A Hypothesis Test Robust to Cluster Heterogeneity

By Akiva Yonah Meiselman*

Draft: October 20, 2021 Click here for most recent version

I propose a hypothesis test for clustered samples. Hypothesis tests of linear models can be misleading when critical values are selected from a distribution that does not match the test statistic. I derive my test by inverting the distribution of the test statistic under a standard assumption about the errors. Previous tests can be severely misspecified in samples with few clusters, few ever-treated clusters, cluster size outliers, or treatment intensity outliers. I show that my proposed test is an exact test regardless of these features. Using simulations, I also demonstrate where this adjustment is most impactful in achieving exact tests compared to previous hypothesis tests.

Researchers often find that their samples include observations that are not independent. Rather, the observations are grouped into independent clusters. Common research designs based on differences-in-differences or otherwise based on fixed effects use the relationships between observations to control for unobservable characteristics. Statistical inference in such samples must account for clustering. Failing to do so can lead to dramatically inaccurate standard errors, confidence intervals, and p-values.

Since White (1984), there have been tools that allow researchers to perform cluster-robust inference. Those tools have asymptotic justifications. We can be sure that they are valid in samples that are effectively large, in the sense of their

* University of Texas at Austin, Bernard and Audre Rapoport Building 3.134G, 2225 Speedway, Austin, TX 78705, Email: meiselman@utexas.edu.

asymptotic behavior. However, it is not clear that those tools work when the effective sample size is small.

When the number of clusters is small, when the number of clusters with treatment variation is small, when there are cluster size outliers, and when there are treatment intensity outliers, the asymptotic behavior of cluster-robust test statistics cannot be relied upon. For example, consider a differences-in-differences analysis of a tax policy using a 30-year panel of US states. Suppose that data could only be collected for 9 states; or that only 4 states ever implemented the tax; or that a few states contained a large fraction of the affected population; or that one state implemented a much larger tax. In all of these cases, some adjustment to the conventional cluster-robust hypothesis test would be necessary. No previous test is robust to these features.

In this paper, I propose a hypothesis test for clustered samples with cluster-level fixed effects that rejects true hypotheses at the correct rate even in samples with few clusters, few clusters with treatment variation, cluster size outliers, and treatment intensity outliers. I focus on the case with cluster-level fixed effects because clustering and fixed effects are frequently paired together in applied research.

In Section III, I develop my test by inverting the distribution of the test statistic under an assumption about the errors often made in adjustments to cluster-robust inference. My test therefore has a finite-sample justification in addition to its asymptotic justification under the standard assumptions for consistent cluster-robust inference.

Previous studies have recommended that when the effective sample size is small, researchers should use hypothesis tests based on a bootstrap or based on approximating the distribution of the test statistic with a t-distribution. Both of these approaches have advantages. However, in Section IV I demonstrate using Monte Carlo simulations that they are all still vulnerable to samples with few clusters, few clusters with treatment variation, cluster size outliers, and treatment intensity outliers. My test does not share this vulnerability.

In Section I, I give context for my contribution to the literature on cluster-robust inference. In Section II, I describe the model with clustering and cluster-level fixed effects that is the setting for this paper. In Section III, I introduce my test and show that it is exact. In Section IV, I present evidence from Monte Carlo simulations that my test rejects true hypotheses at the correct rate even when other tests fail to do so. In Section V, I illustrate how my test functions in an empirical setting from Abouk and Adams (2013), and in Section VI, I conclude.

I. Literature

Much attention has been paid to cluster-robust inference in the applied literature since Bertrand, Duflo and Mullainathan (2004), but consistent cluster-robust variance estimators (CRVEs) were developed much earlier by White (1984), Liang and Zeger (1986), and Arellano (1987).

Many studies of clustering have focused on attaining asymptotically valid inference in clustered samples. White (1984) shows that, when clusters are equally sized and homogenous, the basic CRVE (henceforth $\hat{V}_{CR0}$) can consistently estimate the variance of the OLS estimator. Hansen (2007) relaxes the cluster homogeneity assumption, showing that equal-sized clusters alone allow the CRVE to be consistent and that it converges at a rate determined by the number of clusters G. He recommends critical values be drawn from a t-distribution with $G-1$ degrees of freedom.

In a recent paper that I draw on, Carter, Schnepel and Steigerwald (2017) show that $\hat{V}_{CR0}$ is consistent even when cluster sizes vary, but the rate of convergence is instead determined by G^*, what Carter, Schnepel and Steigerwald (2017) call the "effective number of clusters." They recommend calculating G^* as a diagnostic tool. If G^* is large, inference can rely upon the asymptotic properties of the test statistic to determine its behavior, so critical values can reasonably be drawn from the standard normal distribution $N(0,1)$. In Section III, I will show that my proposed test is asymptotically valid based on the same logic as in Carter,

Schnepel and Steigerwald (2017).

The main advantage of my test over other asymptotically-valid tests is that my test performs better that those tests when the asymptotic properties of the test statistic do not determine its behavior. This will be the case in samples with a small number of clusters, a small number of clusters with treatment variation, large cluster size outliers, or large treatment intensity outliers. This is the first paper to explicitly target exact inference in a finite, clustered sample. However, there have been other studies which make finite-sample adjustments to conventional cluster-robust hypothesis tests.

Bell and McCaffrey (2002) address finite-sample cluster-robust inference by developing two additional CRVEs ($\hat{V}_{CR2}$ and $\hat{V}_{CR3}$), aimed at reducing bias in the variance estimation step. I build on their framework, making a similar but weaker assumption and discarding an approximation embedded in their test.

Taking a different approach, Cameron, Gelbach and Miller (2008) generate a reference distribution for the test statistic through a resampling method, the wild cluster bootstrap with restricted residuals (henceforth WCR)[1]. Djogbenou, MacKinnon and Nielsen (2019) show that, in addition to performing well in simulations, WCR is a formal asymptotic refinement of the conventional test based on $\hat{V}_{CR0}$ and Hansen's $G-1$ degrees of freedom.

Many applied economists use Cameron and Miller (2015) as a guide for how to handle cluster-robust inference. In Section III, I discuss the tests they recommend, including those discussed above, derived from Hansen (2007), Bell and McCaffrey (2002), Carter, Schnepel and Steigerwald (2017), and Cameron, Gelbach and Miller (2008). Unlike these other tests, my test relies neither on an approximating t-distribution nor on resampling; rather, I find the exact distribution of the test statistic. In Section IV, I show that my test outperforms these other tests in Monte Carlo simulations.

[1]The "restricted residuals" used in WCR are the residuals from the model estimated subject to the restriction that the null hypothesis is true.

II. Model

In this section, I describe a linear model with clustering in a single dimension and cluster-level fixed effects. I include fixed effects because they are a common feature of models where there may also be concerns about clustering. Consider the model:

$$y_{ig} = x_{ig}\beta + \gamma_g + \epsilon_{ig}$$

where x_{ig} is a $(1 \times K)$ vector of covariates and γ_g is a cluster-level fixed effect. Clusters are indexed by g, and individual observations are indexed by i. Let N_g be the (deterministic) number of observations in cluster g. Additionally, for ease of notation:

$$Y_g = \begin{bmatrix} y_{1,g} \\ y_{2,g} \\ \dots \\ y_{N_g,g} \end{bmatrix}, \quad X_g = \begin{bmatrix} x_{1,g} \\ x_{2,g} \\ \dots \\ x_{N_g,g} \end{bmatrix}, \quad \epsilon_g = \begin{bmatrix} \epsilon_{1,g} \\ \epsilon_{2,g} \\ \dots \\ \epsilon_{N_g,g} \end{bmatrix}$$

And similarly let Y (an $(N \times 1)$ matrix), X (an $(N \times k)$ matrix), and ϵ (an $(N \times 1)$ matrix) stack up the outcomes, covariates, and errors of all the clusters, so that X_g contains the rows of X corresponding to cluster g.

For standard fixed-effects estimation of β, the fixed effects γ_g are absorbed. Let $\ddot{Y}_g = Y_g - \frac{1}{N_g}\sum_{i=1}^{N_g} y_{ig}$, $\ddot{X}_g = X_g - \frac{1}{N_g}\sum_{i=1}^{X_g} x_{ig}$, and $\ddot{\epsilon}_g = \epsilon_g - \frac{1}{N_g}\sum_{i=1}^{N_g} \epsilon_{ig}$. Assuming that $\mathbb{E}(\epsilon_{ig} \mid x_{jg}) = 0$, the fixed effects estimator can consistently estimate β:

$$\hat{\beta} = (\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T\ddot{Y}$$

For inference on β, I examine two-sided tests of hypotheses with the form $H_0 : c_0^T\beta = a_0$. I normalize c_0 so that $c_0^T c_0 = 1$. Inference involves calculating a test

statistic t_0 and comparing it to a critical value q^*. An exact test will reject a true hypothesis with some probability α, which is the "size" of the test.

Calculating t_0 begins with estimating $\hat{V}(\hat{\beta})$. The true variance of $\hat{\beta}$ is given by:

$$V(\hat{\beta}) = (\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T\mathbb{E}(\ddot{\epsilon}\ddot{\epsilon}^T)\ddot{X}(\ddot{X}^T\ddot{X})^{-1}$$

In a sample of independent, identically distributed observations, inference could rely on the assumption that the errors are all mutually independent. However, in this clustered setting, I make only the (standard) weaker assumption that the errors are uncorrelated across clusters:

$$\mathbb{E}(\epsilon_g\epsilon_{g'}^T) = 0, \forall g \neq g'$$

Let $\hat{\epsilon}_g = \ddot{Y}_g - \ddot{X}_g\hat{\beta}$ be the residuals for cluster g. The simplest cluster-robust variance estimator, $\hat{V}_{CR0}$, takes the form:

$$\hat{V}(\hat{\beta}) = (\ddot{X}^T\ddot{X})^{-1}\left(\sum_g \ddot{X}_g^T\hat{\epsilon}_g\hat{\epsilon}_g^T\ddot{X}_g\right)(\ddot{X}^T\ddot{X})^{-1}$$

Finally, a test statistic can be generated, using the parameter estimator $\hat{\beta}$ and the variance estimator $\hat{V}(\hat{\beta})$:

$$t_0 = \frac{c_0^T\hat{\beta} - a_0}{\sqrt{c_0^T\hat{V}(\hat{\beta})c_0}}$$

If the hypothesis H_0 is true, then generating a test statistic with a large magnitude should be relatively unlikely. So t_0 can be compared to some critical value q^*, and H_0 is rejected if $|t_0| > q^*$.

III. Hypothesis Tests

III.1. My Proposed Test

I propose a new method of testing linear hypotheses, which I develop here. My test accomplishes two goals. First, under a standard assumption about the errors, my test is exact; that is, my test rejects true hypotheses with probability equal to the nominal test size α. Second, under the (weaker) assumptions that allow all cluster-robust hypothesis tests to be consistent, my test is also consistent in the sense that its rejection rate converges in probability to α. In other words, my test maintains the good asymptotic properties of previous cluster-robust hypothesis tests.

In order to perform an exact hypothesis test, I would like to find a critical value $q^*(H_0, \alpha)$ such that, if H_0 is true, then:

$$P(|t_0| > q^*(H_0, \alpha)) = \alpha$$

The optimal method for selecting critical values would be some $q^*(.,.)$ that gives an exact test for any hypothesis H_0 and any test size α.

If $F_{t_0^2}(.)$, the CDF of t_0^2, was known, I could back out $q^*(.,.)$:

$$F_{t_0^2}((q^*(H_0, \alpha)^2) = 1 - \alpha$$
$$q^*(H_0, \alpha) = \sqrt{F_{t_0^2}^{-1}(1 - \alpha)}$$

The intuition for my test is that I make an assumption that is strong enough to determine the distribution of $F_{t_0^2}(.)$.

ASSUMPTION 1: *The errors are normal and homoskedastic with intracluster*

correlations that are constant within and across clusters:

$$\epsilon_g \sim N(0, \sigma^2\Omega_g), \quad \Omega_g = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ \rho & \rho & \dots & 1 \end{bmatrix}$$

Note that σ and ρ may be unknown. I refer to this assumption as "standard" because it is a slightly weaker version of the same assumption that several other papers use to adjust cluster-robust inference in finite samples (Carter, Schnepel and Steigerwald, 2017; Bell and McCaffrey, 2002). With this assumption, $F_{t_0^2}(.)$ may be found.

THEOREM 1: *Suppose that Assumption 1 holds and that H_0 is true. Then:*

$$F_{t_0^2}(q) = L(q; X, H_0)$$

where $L(.;.,.)$ is known.

I prove Theorem 1 by deriving $L(.;.,.)$ in Appendix A. To implement my test[2]:

1) Calculate the test statistic $t_0 = \frac{c_0^T\hat{\beta} - a_0}{\sqrt{c_0^T\hat{V}(\hat{\beta})c_0}}$

2) Find $L(q; X, H_0)$, the CDF of t_0^2 under Assumption 1 and H_0

3) Determine[3] the critical value q^* such that $L((q^*)^2; X, H_0) = 1 - \alpha$

4) Reject H_0 if $|t_0| > q^*$

Above, I calculate a test statistic using the variance estimator $\hat{V}_{CR0}$. In Section IV, I use two additional variants of my test that are based on different variance estimators. These variants use $\hat{V}_{CR2}$ and $\hat{V}_{CR3}$, the estimators given by Bell and

[2]My test is available in R as the function "p.value.meis()" in the package "clubsoda", available through github.

[3]Since $L(q; X, H_0)$ is increasing in q, $L(.; X, H_0)$ can easily be inverted numerically.

McCaffrey (2002). The proof of Theorem 1 in Appendix A holds for both of these variants; they are exact tests under Assumption 1.

My test using $\hat{V}_{CR0}$ is also asymptotically valid under the relatively weak assumptions described by Carter, Schnepel and Steigerwald (2017). They demonstrate that, when using $\hat{V}_{CR0}$, the test statistic converges to a standard normal distribution: $t_0 \xrightarrow{d} N(0,1)$. In Appendix B, I build on the result from Carter, Schnepel and Steigerwald (2017) to prove that my test is asymptotically valid. In large samples, my test will reject a true hypothesis with probability α. Proving that my tests which do not use $\hat{V}_{CR0}$ are also asymptotically valid remains an area of future work.

III.2. Other Tests

In Section IV, I compare my test's performance with the performance of other tests from the literature on inference in clustered samples. In this section, I briefly discuss how those other tests work and how they relate to my tests. Specifically, I will look at the previous tests recommended in Cameron and Miller (2015). These tests can be roughly divided into analytic tests, which select a critical value for t_0 from a known distribution, and resampling-based tests, which generate a simulated distribution of test statistics from which critical values are drawn.

In a test I refer to as "Hansen", derived from Hansen (2007), it is recommended to estimate $\hat{V}(\hat{\beta})$ with $\hat{V}_{CR3}$ and to select critical values for the test statistic t_0 from $T(G-1)$, a t-distribution with $G-1$ degrees of freedom.

The test from Bell and McCaffrey (2002), henceforth "BM", involves estimating $\hat{V}(\hat{\beta})$ with $\hat{V}_{CR2}$ and selecting critical values for t_0 from $T(m)$, where m is calculated according to a "Satterthwaite approximation" of t_0. For an explanation of how the Satterthwaite approximation works and what assumptions it relies on, see Appendix C.

Cameron and Miller (2015) also recommend a test, henceforth "CSS", derived from Carter, Schnepel and Steigerwald (2017). In this test, $\hat{V}(\hat{\beta})$ is estimated

with $\hat{V}_{CR0}$, and critical values for t_0 are selected from $T(G^*)$, where G^* is called the "effective number of clusters". In Appendix C, I show how G^* is a simplified version of the Satterthwaite approximation.

Hansen, BM, and CSS all approximate the test statistic t_0 as a t-distribution. By contrast, the last method recommended by Cameron and Miller (2015) is a resampling method, the wild cluster bootstrap. Using this method, $\hat{V}(\hat{\beta})$ is estimated with $\hat{V}_{CR0}$, and then over many bootstrap iterations, the residuals are resampled by multiplying them by values drawn from an auxiliary distribution with mean 0 and variance 1. A critical value t^* is then selected from the bootstrapped distribution of test statistics.

For choosing among the various specifications of the wild cluster bootstrap, I follow Djogbenou, MacKinnon and Nielsen (2019), a more recent study that tested many variants in simulations. They recommend resampling the restricted residuals (the residuals from the restricted model, subject to H_0), with the auxiliary distribution being either the Rademacher distribution or the Mammen distribution. I refer to these tests as "WCR-R" and "WCR-M", respectively.

There are some parallels between my test and previous analytic tests in the literature. However, these other methods all approximate the distribution of the test statistic. My test does not use an approximation. Rather, I have made an assumption that is strong enough to fully determine the distribution of the test statistic. In the next section, I will demonstrate that this approach makes my test perform better in many samples; my test rejects true hypotheses at the correct rate even when other tests fail to do so.

IV. Simulations

In this section, I show the results of Monte Carlo simulations that demonstrate that that my test is exact and that previous tests fail to reject true hypotheses at the correct rate in certain kinds of samples. Specifically, I focus on samples with few clusters, few clusters with treatment variation, cluster size outliers, and

treatment intensity outliers.

Depending on the specification, I vary certain features of the design matrix X:

- G = number of clusters
- J = number of clusters w/ treatment variation
- N_1 = size of first cluster
- ϕ = treatment intensity of first cluster

In my simulation experiments, I use the following data-generating process:

$$
\begin{aligned}
x_{1ig} &= \mathbb{1}(g \leq J) \times \phi^{\mathbb{1}(g=1)} \times \frac{x^*_{1ig} - 8}{4}, \quad x^*_{1ig} \sim \chi^2_8 \\
x_{2ig} &= \frac{x^*_{2ig} - 8}{4}, \quad x^*_{2ig} \sim \chi^2_8 \\
y_{ig} &= \beta_0 + \beta_1 x_{1ig} + \beta_2 x_{2ig} + \gamma_g + \epsilon_{ig}
\end{aligned} \tag{1}
$$

Following Djogbenou, MacKinnon and Nielsen (2019), the covariates x_{1ig} and x_{2ig} are generated with distributions that are both skewed and leptokurtic. This highlights the fact that my main result does not require normally-distributed covariates.

For each cluster beyond the first, there are 5 obserations in that cluster – that is, for $g > 1$, $N_g = 5$. I set $\beta_0 = 1$, $\beta_1 = 2$, and $\beta_2 = 3$. Since fixed effects are absorbed before any estimation, the values of γ_g do not affect estimation or inference, so for simplicity I set $\gamma_g = 0$ for all g. Unless otherwise specified, the specification parameters have default values $G = 200$, $J = 200$, $N_1 = 5$, and $\phi = 1$.

For now, I generate $\epsilon_{ig} \sim N(0, 1)$; this DGP meets the conditions of Assumption 1. In Section IV.IV.1, I will alter this specification with several violations of Assumption 1. Besides simply confirming what I showed Theorem 1, this set of simulations serves to demonstrate the conditions where previous tests tend to over- or under-reject true hypotheses.

In each simulation, I generate a sample according to the DGP, and I estimate $\hat{\beta}$ with the standard fixed effects estimator. Then, I test the true hypothesis $H_0 : \beta_1 = 2$ using my test as well as each of the tests recommended by Cameron and Miller (2015). Different tests require estimating $\hat{V}(\hat{\beta})$ using different variance estimators and comparing test statistics to critical values selected according to different methods. I discuss the different tests in Section III.

First, I present results for simulations in which the number of clusters G takes on the following values: $G = 5, 10, 20, 50, 100, 200$. In Figure 1, I plot the rejection rates in this DGP for each of the tests discussed in Section III. In general, the degree of a given test's over- or under-rejection depends on the size of the test α. I therefore show rejection rates for 5% tests in Panel 1a and 1% tests in Panel 1b. As the number of clusters gets small, CSS tends to overreject and WCR-M tends to underreject fairly dramatically. Hansen and BM look more reasonable but also tend to underreject for small G. WCR-R rejects at the correct rate except when $G = 5$, where it seems to fail completely. When a sample has a small number of clusters, my test is the only exact test.

Next, in Figure 2, I show rejection rates from simulations where I vary J, the number of clusters with treatment variation. As seen in (1), for clusters beyond the J-th cluster, I simply multiply the value of x_{1ig} by 0. Hansen and CSS both overreject and WCR-M underrejects at $J \leq 20$. BM and WCR-R both struggle at $J = 5$. When a sample has a small number of clusters with treatment variation, my test is the only exact test.

Figure 3 plots rejection rates for different degrees of cluster size heterogeneity, where $N_1 = 20, 100, 200, 500$. In each specification, there are 200 clusters, and for $g > 1$, $N_g = 5$. So when $N_1 = 500$, the first cluster contains about a third of the observations in the sample. In that most extreme case, BM and CSS underreject, Hansen overrejects, and WCR-M overrejects for a 5% test only. My test is exact here, and WCR-R seems to at least be resilient to this form of cluster size heterogeneity.

In Figure 4, I show results for several values of the treatment intensity outlier parameter, so $\phi = 1, 5, 9, 13, 18, 24, 30$. Recall that the value of x_{1ig} is multiplied by ϕ for $g = 1$ only, so that a large value of ϕ creates a cluster that is an outlier in terms of variance in the treatment variable. Hansen overrejects for $\phi \geq 9$ and BM and CSS both underreject for $\phi \geq 9$. Around $\phi = 18$, both WCR-R and WCR-M begin to substantially underreject. In the presence of a large treatment intensity outlier, my test is the only exact test.

As discussed in Section III, several of the methods (Hansen, CSS, and BM) approximate the test statistic t_0 with a t-distribution. Of those, Hansen and CSS can underreject or overreject, depending on the specification. Why does this happen? The approximation to a t-distribution depends implicitly on having an unbiased variance estimator. However, Hansen uses $\hat{V}_{CR3}$, which is biased up in this DGP (corresponding to underrejection), and CSS uses $\hat{V}_{CR0}$, which is biased down in this DGP (corresponding to overrejection). It is also true that both of these methods can select inappropriate critical values due to the approximation itself. The reason that Hansen and CSS underreject in some specifications and overreject in other specifications is that the bias in the variance estimator causes the rejection rate to move in one direction and misspecified critical value selection causes the rejection rate to move in the opposite direction. For a deeper discussion of the approximation of the test statistic t_0 with a t-distribution, see Section C.

IV.1. Robustness

So far, the Monte Carlo simulations have met the conditions of Assumption 1. Here, I present additional simulation evidence regarding the robustness of my test to violations of Assumption 1.

Since my test is asymptotically valid, violations of Assumption 1 can only affect the performance of my test in samples with few clusters, few clusters with treatment variation, cluster size outliers, or treatment intensity outliers.

I focus on three violations of Assumption 1:

- Non-normal errors
- Serial correlation, where ϵ_{ig} is an AR(1) process
- Heteroskedasticity

These three violations correspond roughly to different parts of Assumption 1: normality, constant intracluster correlation, and homoskedasticity. It may be that normality of the errors can be relaxed when N_g, the number of observations per cluster, is large, and proving this is an area of future work. Still, I test robustness to non-normal errors here. Bertrand, Duflo and Mullainathan (2004) highlight serial correlation as an important potential problem in differences-in-differences analyses of panel data, and CRVEs are powerful tools for addressing serial correlation. For that reason, it seems natural to check test performance when ϵ_{ig} is serially correlated as an AR(1) process. MacKinnon and Webb (2018) find that a simple analytic test using a CRVE is less reliable when the errors are heteroskedastic. Following that paper as well as several other simulation studies of cluster-robust inference (Cameron, Gelbach and Miller, 2008; Djogbenou, MacKinnon and Nielsen, 2019), I also test robustness to the error variance differing across clusters.

In Figure 5, I plot rejection rates when the errors have a normal distribution and when they have non-normal distributions. I selected distributions with substantially different third and fourth moments than the normal distribution. The Laplace distribution is leptokurtic, the uniform distribution is platykurtic, and the log-normal distribution is skewed right. When $J = 200$ and $\phi = 1$ (panels 5a and 5d), so that the number of clusters with treatment variation is large and there are no treatment intensity outliers, the different error distributions do not matter and every test rejects at the correct rate because all of the tests (including my test) are asymptotically valid. When $J = 5$ (panels 5b and 5e) and when $\phi = 20$ (panels 5c and 5f), my test is not quite exact for non-normal error distributions, but it performs better than any of the other tests.

Next, I check test performance when the errors are serially correlated. Figure 6 shows the results of simulations where the errors are distributed as a stationary AR(1) process: $\epsilon_{i,g} = \psi\epsilon_{i-1,g} + \sqrt{1-\psi^2}\epsilon^*_{i,g}$, where $\epsilon^*_{i,g} \sim N(0,1)$. It bears repeating that when $J = 200$ and $\phi = 1$ (panels 6a and 6d), every test rejects at the correct rate because all of the tests (including my test) are asymptotically valid. When $J = 5$ (panels 6b and 6e) and when $\phi = 20$ (panels 6c and 6f), my test is still nearly exact.

Finally, in Figure 7, I show the rejection rates for heteroskedastic errors. Specifically, I use an error distribution very similar to the one in Djogbenou, MacKinnon and Nielsen (2019): $\epsilon_{ig} = (1+\xi(x^*_{1ig})^2))^{\frac{1}{2}}\epsilon^*_{ig}$, where $\epsilon^*_{ig} \sim N(0,1)$. The error variance is higher for observations with greater magnitudes of x_{1ig}. When $J = 5$ (panels 7b and 7e), my test overrejects when ξ is high. When $\phi = 20$ (panels 7c and 7f), my test underrejects when ξ is high. Even in these cases, my test performs about as well as any other test.

My test is only plausibly vulnerable to violations of Assumption 1 when the test statistic t_0 is not behaving asymptotically. I have shown in this section that several straightforward violations of Assumption 1 don't seem to affect the performance of my test very much. In particular, my test is less affected by these violations than the other tests are affected by samples with few clusters, few clusters with treatment variation, cluster size outliers, or treatment intensity outliers.

V. Empirical Application

In this section, I apply my test to the empirical setting in Abouk and Adams (2013). First, I estimate their model and demonstrate some of the features of the sample that make the differences between methods of hypothesis testing meaningful. After that, I perform a simulation exercise using their data to discern how well different tests work in a context like this one.

Abouk and Adams (2013) estimate the impact on fatal car crashes of state bans on texting-while-driving. They find that the enforcement mechanisms behind some of the bans were stronger than others, and they define two dummy variables $StrongBan_{s,t}$ and $WeakBan_{s,t}$ based on those two types of bans.

If $StrongBan_{s,t} = 1$, then police officers could pull over any driver for texting-while-driving as a primary offense. If $WeakBan_{s,t} = 1$, then texting-while-driving was either a secondary offense only or the ban only applied to a small subset of the population (e.g., drivers under 21 years old). By definition, no state has $StrongBan_{s,t} = WeakBan_{s,t} = 1$, and it happens that many states had $StrongBan_{s,t} = WeakBan_{s,t} = 0$.

I replicate their estimation:

$$\log FatalCrashes_{s,t} = \beta_{Strong} StrongBan_{s,t} + \beta_{Weak} WeakBan_{s,t} + w_{s,t}\delta + \gamma_s + \theta_t + \epsilon_{s,t}$$

where $FatalCrashes_{s,t}$ is the number of fatal, single-vehicle, single-occupant car crashes in state s in year t, $w_{s,t}$ are a set of time-varying demographic and economic characteristics (e.g., unemployment rate), γ_s are state fixed effects, and θ_t are month-by-year fixed effects. The errors $\epsilon_{s,t}$ are clustered by state.

Table 9 shows coefficient estimates for $\hat{\beta}_{Strong}$ and $\hat{\beta}_{Weak}$. A "strong" ban reduced crashes by 8.1%. A "weak" ban is associated with an increase in crashes by 7.5%, but the estimate is imprecise.

I also give three different measures of the effective sample size. There were 49 states in the sample. I calculate the effective number of clusters G^* according to Carter, Schnepel and Steigerwald (2017) and the Satterthwaite-approximated degrees of freedom m from Bell and McCaffrey (2002). All three of these measures are used as the degrees of freedom v in different hypothes.

One way to think about these measures is as measures of the variance of the test statistic t_0. They represent the number of observations such that, if the sample

was i.i.d., the variance of the t_0 would match the current sample. The measures differ between coefficients because test statistics based on different hypotheses have different behavior. The measures differ within a coefficient because they make different assumptions and simplifications [4].

These measures suggest that the asymptotic properties of the standard CRVE might not dominate the behavior of a cluster-robust hypothesis test. Therefore, the differences between different methods of hypothesis testing might matter, especially for $\hat{\beta}_{Weak}$, where the Satterthwaite degrees of freedom is only 5.7.

In the bottom half of Table 9, I show the p-values given by each test discussed in Section III. $\hat{\beta}_{Strong}$ is significantly different from 0 at the 5% level according to every test. $\hat{\beta}_{Weak}$ is significantly different from 0 at the 5% level according to the test that makes no adjustment to White (1984), using $\hat{V}_{CR0}$ with critical values drawn from the standard normal distribution, but it is not significant according to any other test. The adjustments to the CRVE and to critical value selection matter when the effective sample size is small.

It is not possible to know the correct p-value because we do not know the distribution of the errors $\epsilon_{s,t}$ and so we cannot calculate probabilities of precise quantiles of the test statistic t_0. However, I have performed a simulation exercise to learn which tests work well in samples like this one.

In this exercise, I randomize the values of the treatment variables $StrongBan_{s,t}$ and $WeakBan_{s,t}$. In each simulation, each state s is randomly assigned another state $\tilde{s}$ (without replacement). Then I assign $RStrongBan_{s,t} = StrongBan_{\tilde{s},t}$, and I assign $RWeakBan_{s,t} = WeakBan_{\tilde{s},t}$, so that each state s is assigned the entire sequence of treatment variables from state $\tilde{s}$.

I have assigned these treatments randomly and after the fact, so it is safe to assume that $\beta_{RStrong} = \beta_{RWeak} = 0$. Since $H_0 : \beta_{RStrong} = 0$ and $H_0 : \beta_{RWeak} = 0$ are true hypotheses, an exact hypothesis test should reject with probability

[4] For a deeper discussion of the assumptions and simplifications behind the the effective number of clusters G^* from Carter, Schnepel and Steigerwald (2017) and the Satterthwaite-approximated degrees of freedom m from Bell and McCaffrey (2002), see Appendix C

equal to the nominal size of the test. Figure 8 shows the rejection rates for each test method across all the simulations.

Hansen substantially overrejects $H_0 : \beta_{RWeak} = 0$ for a test size of $\alpha = 0.01$, while CSS substantially overrejects $H_0 : \beta_{RStrong} = 0$ for a test size of $\alpha = 0.01$. CSS also seems to overreject somewhat for a test size of $\alpha = 0.05$ on either coefficient.

The variant of my test that uses $\hat{V}_{CR3}$ is about as close to exact as any other test for both coefficients and test sizes. Also, it's worth noting that in Figure 8 the rejection rates of all variants of my test tend to be between those of the other tests. This comports with the Monte Carlo results in Section IV, where some tests over-reject and some under-reject.

The important features of this empirical setting are not unique. Staggered changes in public policy across states or counties provide useful natural experiments for learning about the impact of public policy. A good hypothesis test is necessary for understanding how much can be learned from a given experiment.

VI. Conclusion

In this paper, I have proposed a hypothesis test for inference in clustered samples with cluster-level fixed effects. My test is robust to samples with few clusters, few clusters with treatment variation, cluster size outliers, or treatment intensity outliers. Indeed, my test is exact under normal, i.i.d. errors, and in simulations it seems to perform well compared to other methods from the literature.

Many samples are large enough that the conventional cluster-robust test will work fine, drawing critical values from the standard normal distribution. However, some samples are not as large as they seem, in the sense of the distribution of the test statistic.

Samples with many observations but few clusters require adjustment to the conventional test. Samples with many clusters but only a few with treatment variation require adjustment. And samples where the residual treatment variation

(conditional on covariates) is concentrated in a small number of clusters require adjustment. Covariates can hide the features that change the behavior of the test statistic. It is worthwhile to use a test that is simply robust to these features.

Inference in finite samples is limited by heteroskedasticity and other unknown features of the unobserved errors. I provide a hypothesis test that is as robust as possible to the known, observable features of the sample.

REFERENCES

Abouk, Rahi, and Scott Adams. 2013. "Texting Bans and Fatal Accidents on Roadways: Do They Work? Or Do Drivers Just React to Announcements of Bans?" *American Economic Journal: Applied Economics*, 5(2): 179–199.

Arellano, M. 1987. "Computing Robust Standard Errors for Within-groups Estimators." *Oxford Bulletin of Economics and Statistics*, 49(4): 431–434.

Bell, Robert M., and Daniel F. McCaffrey. 2002. "Bias reduction in standard errors for linear regression with multi-stage samples." *Survey Methodology*, 28(2): 169–182. ISBN: 0714-0045.

Bertrand, M., E. Duflo, and S. Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *The Quarterly Journal of Economics*, 119(1): 249–275.

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources*, 50(2): 317–372.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics*, 90(3): 414–427.

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald. 2017. "Asymptotic Behavior of a t -Test Robust to Cluster Heterogeneity." *The Review of Economics and Statistics*, 99(4): 698–709.

Djogbenou, Antoine A., James G. MacKinnon, and Morten Ørregaard Nielsen. 2019. "Asymptotic theory and wild bootstrap inference with clustered errors." *Journal of Econometrics*, 212(2): 393–412.

Hansen, Christian B. 2007. "Asymptotic properties of a robust variance matrix estimator for panel data when is large." *Journal of Econometrics*, 141(2): 597–620.

Imhof, J. P. 1961. "Computing the Distribution of Quadratic Forms in Normal Variables." *Biometrika*, 48(3/4): 419.

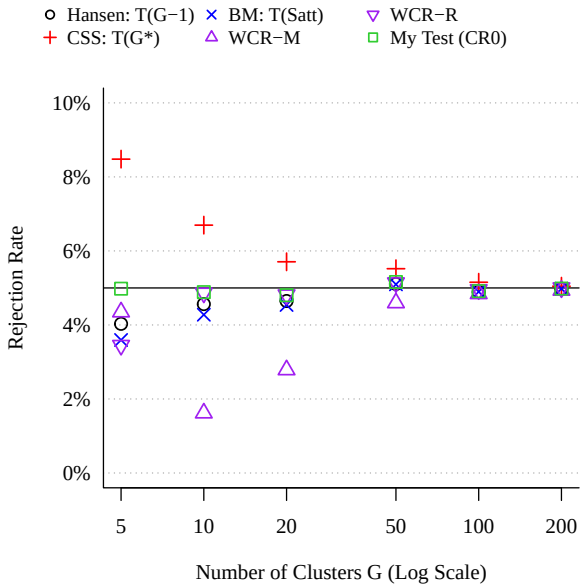
Liang, Kung-Yee, and Scott L. Zeger. 1986. "Longitudinal data analysis using generalized linear models." *Biometrika*, 73(1): 13–22.

MacKinnon, James G., and Matthew D. Webb. 2018. "The wild bootstrap for few (treated) clusters." *The Econometrics Journal*, 21(2): 114–135.

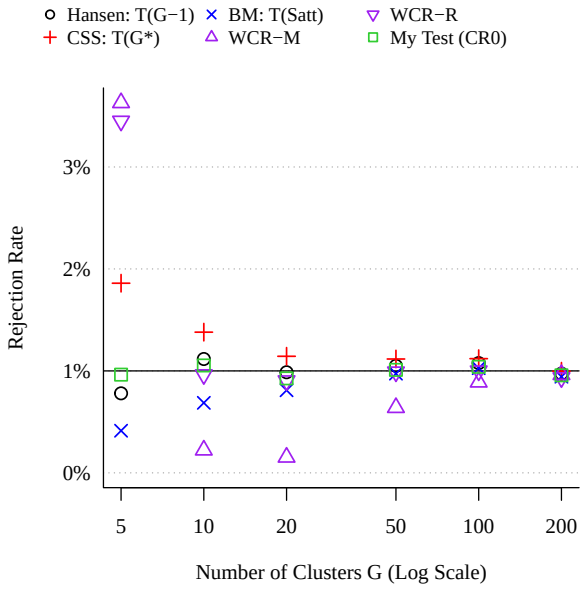
White, Halbert. 1984. *Asymptotic theory for econometricians. Economic theory, econometrics, and mathematical economics*, Orlando:Academic Press.

Figure 1. : Comparison of Hypothesis Tests With Varying G

(a) Test Size $\alpha = 0.05$



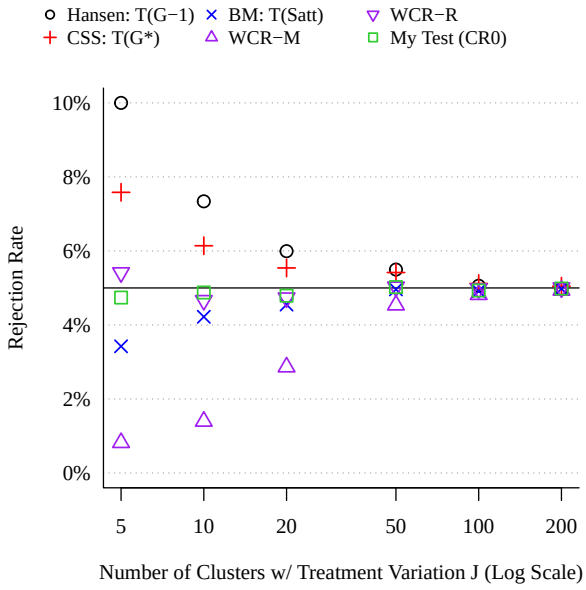
(b) Test Size $\alpha = 0.01$



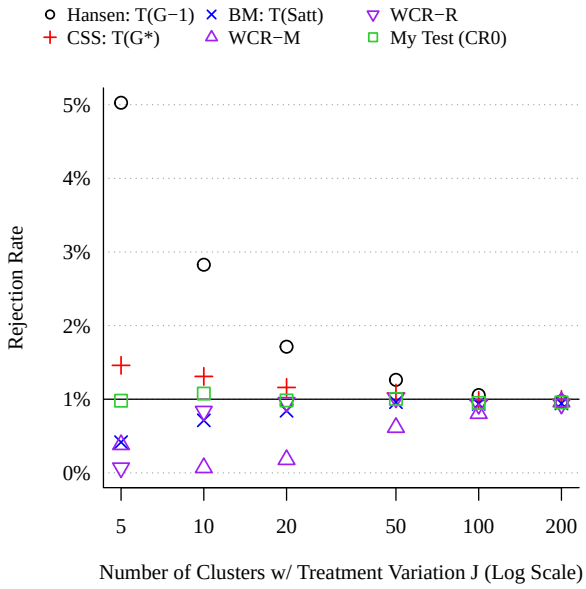
Notes: In all of these simulations, $J = 200$, $N_1 = 5$, and $\phi = 1$. The errors are i.i.d. standard normal.

Figure 2. : Comparison of Hypothesis Tests With Varying J

(a) Test Size $\alpha = 0.05$



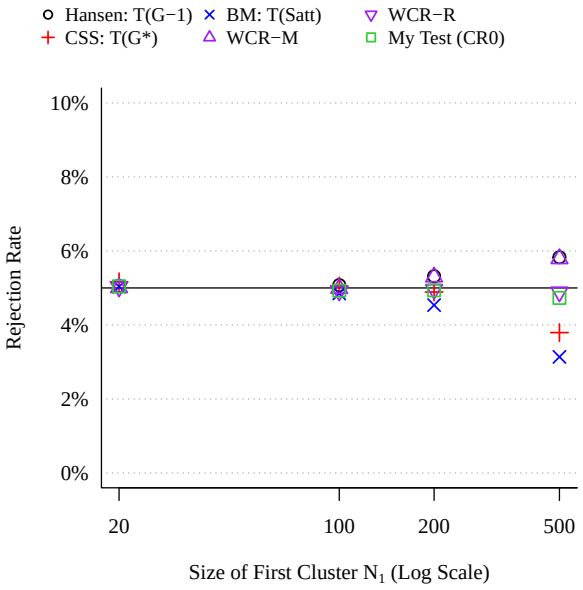
(b) Test Size $\alpha = 0.01$



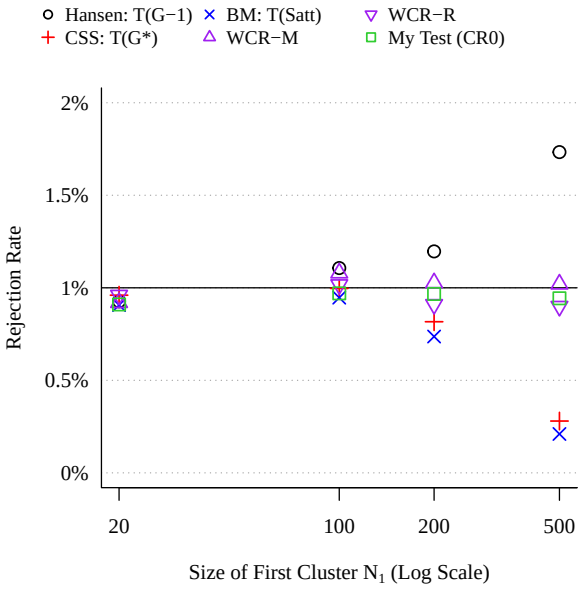
Notes: In all of these simulations, $G = 200$, $N_1 = 5$, and $\phi = 1$. The covariate x_{1ig} is distributed as $x_{1ig} = \mathbb{1}(g \leq J) \times \phi^{\mathbb{1}(g=1)} \times \frac{x^*_{1ig}-8}{4}$, where $x^*_{1ig} \sim \chi^2_8$. The errors are i.i.d. standard normal.

Figure 3. : Comparison of Hypothesis Tests With Varying N_1

(a) Test Size $\alpha = 0.05$



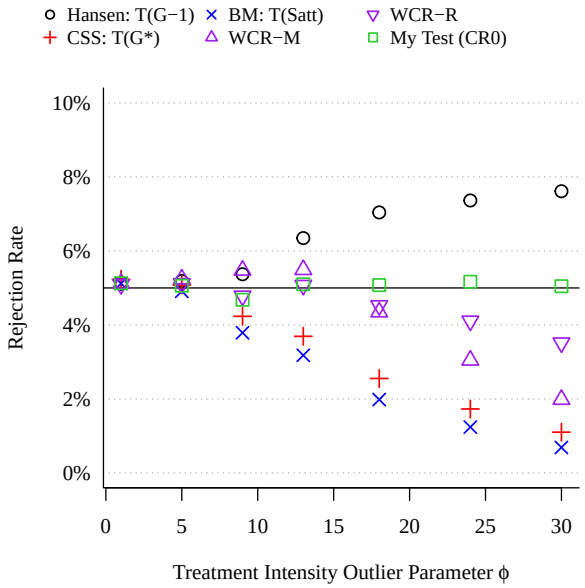
(b) Test Size $\alpha = 0.01$



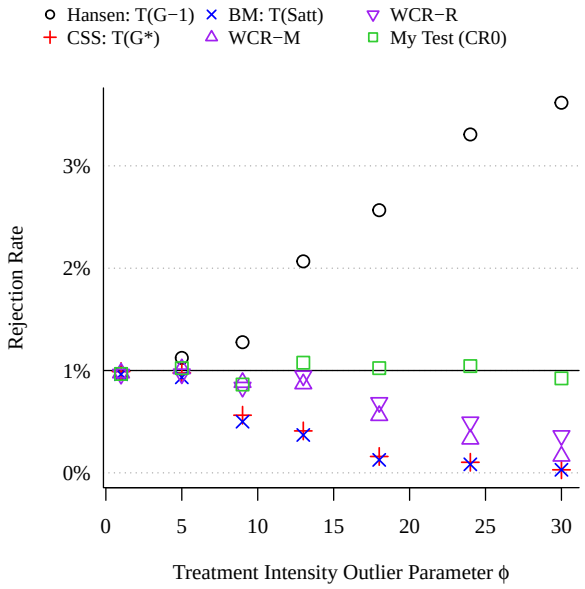
Notes: In all of these simulations, $G = 200$, $J = 200$, and $\phi = 1$. The errors are i.i.d. standard normal.

Figure 4. : Comparison of Hypothesis Tests With Varying ϕ

(a) Test Size $\alpha = 0.05$

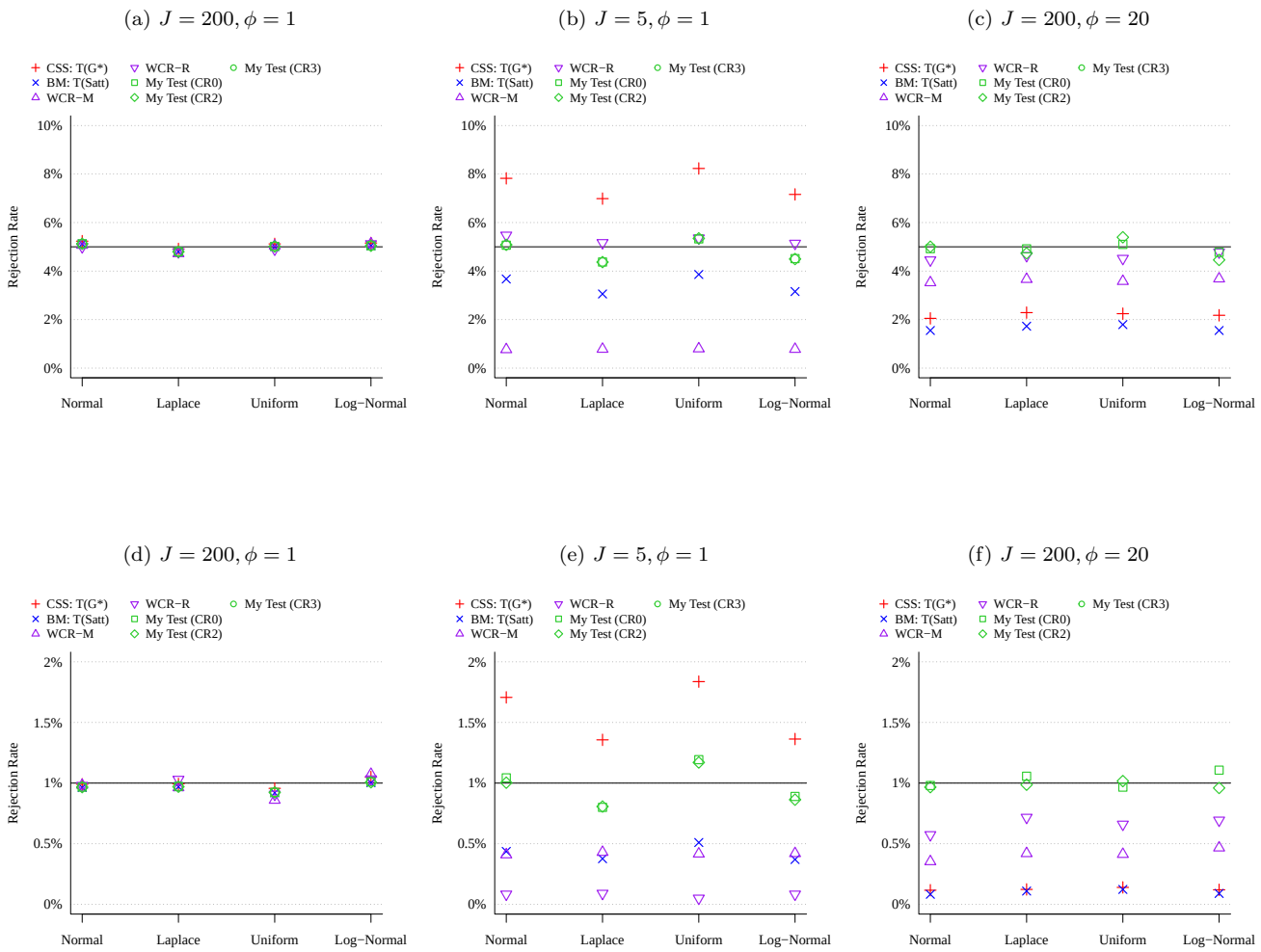


(b) Test Size $\alpha = 0.01$



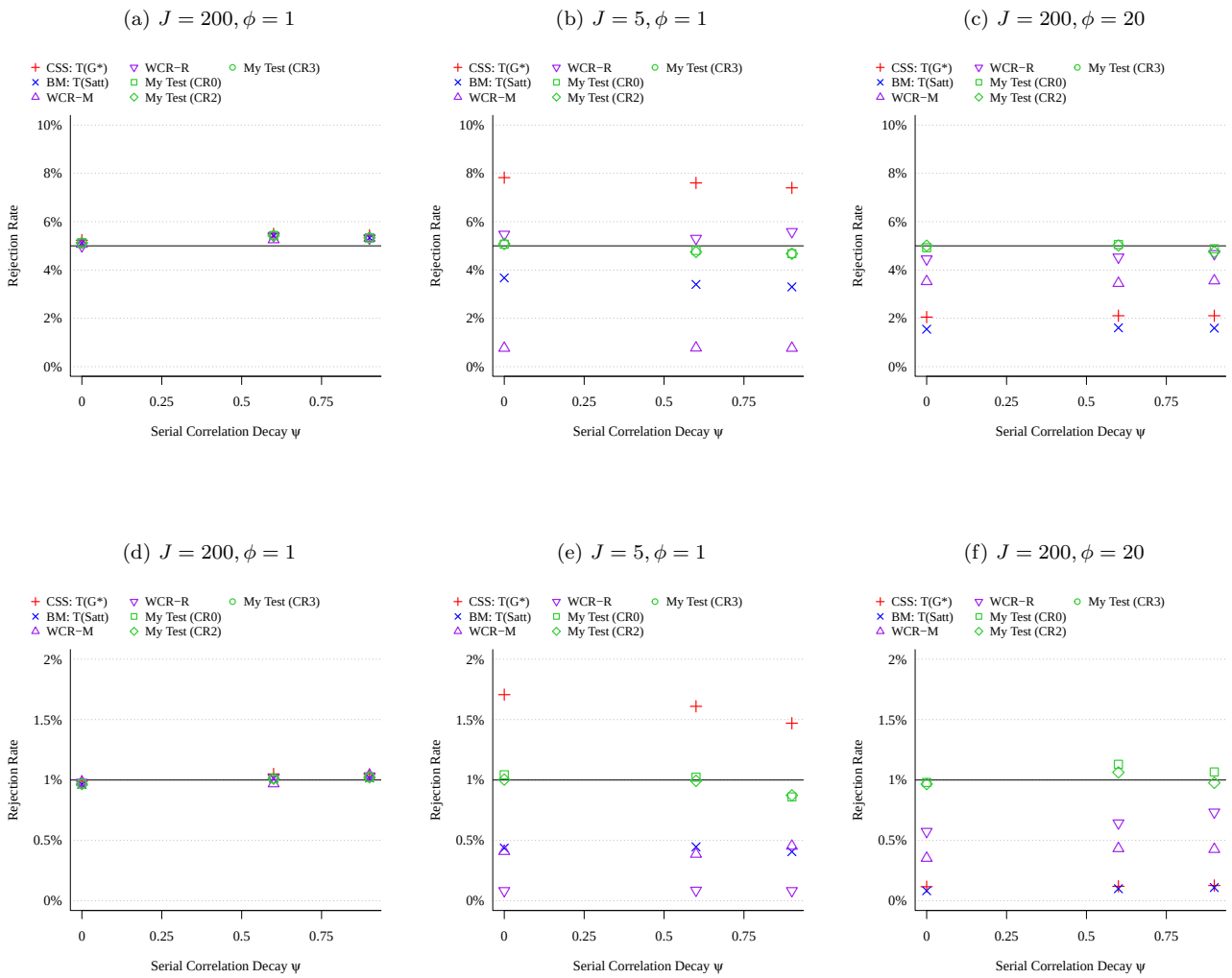
Notes: In all of these simulations, $G = 200$, $J = 200$, and $N_1 = 5$. The covariate x_{1ig} is distributed as $x_{1ig} = \mathbb{1}(g \leq J) \times \phi^{\mathbb{1}(g=1)} \times \frac{x^*_{1ig}-8}{4}$, where $x^*_{1ig} \sim \chi^2_8$. The errors are i.i.d. standard normal.

Figure 5. : Comparison of Hypothesis Tests When the Error Distribution Varies



Notes: In all of these simulations, $G = 200$ and $N_1 = 5$. The errors are distributed as recentered and rescaled normal, Laplace, uniform, and log-normal distributions such that $\mathbb{E}(\epsilon_{ig}) = 0$ and $V(\epsilon_{ig}) = 1$.

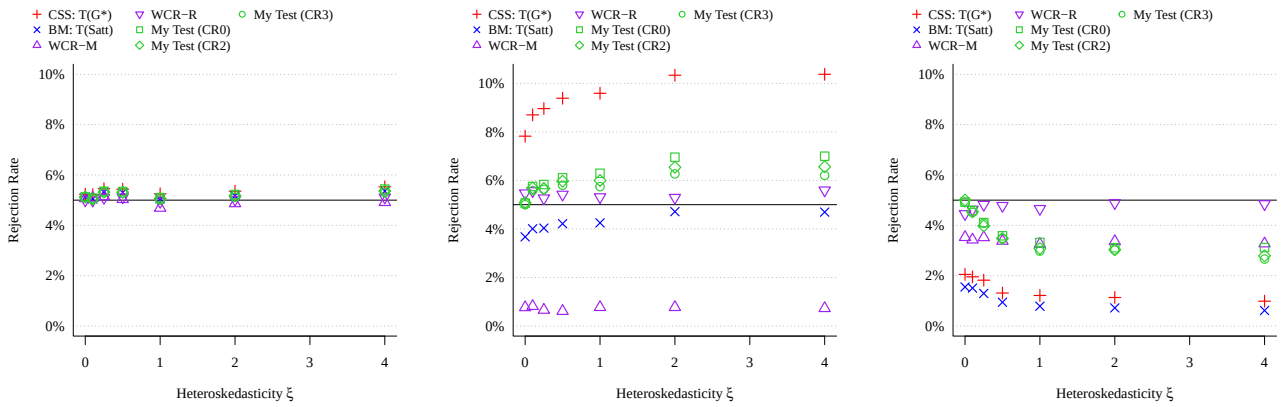
Figure 6. : Comparison of Hypothesis Tests When the Errors are AR(1)



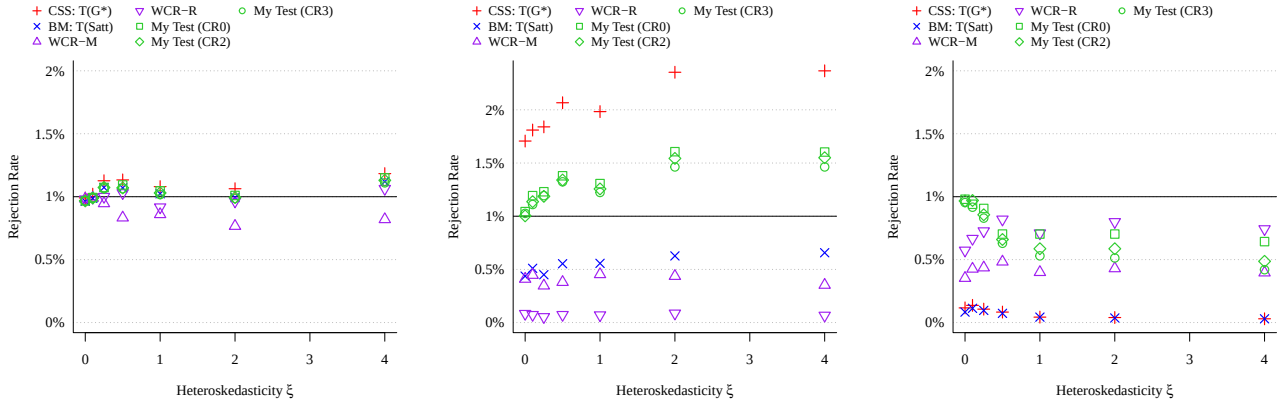
Notes: In all of these simulations, $G = 200$ and $N_1 = 5$. The errors are distributed as $\epsilon_{i,g} = \psi\epsilon_{i-1,g} + \sqrt{1-\psi^2}\epsilon^*_{i,g}$, where $\epsilon^*_{i,g} \sim N(0,1)$.

Figure 7. : Comparison of Hypothesis Tests When the Errors are Heteroskedastic

(a) $J = 200, \phi = 1$ (b) $J = 5, \phi = 1$ (c) $J = 200, \phi = 20$



(d) $J = 200, \phi = 1$ (e) $J = 5, \phi = 1$ (f) $J = 200, \phi = 20$

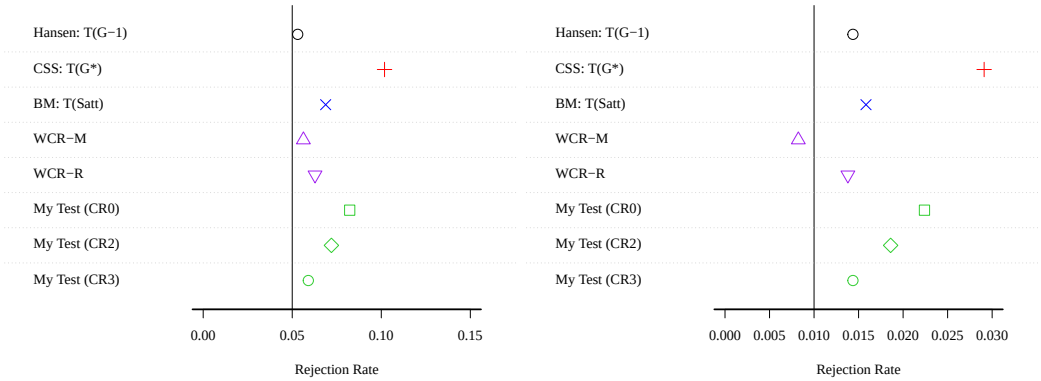


Notes: In all of these simulations, $G = 200$ and $N_1 = 5$. The errors are distributed as $\epsilon_{ig} = (1 + \xi(x^*_{1ig})^2))^{\frac{1}{2}}\epsilon^*_{ig}$, where $\epsilon^*_{ig} \sim N(0, 1)$.

Figure 8. : Simulation Exercise Based On Abouk and Adams (2013)

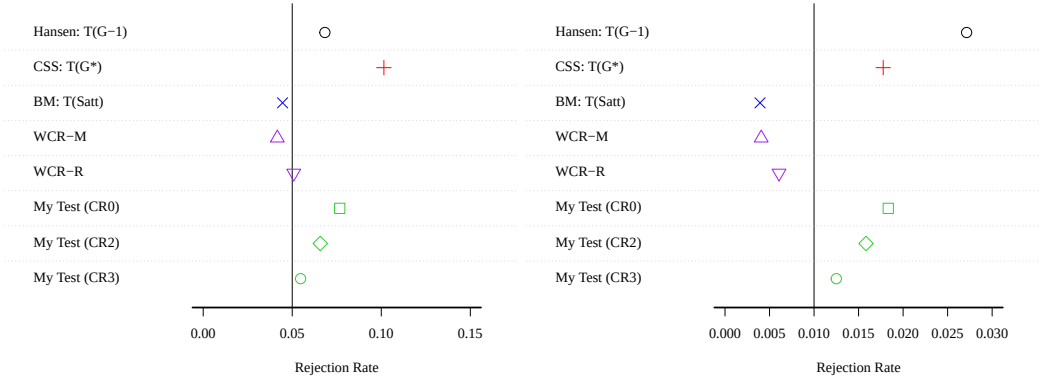
(a) $H_0 : \beta_{RStrong} = 0, \quad \alpha = 0.05$

(b) $H_0 : \beta_{RStrong} = 0, \quad \alpha = 0.01$



(c) $H_0 : \beta_{RWeak} = 0, \quad \alpha = 0.05$

(d) $H_0 : \beta_{RWeak} = 0, \quad \alpha = 0.01$



Notes: These plots show the rejection rate of each test when the values of the treatment variables $StrongBan_{s,t}$ and $WeakBan_{s,t}$ have been randomized.

Figure 9. : Estimates From Abouk and Adams (2013)

	Dependent Variable: Log(Number of Crashes)	
	Independent Variable:	
	Strong Ban	Weak Ban
Coefficient Estimate	-0.081	0.075
	(0.025)	(0.037)
# of Clusters (States)	49	49
G*, CSS Effective # of Clusters	11.6	7.2
m, BM-Satterthwaite DOF	10.7	5.7
P-Values for H0: β=0		
No Adjustment	0.001	0.038
Hansen, T(G-1)	0.011	0.212
CSS, T(G*)	0.007	0.075
BM, T(m)	0.012	0.152
WCR-R Bootstrap	0.019	0.193
WCR-M Bootstrap	0.035	0.189
My Test (CR0)	0.008	0.111
My Test (CR2)	0.009	0.146
My Test (CR3)	0.011	0.190

Notes: Both columns contain estimates from the same regression. That regression also included state and year fixed effects as well as a set of time-varying state-level economic and demographic controls. Standard errors are clustered by state. The effective number of clusters is calculated according to Carter, Schnepel and Steigerwald (2017) and the Satterthwaite-approximated degrees of freedom is calculated according to Bell and McCaffrey (2002). P-values are calculated for all of the hypothesis tests discussed in Section III.

APPENDIX A Proof of Theorem 1

The null hypothesis H_0 is true if $c_0^T\beta = a_0$. Assumption 1 holds when $\epsilon_g \sim N(0, \sigma^2\Omega_g)$, where $\Omega_g = (1-\rho)I_g + \rho\iota_g\iota_g^T$ (σ and ρ may be unknown).

Theorem 1 says that when H_0 is true and Assumption 1 holds, the CDF of the squared test statistic is a known function of the design matrix X and the hypothesis H_0:

$$P(t_0^2 < q \mid X) = L(q; X, H_0)$$

In this section, I will prove Theorem 1 by deriving $L(.;.,.)$. I give some notation to help with the analysis, I prove that $L(.;.,.)$ is known in principle, and I show how $L(.;.,.)$ can be calculated quickly in practice.

A.1 Notation

Let N_g be the number of observations in cluster g, and let $N = \sum_g N_g$. Then let I_g be an identity matrix of size N_g, and let ι_g be a column vector of length N_g whose elements are all 1. So then let $M_g = I_g - \frac{1}{N_g}\iota_g\iota_g^T$, and note that $M_g\iota_g = 0$.

Recall that:

$$y_{ig} = x_{ig}\beta + \gamma_g + \epsilon_{ig}$$

For identification, we have that $\mathbb{E}(\epsilon_{ig} \mid x_{jg}) = 0$, and for inference, we have that $\mathbb{E}(\epsilon_{ig}\epsilon_{jg'}) = 0$. Y_g is $(N_g \times 1)$, stacking up the dependent variable within cluster g, and X_g is $(N_g \times K)$, stacking up the covariates within cluster g. Then $\epsilon_g = Y_g - X_g\beta - \gamma_g\iota_g$.

For fixed effects absorption – that is, to absorb γ_g – the cluster-level means of y_{ig} and x_{ig} are subtracted from the individual-level y_{ig} and x_{ig}, respectively.

Stated another way:

$$\begin{aligned}
\ddot{Y}_g &= M_g Y_g \\
\ddot{X}_g &= M_g X_g \\
\ddot{\epsilon}_g &= M_g \epsilon_g = M_g(Y_g - X_g - \gamma_g \iota_g) = \ddot{Y}_g - \ddot{X}_g \beta
\end{aligned}$$

Also recall that $\hat{\beta}$ is the fixed effects estimator of β, and $\hat{V}(\hat{\beta})$ is the CRVE:

$$\begin{aligned}
\hat{\beta} &= (\ddot{X}^T \ddot{X})^{-1} \ddot{X}^T \ddot{Y} \\
\hat{\epsilon}_g &= \ddot{Y}_g - \ddot{X}_g \hat{\beta} \\
\hat{V}(\hat{\beta}) &= (\ddot{X}^T \ddot{X})^{-1} \left(\sum_g \ddot{X}_g^T A_g \hat{\epsilon}_g \hat{\epsilon}_g^T A_g^T \ddot{X}_g \right) (\ddot{X}^T \ddot{X})^{-1}
\end{aligned}$$

where (A_g) are adjustment matrices according to Bell and McCaffrey (2002); for $\hat{V}_{CR0}$, $A_g = I_g$.

Finally, recall the hypothesis H_0 and the test statistic t_0:

$$\begin{aligned}
H_0 &: c_0^T \beta = a_0 \\
t_0 &= \frac{c_0^T \hat{\beta} - a_0}{\sqrt{c_0^T \hat{V}(\hat{\beta}) c_0}}
\end{aligned}$$

A.2 Main Proof

First, I show that the CDF of t_0^2 at a particular quantile q can be written as the CDF at 0 of a linear combination of χ^2 random variables. Second, I show that it is possible to determine the coefficients of that linear combination. Third, I give a formula for $L(q; X, H_0)$.

Since the hypothesis H_0 holds:

$$\begin{aligned} t_0^2 &= \frac{(c_0^T\hat{\beta} - a_0)^2}{c_0^T\hat{V}(\hat{\beta})c_0} \\ &= \frac{(c_0^T(\hat{\beta} - \beta))^2}{c_0^T\hat{V}(\hat{\beta})c_0} \\ &= \frac{c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T\ddot{\epsilon}\ddot{\epsilon}^T\ddot{X}(\ddot{X}^T\ddot{X})^{-1}c_0}{c_0^T(\ddot{X}^T\ddot{X})^{-1}\left(\sum_g \ddot{X}_g^T A_g\hat{\epsilon}_g\hat{\epsilon}_g^T A_g^T\ddot{X}_g\right)(\ddot{X}^T\ddot{X})^{-1}c_0} \end{aligned}$$

Let $H = \ddot{X}(\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T$, let I be an identity matrix of size N, and let $(I-H)_g$ be the rows of $(I-H)$ corresponding to cluster g. Then using the fact that $\hat{\epsilon}_g = \ddot{\epsilon}_g - \ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T\ddot{\epsilon} = (I-H)_g\ddot{\epsilon}$, I continue:

$$\begin{aligned} t_0^2 &= \frac{c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T\ddot{\epsilon}\ddot{\epsilon}^T\ddot{X}(\ddot{X}^T\ddot{X})^{-1}c_0}{c_0^T(\ddot{X}^T\ddot{X})^{-1}\left(\sum_g \ddot{X}_g^T A_g(I-H)_g\ddot{\epsilon}\ddot{\epsilon}^T(I-H)_g^T A_g^T\ddot{X}_g\right)(\ddot{X}^T\ddot{X})^{-1}c_0} \\ &= \frac{\ddot{\epsilon}^T\ddot{X}(\ddot{X}^T\ddot{X})^{-1}c_0c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T\ddot{\epsilon}}{\ddot{\epsilon}^T\left(\sum_g(I-H)_g^T A_g^T\ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}c_0c_0^T(\ddot{X}^T\ddot{X})^{-1}\ddot{X}_g^T A_g(I-H)_g\right)\ddot{\epsilon}} \end{aligned}$$

Now let $d_0 = \ddot{X}(\ddot{X}^T\ddot{X})^{-1}c_0$, and for $g \geq 1$, let $d_g = (I-H)_g^T A_g^T\ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}c_0$. Then it follows that:

$$\begin{aligned} t_0^2 &= \frac{\ddot{\epsilon}^T d_0 d_0^T\ddot{\epsilon}}{\ddot{\epsilon}^T\left(\sum_g d_g d_g^T\right)\ddot{\epsilon}} \\ P(t_0^2 < q \mid X) &= P\left(\frac{\ddot{\epsilon}^T d_0 d_0^T\ddot{\epsilon}}{\ddot{\epsilon}^T\left(\sum_g d_g d_g^T\right)\ddot{\epsilon}} < q \mid X\right) \\ &= P\left(\frac{1}{q}\ddot{\epsilon}^T\left(d_0 d_0^T\right)\ddot{\epsilon} - \ddot{\epsilon}^T\left(\sum_g d_g d_g^T\right)\ddot{\epsilon} < 0 \mid X\right) \\ &= P\left(\ddot{\epsilon}^T\left(\frac{1}{q}d_0 d_0^T - \sum_g d_g d_g^T\right)\ddot{\epsilon} < 0 \mid X\right) \end{aligned}$$

Let $D_+ = [d_0 \quad d_1 \dots d_g \dots d_G]$ and $D_- = [\frac{1}{q} d_0 \quad -d_1 \cdots -d_g \cdots -d_G]$, so that:

$$P(t_0^2 < q \mid X) = P\left(\ddot{\epsilon}^T \left(D_+ D_-^T\right) \ddot{\epsilon} < 0 \mid X\right) \tag{2}$$

Now using Assumption 1, the errors before fixed effects absorption are distributed $\epsilon_g \sim N(0, \sigma^2 \Omega_g)$. Then the errors after absorption, $\ddot{\epsilon}_g = M_g \epsilon_g$, are distributed $\ddot{\epsilon}_g \sim N(0, \sigma^2 \ddot{\Omega}_g)$, where:

$$\begin{aligned}
\ddot{\Omega}_g &= M_g \Omega_g M_g \\
&= \left(I_g - \frac{1}{N_g} \iota_g \iota_g^T\right) \left((1-\rho) I_g + \rho \iota_g \iota_g^T\right) M_g \\
&= \left(I_g - \frac{1}{N_g} \iota_g \iota_g^T\right) \left((1-\rho) I_g\right) M_g \\
&= (1-\rho) \left(I_g - \frac{1}{N_g} \iota_g \iota_g^T\right) M_g \\
\ddot{\Omega}_g &= (1-\rho) M_g M_g
\end{aligned}$$

Then let M be an $(N \times N)$ block-diagonal matrix where the g-th block is M_g. And let η be joint normal with $\eta \sim N(0, I)$, such that $\ddot{\epsilon} = \sigma (1-\rho)^{\frac{1}{2}} M \eta$. Substituting this into (2):

$$\begin{aligned}
L(q; X, H_0) &= P(t_0^2 < q \mid X) \\
&= P\left(\eta^T M \sigma (1-\rho)^{\frac{1}{2}} \left(D_+ D_-^T\right) \sigma (1-\rho)^{\frac{1}{2}} M \eta < 0 \mid X\right) \\
&= P\left(\eta^T M D_+ D_-^T M \eta < 0 \mid X\right)
\end{aligned}$$

At this point, the right-hand side is a function only of known quantities and random variables with known distributions. Neither σ nor ρ appear; the behavior of the test statistic t_0 does not depend on them.

A.3 Calculating $L(q; X, H_0)$ in Practice

I have shown that $L(q; X, H_0) = P\left(\eta^T M D_+ D_-^T M \eta < 0 \mid X\right)$. In this section, I will explain a method for calculating $L(q; X, H_0)$ quickly in practice.

Let $Q_{(N\times N)} = M D_+ D_-^T M$. Then let S be the orthogonal[5] matrix of eigenvectors of Q, let λ^* be an $(N \times 1)$ column vector whose elements are the eigenalues of Q, and let Λ be an $(N \times N)$ diagonal matrix whose diagonal elements are also the eigenvalues of Q. Note that since S is orthogonal, $S\eta \sim \eta$. Then:

$$\begin{aligned} P(t_0^2 < q \mid X) &= P\left(\eta^T S \Lambda S^T \eta < 0 \mid X\right) \\ &= P\left(\eta^T \Lambda \eta < 0 \mid X\right) \end{aligned}$$

Let w be an N-vector of independent random variables such that $\forall i, w_i \sim \chi_1^2$. Then:

$$P(t_0^2 < q \mid X) = P\left(w^T \lambda^* < 0 \mid X\right) \tag{3}$$

Thus, I have shown that the CDF of t_0^2 at q can be written as the CDF at 0 of a linear combination of independent χ_1^2 random variables. Next, I find the non-zero elements of λ^*; it will be the case that λ^* has no more than $G + 1$ non-zero elements.

In principle, the vector of eigenvalues λ^* can be found by eigendecomposing Q. However, since Q is $(N \times N)$, that might be inconvenient in practice. Instead, it is sufficient to find the non-zero eigenvalues of $D_-^T M D_+$, which are the same as the non-zero eigenvalues of $Q = M D_+ D_-^T M$.

To see why, suppose that λ_j is a non-zero eigenvalue of $M D_+ D_-^T M$ correspond-

[5] An orthogonal S can always be found because $Q = M\left(\frac{1}{q} d_0 d_0^T - \sum_g d_g d_g^T\right) M$ is symmetric.

ing to the eigenvector s_j. Note that M is idempotent. Therefore:

$$\begin{aligned} MD_+D_-^TMs_j &= \lambda_j s_j \\ D_-^TMMD_+D_-^TMs_j &= D_-^TM\lambda_j s_j \\ D_-^TMD_+(D_-^TMs_j) &= \lambda_j(D_-^TMs_j) \end{aligned}$$

Thus, λ_j is an eigenvalue of $D_-^TMD_+$ corresponding to the eigenvector $D_-^TMs_j$. And since $D_-^TMD_+$ is a $(G+1\times G+1)$ matrix, it has no more than $G+1$ non-zero eigenvalues. Letting λ be a $(G+1\times 1)$ vector whose elements are the eigenvalues of $D_-^TMD_+$, and in an abuse of notation letting w now be $(G+1\times 1)$, we have that:

$$P(t_0^2 < q \mid X) = P\left(w^T\lambda < 0 \mid X\right) \tag{4}$$

The CDF of a linear combination of independent χ_1^2 random variables $w^T\lambda$ is given by Imhof (1961)[6]:

$$P\left(w^T\lambda < 0 \mid X\right) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{\sin\left(\frac{1}{2}\sum_{j=1}^{G+1}\tan^{-1}\left(\lambda_j u\right)\right)}{u\prod_{j=1}^{G+1}\left(1+\lambda_j^2u^2\right)^{\frac{1}{4}}}du$$

So the CDF of t_0^2 at q can be written as the CDF at 0 of a linear combination of $G+1$ independent χ_1^2 random variables, and it is possible to calculate the coefficients λ as a function of the design matrix X, the hypothesis H_0, and the

[6]This can be calculated quickly by numerical integration, with a high degree of precision, using the *imhof()* function from the R package *CompQuadForm*.

quantile q. In summary:

$$d_0 = \ddot{X}(\ddot{X}^T\ddot{X})^{-1}c_0, \quad d_g = (I-H)_g^T A_g^T \ddot{X}_g(\ddot{X}^T\ddot{X})^{-1}c_0,$$
$$D_+ = [d_0 \quad d_1 \dots d_g \dots d_G], \quad D_- = [\frac{1}{q}d_0 \quad -d_1 \cdots -d_g \cdots -d_G],$$
M block-diagonal, with g-th block M_g,
λ are the eigenvalues of $D_-^T M D_+$, and

$$L(q; X, H_0) = P(t_0^2 < q \mid X)$$
$$= \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{\sin\left(\frac{1}{2}\sum_{j=1}^{G+1}\tan^{-1}(\lambda_j u)\right)}{u\prod_{j=1}^{G+1}\left(1+\lambda_j^2 u^2\right)^{\frac{1}{4}}}du$$

And this is what I set out to find.

My test involves selecting a critical value:

$$q^*(\alpha; X, H_0) = \sqrt{L^{-1}(1-\alpha; X, H_0)}$$

And I reject H_0 if $|t_0| > q^*$. Thus, under Assumption 1, my test is exact, so that the rate at which a true hypothesis is rejected is equal to the nominal size of the test:

$$P\left(|t_0| > q^*\right) = P\left(t_0^2 > (q^*)^2\right) = 1 - L((q^*)^2; X, H_0)$$
$$= \alpha$$

APPENDIX B Asymptotic Validity

Recall the linear model with clustering and cluster-level fixed effects:

$$y_{ig} = x_{ig}\beta + \gamma_g + \epsilon_{ig}$$

Furthermore, recall that $t_0 = \frac{c_0^T(\hat{\beta}-\beta)}{\sqrt{c_0^T \hat{V}_{CR0} c_0}}$. Also, for my test, a critical value $q^*(\alpha; X, H_0)$ is selected such that, by Theorem 1, when Assumption 1 holds and the hypothesis H_0 is true, the test is exact:

$$P(|t_0| > q^*) = \alpha$$

In this section, I show that my test is asymptotically valid without Assumption 1. This result draws on the first theorem from Carter, Schnepel and Steigerwald (2017). I refer to their first assumption as CSS.A1 and their second assumption as CSS.A2, summarized here:

- CSS.A1 ensures that the errors have finite fourth moments
- CSS.A2 ensures that the observations aren't too concentrated in a small number of clusters; for example, the number of clusters $G \to \infty$ as the number of observations in the sample $N \to \infty$

Let $\alpha^* = P\left(|t_0| > q^*(\alpha; X, H_0)\right)$ be the rejection rate of my test – the rate at which my test rejects a true hypothesis H_0.

THEOREM B1: *Suppose that CSS.A1 and CSS.A2 hold and that H_0 is true. Then α^* converges to the nominal test size α:*

$$\alpha^* \xrightarrow{p} \alpha$$

Recall that $\mathbb{E}(\epsilon_{ig}\epsilon_{jg'}) = 0$, so that the errors are uncorrelated across clusters. According to the first theorem in Carter, Schnepel and Steigerwald (2017), it follows that $t_0 \xrightarrow{d} N(0,1)$.

Now, consider a counterfactual data generating process:

$$\tilde{y}_{ig} = x_{ig}\beta + \gamma_g + \tilde{\epsilon}_{ig}$$

where $\tilde{\epsilon} \sim N(0, I)$, so that the counterfactual errors are normal, i.i.d., and homoskedastic. Let $\tilde{t}_0$ be the test statistic that would be generated for H_0 using $\hat{V}_{CR0}$ in the counterfactual data generating process where the errors are $\tilde{\epsilon}$ rather than ϵ:

$$\begin{aligned}
\tilde{\beta} &= \beta + (\ddot{X}^T\ddot{X})^{-1}\ddot{X}^T M\tilde{\epsilon} \\
\hat{\tilde{\epsilon}} &= \ddot{Y} - \ddot{X}\tilde{\beta} \\
\tilde{V} &= (\ddot{X}^T\ddot{X})^{-1}\left(\sum_g \ddot{X}_g^T\hat{\tilde{\epsilon}}_g\hat{\tilde{\epsilon}}_g^T\ddot{X}_g\right)(\ddot{X}^T\ddot{X})^{-1} \\
\tilde{t}_0 &= \frac{c_0^T(\tilde{\beta} - \beta)}{\sqrt{c_0^T\tilde{V}c_0}}
\end{aligned}$$

Since $\tilde{\epsilon}$ meets the conditions of CSS.A1 and CSS.A2, the first theorem in Carter, Schnepel and Steigerwald (2017) applies, so $\tilde{t}_0 \xrightarrow{p} N(0,1)$. And whereas $\tilde{\epsilon}$ also meets the conditions of Assumption 1, it is also the case that Theorem 1 applies, so that $P\left(|\tilde{t}_0| \geq q^*(\alpha; X, H_0) \mid X\right) = \alpha$.

Note that for the case of stochastic X with PDF $f(.)$, it follows from Theorem 1 and the law of total probability that:

$$\begin{aligned}
P(|\tilde{t}_0| \geq q^*(\alpha; X, H_0)) &= \int P(|\tilde{t}_0| \geq q^* \mid X) f(X) dX \\
&= \int \alpha f(X) dX \\
&= \alpha
\end{aligned}$$

Therefore:

$$\begin{aligned}
\alpha^* - \alpha &= P\left(|t_0| \geq q^*\right) - P\left(|\tilde{t}_0| \geq q^*\right) \\
&= \left(1 - P\left(t_0 < q^*\right) + P\left(t_0 < -q^*\right)\right) \\
&\quad - \left(1 - P\left(\tilde{t}_0 < q^*\right) + P\left(\tilde{t}_0 < -q^*\right)\right) \\
&\xrightarrow{p} \left(\Phi(-q^*) - \Phi(q^*)\right) - \left(\Phi(-q^*) - \Phi(q^*)\right) \\
\alpha^* - \alpha &\xrightarrow{p} 0 \\
\alpha^* &\xrightarrow{p} \alpha
\end{aligned}$$

where $\Phi(.)$ is the CDF of the standard normal distribution. In summary:

1) The true test statistic t_0 and the counterfactual test statistic $\tilde{t}_0$ converge to the same distribution.

2) My test is exact for the counterfactual DGP.

3) My test converges to an exact test for any DGP meeting CSS.A1 and CSS.A2.

APPENDIX C Approximating t_0 as $T(v)$

In this section, I will discuss how previous tests have selected critical values for the test statistic t_0 by approximating its distribution as $T(v)$, a t-distribution with v degrees of freedom. Bell and McCaffrey (2002) use $T(m)$, where m is the degrees of freedom from a Satterthwaite approximation, and Carter, Schnepel and Steigerwald (2017) use $T(G^*)$, where G^* is the effective number of clusters. The purpose of this section will be to make sense of the assumptions and simplifications that are necessary to rationalize those tests in a finite sample.