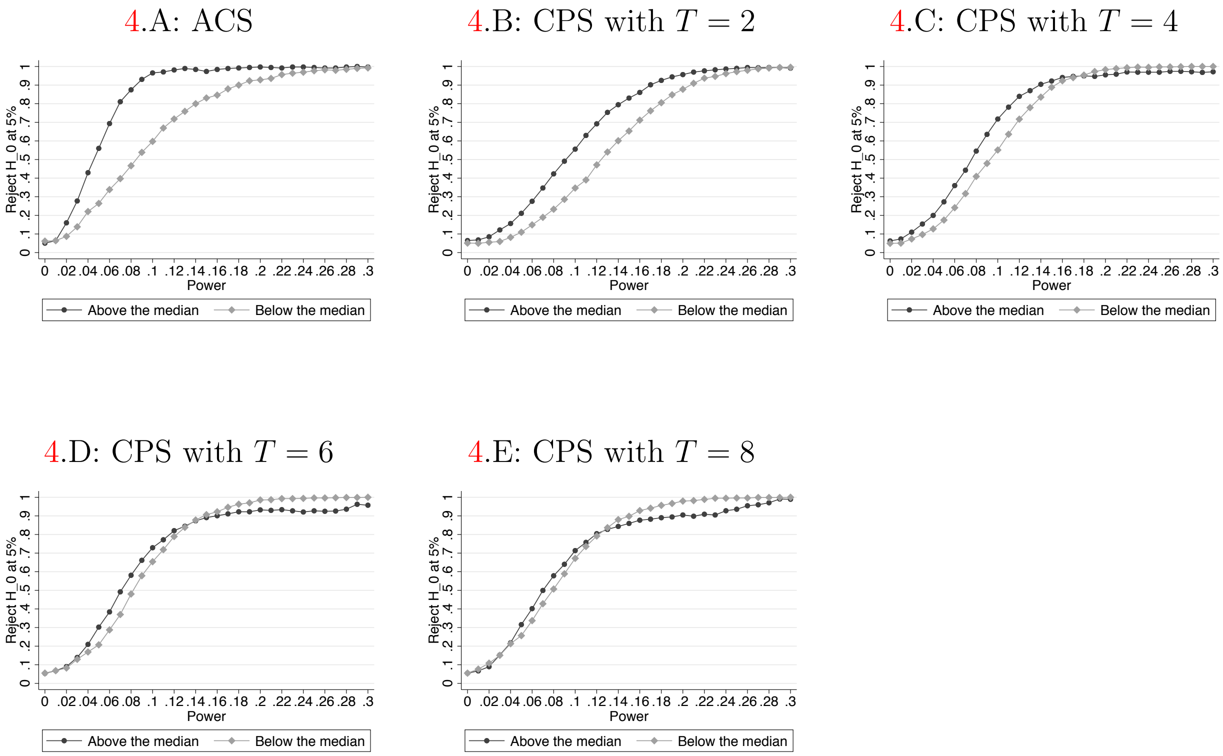
3.E: $N = 25,\ \rho = 1\%$

3.F: $N = 25,\ \rho = 4\%$

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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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



4.A: ACS

4.B: CPS with $T = 2$

4.C: CPS with $T = 4$

4.D: CPS with $T = 6$

4.E: CPS with $T = 8$

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

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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

	Robust OLS		Bootstrap w/o correction		Bootstrap with correction	
ρ	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	0.036	0.050	0.002
1%	0.193	0.032	0.051	0.018	0.052	0.001
4%	0.418	0.062	0.051	0.007	0.052	0.002
Panel B: $N = 25$						
0.01%	0.052	0.002	0.052	0.030	0.055	0.005
1%	0.193	0.032	0.054	0.016	0.056	0.005
4%	0.424	0.055	0.055	0.006	0.055	0.006

Notes: This table presents results from the MC simulations explained in Section 5, with 100 groups and $M \in [50, 200]$. In all simulations, only one group is treated. Each line presents simulation for different values of intra-group correlation. 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.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

Table 2: **Simulations with the ACS Survey**

	Inference Method					
			Bootstrap		Bootstrap	
	Robust OLS		w/o correction		with correction	
Outcome	Mean	Diff	Mean	Diff	Mean	Diff
Variable	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: ACS with PUMA level interventions					
Employment	0.068^{xxx}	0.002	0.051	-0.077***	0.050	-0.002
	(0.003)	(0.010)	(0.003)	(0.005)	(0.003)	(0.005)
Log(wages)	0.078^{xxx}	0.007	0.050	-0.084***	0.051	-0.001
	(0.003)	(0.012)	(0.003)	(0.005)	(0.003)	(0.006)
	Panel B: ACS with state level interventions					
Employment	0.052	0.007	0.050	-0.097***	0.048	0.005
	(0.008)	(0.016)	(0.014)	(0.023)	(0.008)	(0.016)
Log(wages)	0.071^{x}	-0.002	0.056	-0.110***	0.054	-0.011
	(0.011)	(0.021)	(0.019)	(0.034)	(0.010)	(0.020)

Notes: This table presents rejection rates for the simulations using ACS data, as explained in Section 6.1. We consider the following 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 (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). Standard errors clustered at the treated group level. * Significantly different from 0 at 10%, ** Significantly different from 0 at 5%, * Significantly different from 0 at 1%; ^{x} Significantly different from 0.05 at 10%, ^{xx} Significantly different from 0.05 at 5%, ^{x} Significantly different from 0.05 at 1%.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_00759

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

	Inference Method					
	Robust OLS		Bootstrap w/o correction		Bootstrap with correction	
Outcome	Mean	Diff	Mean	Diff	Mean	Diff
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 2 years						
Employment	0.046	-0.007	0.048	-0.045***	0.053	0.009
	(0.007)	(0.011)	(0.008)	(0.011)	(0.007)	(0.012)
Log(wages)	0.068^{xxx}	-0.001	0.048	-0.034***	0.055	0.015
	(0.006)	(0.013)	(0.006)	(0.012)	(0.006)	(0.012)
Panel B: 4 years						
Employment	0.063^{xx}	0.012	0.045	-0.035***	0.053	-0.004
	(0.007)	(0.012)	(0.007)	(0.013)	(0.006)	(0.013)
Log(wages)	0.100^{xxx}	0.033	0.057	-0.034**	0.057	0.014
	(0.011)	(0.021)	(0.008)	(0.015)	(0.007)	(0.016)
Panel C: 6 years						
Employment	0.087^{xxx}	-0.006	0.053	-0.045***	0.053	-0.016
	(0.008)	(0.017)	(0.007)	(0.013)	(0.006)	(0.013)
Log(wages)	0.141^{xxx}	0.053**	0.053	-0.045***	0.053	-0.004
	(0.014)	(0.027)	(0.009)	(0.015)	(0.009)	(0.016)
Panel D: 8 years						
Employment	0.132^{xxx}	0.022	0.053	-0.046***	0.048	-0.015
	(0.013)	(0.023)	(0.008)	(0.014)	(0.007)	(0.014)
Log(wages)	0.207^{xxx}	0.022	0.054	-0.041**	0.054	-0.007
	(0.015)	(0.033)	(0.010)	(0.019)	(0.010)	(0.019)

Notes: This table presents rejection rates for the simulations using CPS data, as explained in Section 6.2. We consider the following 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). Standard errors clustered at the treated group level. * Significantly different from 0 at 10%, ** Significantly different from 0 at 5%, * Significantly different from 0 at 1%; ^{x} Significantly different from 0.05 at 10%, ^{xx} Significantly different from 0.05 at 5%, ^{x} Significantly different from 0.05 at 1%.