ISSN 1936-5349 (print)
ISSN 1936-5357 (online)

HARVARD

JOHN M. OLIN CENTER FOR LAW, ECONOMICS, AND BUSINESS

ON INFERENCE WHEN USING STATE CORPORATE LAWS FOR IDENTIFICATION

Holger Spamann

Discussion Paper No. 1024

12/2019

Harvard Law School
Cambridge, MA 02138

This paper is also Discussion Paper 2019-12 of the
Harvard Law School Program on Corporate Governance

On Inference When Using State Corporate Laws for Identification

Holger Spamann *

Harvard Law School and Wissenschaftskolleg zu Berlin

December 5, 2019

Abstract

A popular research design identifies the effects of corporate governance by (changes in) state laws, clustering standard errors by state of incorporation. Using Monte-Carlo simulations, this paper shows that conventional statistical tests based on these standard errors dramatically overreject: in a typical design, randomly generated "placebo laws" are "significant" at the 1/5/10% level 9/21/30% of the time. This poor coverage is due to the extremely unequal cluster sizes, especially Delaware's concentration of half of all incorporations. Fixes recommended in the literature fail, including degrees-of-freedom corrections and the cluster wild bootstrap. The paper proposes a permutation test for valid inference.

Keywords— Anti-Takeover Laws, Corporate Governance, Cluster-Robust Inference, Monte Carlo, Placebo Laws, Permutation Test
JEL Classification— C12, G34, G38, K22

*hspamann@law.harvard.edu. For very helpful comments and suggestions, I thank Bobby Bartlett, Emiliano Catan, Jonah Gelbach, James G. MacKinnon, Justin McCrary, Matt Webb, and Alwyn Young.

1 Introduction

In the United States, many important corporate governance features are laid down in state laws. A large empirical literature in corporate finance studies the effects of (changes in) these laws on corporate actions and performance in a firm-level (difference-in-difference) framework. Interest is in the laws for their own sake or, more commonly, as exogenous variation in general economic determinants such as managerial slack. In 2018, Karpoff and Wittry counted 78 published articles and working papers using changes in anti-takeover laws alone; this list has kept growing rapidly.[1] The standard design is a linear firm-level panel regression with firm fixed effects and a variety of possibly confounding time-varying factors. Following Bertrand, Duflo, and Mullainathan 2004, it has become standard econometric practice in these papers to cluster the standard errors by state of incorporation before performing inference using the normal distribution or the t-distribution with degrees of freedom equal to the number of clusters minus 1.[2]

This paper shows, however, that this conventional approach to inference dramatically overrejects in this setting. In a typical difference-in-difference setting with real data, tests of randomly generated "placebo laws" reject the true null hypothesis of no effect at the 1/5/10% level *at least* 3/9/16% of the time, depending on the number of "treated" states and using the popular Tobin's q as dependent variable. The mean (median) rejection rate across numbers of "treated" states and whether Delaware is "treated" is 9/21/30% (8/19/28%). Simulations show that this poor coverage is due to the extremely unequal cluster sizes. In simulated cross-sectional data, severe overrejection occurs if and only if clusters are of very uneven size, especially when one cluster contains half the sample like Delaware, where half of all U.S. firms are incorporated. Indeed, while the nominal number of clusters (states) is 51, in the real data the feasible effective number of clusters (Carter, Schnepel, and Steigerwald 2017) is only about 2. Fixes commonly recommended in the literature fail, including degrees-of-freedom corrections and the cluster wild bootstrap. This paper proposes an exact permutation test for valid inference similar to DiCiccio and J. P. Romano 2017; MacKinnon and Webb 2019a.

The present paper is similar in spirit to Bertrand, Duflo, and Mullainathan 2004, Petersen 2009, and others who use Monte Carlo simulations to demonstrate the practical importance of properly accounting for serial and cross-sectional correlation in the error term. When the number of clusters is above 42 or 50, as in regressions using U.S. state laws, the standard advice (e.g., Bertrand, Duflo, and Mullainathan 2004; Petersen 2009; Angrist and Pischke 2008) is to use the clustered "sandwich" variance estimator (White 1984; Liang and Zeger 1986). MacKinnon and Webb 2017 show that this approach fails when cluster sizes are unequal, the more so the further away the fraction of treated clusters is from $\frac{1}{2}$, and instead point to the cluster wild bootstrap proposed by Cameron, Gelbach, and Miller 2008 as the

[1] See, e.g., Bharath and Hertzel 2019; He and Hirshleifer 2019; Demiroglu, Iskenderoglu, and Ozbas 2019; Gutiérrez Urtiaga and Vazquez 2019; Cremers, Guernsey, and Sepe 2019.

[2] On the necessity to cluster by state of incorporation, see section 2.2 below.

solution. With one exception, however, none of the prior literature consider cluster size imbalance as extreme as that in the corporate governance context: More than half of all publicly traded U.S. corporations are incorporated in Delaware, whereas, e.g., MacKinnon and Webb 2019a considered even 19% to be "quite extreme" for the largest cluster. When one cluster contains half the observations, both the sandwich estimator and the cluster wild bootstrap spectacularly fail to control size, as shown by Monte Carlo evidence with a continuous regressor in Djogbenou, MacKinnon, and Nielsen 2019 and for the cluster treatment assignment model in this paper (even if the fraction of treated states is exactly $\frac{1}{2}$). Theory explaining this failure is provided in Carter, Schnepel, and Steigerwald 2017, MacKinnon and Webb 2017, and Djogbenou, MacKinnon, and Nielsen 2019. The latter papers are part of an active theoretical econometric literature on inference with clustered data, relevant parts of which will be reviewed in sections 5 and 6. More specifically, this paper's proposal of a permutation test is related to a recent surge in interest in permutation tests for regression coefficients (DiCiccio and J. P. Romano 2017), including in clustered regression (Canay, J. P. Romano, and Shaikh 2017; Hagemann 2019; MacKinnon and Webb 2019a).

The present paper is also closely related and complementary to Karpoff and Wittry 2018's simulation tests of omitted variable bias in the study of state antitakeover laws. Taking as given the actual distribution of these laws and various firm characteristics (input data), they simulate output data under various assumptions about which laws and firm characteristics have an effect. They then investigate rejection rates for laws that actually have no effect in their data generating process while purposefully *not* controlling for the other features that do. They find that these tests reject the true null of no effect much more frequently than nominal test size. They explain that their findings illustrate omitted variable bias, which arises when relevant *correlated* regressors are omitted, because the various state laws and other features are correlated in the real input data. By contrast, the present paper shows that standard tests suffer from serious problems even with *uncorrelated* regressors. In other words, while Karpoff and Wittry 2018 is about regression specification and estimation, the present paper is about inference. Consistent with this, controlling for Karpoff and Wittry 2018's state law dummies in the placebo law tests only slightly reduces overrejection.

The rest of this paper is structured as follows. Section 2 explains in more detail the popular firm-level linear regression design with state corporate law as the key independent variable, focusing on the design's difference-in-difference panel variant. Section 3 presents the first set of Monte Carlo evidence, namely results from regressing real firm level outcomes on fake Placebo laws in the difference-in-difference setup. Section 4 digs deeper into the inference issue using simulated cross-sectional data to show that the problem originates in the radically different cluster sizes combined with state level disturbances rather than anything specific to the difference-in-difference setup or uncontrolled features in the real data. Section 5 shows that several fixes proposed in the literature, including degrees-of-freedom corrections (Carter, Schnepel, and Steigerwald 2017; Young 2016) and the cluster wild bootstrap (Cameron, Gelbach, and Miller 2008; MacKinnon and Webb 2017),

fail in the setting under consideration. Section 6 proposes an exact permutation test as a solution to the inference problem, similar to DiCiccio and J. P. Romano 2017; MacKinnon and Webb 2019a. Section 7 concludes with a warning about power.

2 The Typical Study Design

2.1 Estimated Equation and Estimator

A typical study estimates an equation of the following type:

$$y_{ij\ldots st} = \alpha_i + \beta D_{st} + \delta' \mathbf{x_{it}} + \gamma' \mathbf{z_{j\ldots st}} + \epsilon_{ij\ldots st} \tag{1}$$

where β is the coefficient of interest, and

- the subscripts are $i \in \{1, \ldots, I\}$ for firms, $j \in \{1, \ldots, J\}$ for industries, $s \in \{1, \ldots, S\}$ for incorporation states, $t \in \{1, \ldots, T\}$ for time (years), and "$\ldots$" in a subscript stand for further possible groupings (e.g., location of firm headquarters);
- D_{st} is a dummy variable for the "treatment," i.e., whether state s, where firm i is incorporated, has the provision in question in year t (usually, the dummy switches on in some year t^* and stays on for the remainder of the sample period $t \geq t^*$ in that state);
- $y_{ij\ldots st}$ is the outcome variable (e.g. Tobin's q) for firm i in year t;
- α_i is a fixed effect for firm i (which would be constrained to $\alpha_i = \alpha \forall i$ in the cross-sectional variant);
- $\mathbf{x_{it}}$ is a vector of firm-level controls (necessarily time-varying in the panel variant) such as size and leverage;
- $\mathbf{z_{j\ldots st}}$ is a vector of industry etc. level controls that usually includes industry and often includes other features such as other state laws (in the panel variant, these controls are again necessarily time-varying and usually contain a set of year-specific fixed effects, such as industry-year fixed effects); and
- $\epsilon_{ij\ldots st}$ is a firm-year specific error term (which is allowed to be correlated within firm and within incorporation state, as discussed below).

If $T = 1$ and hence $t = 1$ constant for all observations, the equation (1) collapses to a cross-section and one must assume $\alpha_i = \alpha \forall i$. The panel variant with $T > 1$ is more popular by far because it allows controlling for unobserved heterogeneity in α_i. Much of the subsequent discussion will therefore focus on the panel variant. That said, the cross-sectional variant will be used in the simulations because this facilitates computation.

Equation (1) is usually estimated using the fixed effect (FE) estimator (presumably because the dependent variable is assumed to react too slowly and unpredictably to use the first or other difference estimator). That is, one estimates by OLS the demeaned equation:

$$\ddot{y}_{ij\ldots st} = \beta \ddot{D}_{st} + \delta' \ddot{\mathbf{x}}_{\mathbf{it}} + \gamma' \ddot{\mathbf{z}}_{\mathbf{j}\ldots\mathbf{st}} + \ddot{\epsilon}_{ij\ldots st} \tag{2}$$

where double dots denote time-specific deviations from the firm-specific mean for firm i. By the Frisch-Waugh-Lovell theorem, regressions with the demeaned data yield the same result as regressions with the raw data and fixed effects α_i.

The investigations below do not include $\mathbf{x}_{\mathbf{it}}$ in any regressions and focus instead on the pruned equation

$$y_{ij\ldots st} = \alpha_i + \beta D_{st} + \gamma' \mathbf{z}_{\mathbf{j}\ldots\mathbf{st}} + \epsilon_{ij\ldots st} \tag{3}$$

and its demeaned counterpart

$$\ddot{y}_{ij\ldots st} = \beta \ddot{D}_{st} + \gamma' \ddot{\mathbf{z}}_{\mathbf{j}\ldots\mathbf{st}} + \ddot{\epsilon}_{ij\ldots st}. \tag{4}$$

The main reason to focus on the pruned equation is computational simplicity, and the omission of $\mathbf{x}_{\mathbf{it}}$ has no other effect on the simulations by construction.[3] That said, there is also a compelling econometric reason not to control for $\mathbf{x}_{\mathbf{it}}$ in the placebo law tests with real data: time-varying firm-level controls are bound to violate the strict exogeneity assumption $\mathbf{E}(\epsilon_{ij\ldots st}|\mathbf{x}_{\mathbf{i}\tau}) = 0 \,\forall\, t, \tau \in \{1, \ldots, T\}$ (which is required for consistent estimation of (1) or (2), see, e.g., Wooldridge 2010).[4] Not controlling for $\mathbf{x}_{\mathbf{it}}$ cannot create omitted variable bias here because the placebo laws are orthogonal to firm characteristics by construction.[5] For the same reason–orthogonality by construction–the tests of placebo laws need not worry about another violation of strict exogeneity recently pointed out by Karpoff and Wittry 2018, which is that developments at a small number of firms triggered many state law changes (usually after heavy lobbying by the firm).

[3] Even without firm-specific controls, the number of regressors tends to be large because of the many group-time interactions contained in $\mathbf{z}_{\mathbf{j}\ldots\mathbf{t}}$ (conventionally group-year specific fixed effects). The computational challenge is solved using the `Stata` package `reghdfe` of Correia 2016.

[4] The reason is that the outcome component $\epsilon_{ij\ldots st}$ in one period (e.g., a loss if the outcome is profitability) is bound to affect time-varying firm-specific characteristics $\mathbf{x}_{\mathbf{i}\tau}$ in future periods $\tau > t$ mechanically (e.g., leverage) or via endogenous firm adjustment (e.g., a reduction in investment) or even to anticipate next period's characteristics (in particular, if the outcome is or contains the stock price, such as Tobin's q). In a test of an actual law rather than a placebo law, a related concern would be that firm-specific time-varying characteristics may themselves be affected by the law and hence may soak up some of the effect of interest – they are "bad controls" (cf. Angrist and Pischke 2008).

[5] For tests of actual laws, one might worry that states' adoption of the laws is correlated with time-varying firm-level characteristics $\mathbf{x}_{\mathbf{it}}$, such that omission of $\mathbf{x}_{\mathbf{it}}$ would create omitted variable bias. If states reacted to changes in "their" firms, however, one would probably also have to expect them to react to $\epsilon_{ij\ldots st}$, such that controlling for $\mathbf{x}_{\mathbf{it}}$ would not remove all bias (cf. the discussion in the next sentence of the main text).

The placebo laws' orthogonality by construction also removes any potential for omitted variable bias from not controlling for variation in state laws *other than* the particular law under study, important though this would be in a test of a real law (Coates 2000; Catan and Kahan 2016; Cain, McKeon, and Solomon 2017; Karpoff and Wittry 2018). Nevertheless, controlling for other state laws can reduce noise and, most importantly, within-state correlation of the errors, which could improve inference. As will be seen in section 3, however, controlling for other state laws has almost no effect on the placebo law results. The simulated data in section 4 are by construction unaffected by any confounding effects of other "laws."

2.2 Inference

As already mentioned, it is now standard practice to cluster the standard errors by state after estimating (4), i.e., to account for the likely non-zero covariance between residuals for the same firm over time and for multiple firms within the same state of incorporation. This is indicated because the treatment assignment is clustered, i.e., perfectly correlated within states (Abadie et al. 2017). Ignoring the clustered treatment assignment would be harmless if errors were uncorrelated within clusters. Errors are virtually guaranteed to be correlated within clusters, however, since a myriad of state court decisions and statutory amendments affect all or many firms within the state simultaneously. Indeed, some papers identify individual decisions or amendments and exploit them for identification (e.g., Cohen and Wang 2013; Cain, McKeon, and Solomon 2017). But one cannot hope to identify and to control for all possibly relevant state-level shocks. Empirically, unreported placebo law tests with firm-level clustering show even worse overrejection than that reported with state-level clustering in Section 3, controls for the five second-generation anti-takeover statutes from Karpoff and Wittry 2018 notwithstanding.

The usual way to cluster is the "sandwich" estimator of the coefficient variance matrix (cf. White 1984; Liang and Zeger 1986):

$$\hat{\mathbf{V}} \equiv (\mathbf{W}'\mathbf{W})^{-\mathbf{1}} \sum_{s=1}^{S} (\mathbf{W}_\mathbf{s}' \hat{\epsilon}_\mathbf{s} \hat{\epsilon}_\mathbf{s}' \mathbf{W}_\mathbf{s}) (\mathbf{W}'\mathbf{W})^{-\mathbf{1}} \tag{5}$$

where $\hat{\epsilon}_\mathbf{s}$ is a column vector of regression residuals for the observations in state s, $\mathbf{W_s}$ is a matrix of covariates for the observations in state s (i.e., each row contains the covariates for one observation $ijst$), and $\mathbf{W} \equiv \left(\mathbf{W_1}' \ldots \mathbf{W_S}'\right)'$.[6] Let $\hat{V}(\hat{\beta})$ denote the estimated variance of $\hat{\beta}$, which is of course the appropriate diagonal element of $\hat{\mathbf{V}}$. For hypothesis tests, focus in the literature and in this paper is on the resulting t-statistic

$$\hat{t} \equiv \frac{\hat{\beta}}{\sqrt{\hat{V}(\hat{\beta})}}. \tag{6}$$

[6] When using the FE estimator, each row of $\mathbf{W_s}$ will be of the form $\left(\ddot{D}_{st}, \ddot{\mathbf{z}}'_{\mathbf{j}\ldots\mathbf{st}}\right)$. However, in the simulations using simple cross-sectional data, each row will be of the form $\left(D_{st}, \mathbf{z}'_{\mathbf{j}\ldots\mathbf{st}}\right)$, where $t = 1$ for all observations.

The justification for the "sandwich" estimator is asymptotic normality of $\hat{t}$. Nevertheless, it is standard to use critical values from a t-distribution and to apply an adjustment factor to $\hat{\mathbf{V}}$ to correct for finite sample bias (see Donald and Lang 2007 for intuition and an exact result under more restrictive assumptions). In particular, `Stata` multiplies $\hat{\mathbf{V}}$ by $\frac{N-1}{N-k} \times \frac{S}{S-1}$, where k is the number of regressors excluding nested fixed effects, and uses critical values from the t-distribution with $S-1$ degrees of freedom. This paper does so too.

Carter, Schnepel, and Steigerwald 2017, B. E. Hansen and S. Lee 2019, and Djogbenou, MacKinnon, and Nielsen 2019 generalize earlier results to show that $\hat{\mathbf{V}}$ is a consistent estimator of the true variance and the resulting t-statistic asymptotically normal even with heterogeneous clusters of unequal sizes, as in the current setting. They also show, however, that this requires not only $S \to \infty$ but also that the share of the observations in the largest cluster asymptotically vanishes. In the present setting, the latter assumption could be questioned because Delaware's preeminence seems to be a fixture of the setting even if one were to imagine the number of states to grow beyond bound. In any event, the greater practical concern is that the rate of convergence with unequal cluster sizes may be much slower than the raw number of clusters would lead one to expect. Carter, Schnepel, and Steigerwald 2017 introduce the feasible effective number of clusters as a guide to this behavior. In the setting considered in the following Section 3 where the nominal number of clusters is 51, the feasible effective number of clusters for testing the effects of one of the five modern anti-takeover statutes of Karpoff and Wittry 2018 is between 1.3 and 3.2. One should thus expect the variance estimate $\hat{\mathbf{V}}$ to be very poor and the resulting t-statistic $\hat{t}$ to be very erratic. While this feasible effective number of clusters is based on a worst-case assumption, the placebo Monte Carlo simulations in the next Section will show that the usual tests do indeed perform extremely poorly.

3 Placebo Laws

To demonstrate that the conventional "sandwich" clustered standard error approach fails when studying the effect of state corporate law changes on corporate outcomes, this section studies the "effect" of *random* "Placebo" laws in a typical data set. Even though the Placebo laws are random and hence have no real effect by construction, the conventional tests reject the null of no effect at rates far higher than their nominal level. The mean (median) rejection rate across specifications is 9/21/30% (8/19/28%). That is, the type I error rate is far higher than the chosen test size would make one believe.

The base data are all firm-year observations of US-listed firms from the CRSP/Compustat merged database for the years 1983-2018 excluding financials and utilities and ADRs. The dependent variable y_{ijst} is 95%-winsorized Tobin's q for firm i in year t.[7] The control variables $\mathbf{z_{jst}}$ always include industry-year jt fixed effects (using the Fama-French 49 industry coarsening of the SIC classification), and in one group of

[7] Tobin's q is constructed as (total assets + market equity - book equity) di-

simulations also include year-state indicator dummies for adoption of each the five second-generation anti-takeover statutes from Karpoff and Wittry 2018.[8] Standard errors are clustered by 51 "states" s of incorporation (the 50 U.S. states and the District of Columbia). As incorporation state is merely a current header variable in CRSP/Compustat but corporations may change their state of incorporation over time, firms' historical incorporation states were scraped from SEC filings using the SEC's EDGAR database, which reaches back to about 1995; for earlier years, incorporation was backfilled from the latest information available. The years have been chosen to avoid complications from earlier generation anti-takeover statutes (cf. Karpoff and Wittry 2018), and Tobin's q has been chosen because of its widespread use in the corporate governance literature (in spite of important theoretical reservations[9]). In any event, beyond illustration, nothing hinges on the choice of variables and time frame, as the results are qualitatively similar with completely artificial data as shown in Section 4.

Starting from the base data just described, the Monte Carlo simulations reported below then generate random "Placebo" laws (i.e., a dummy variable equalling one for treated state-years) enacted in random states in random years (once enacted, they stay "in effect" throughout the sample period, as is usually the case in the real world). The simulations are run separately 2,000 times for each combination of (a) every number of treated states between 1 and 51, (b) whether or not Delaware (DE) is among the treated states (if it is not generally among the treated, it nevertheless will be once the number of treated states is 51), and (c) whether anti-takeover statute dummies from Karpoff and Wittry 2018 (KW) are included as controls. For each run, the requisite number of "treated" states are picked at random (subject to ensuring that Delaware is or is not treated, as the case may be), and then a first treatment year is drawn independently for each of the treated states from a uniform discrete distribution over all sample years. (Unreported results forcing the first treatment year to be year 2 or beyond look virtually identical.) Equation (4) is then estimated using OLS, the t-statistic calculated as in (6) using (5), and then compared to critical values from a t-distribution with 50 degrees of freedom. Rejection rates are then calculated for each combination of treated number, Delaware status, and controls.

[Figure 1 about here.]

vided by total assets, and market and book equity are as defined on Ken French's website https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/variable_definitions.html.

[8]The 49 industry definitions are from http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ftp/Siccodes49.zip. Karpoff and Wittry 2018's five second-generation anti-takeover statutes are control share acquisition laws, business combination laws, fair price laws, directors' duties laws, and poison pill laws.

[9]With firm fixed effects, Tobin's q's scaling of market capitalization by assets is superfluous when assets are fixed, creates unnecessary noise when assets change due to accounting maneuvers, and leads to absurd results when assets change because the firm is growing. Cf. Bartlett and Partnoy 2019.

Figure 1 graphs the results. Here and in all subsequent graphs, the vertical axis has been rescaled by taking square roots of the rejection rates, and horizontal lines drawn at 1%, 5%, and 10%, for more tailored graphical comparison of the tests' nominal size to their empirical size.

As previously mentioned, the conventional test using (6) grossly overrejects. When Delaware is not among the treated states–a frequent case in practice, depicted in the top panels–, the nominal 1/5/10% tests behave virtually like 10/20/30% tests – slightly better when controlling for the five second-generation anti-takeover statutes of Karpoff and Wittry 2018 (KW), and slightly worse when not. The *minimum* rejection rate of the nominal 1% test when Delaware is not treated is 6% when controlling for KW's statutes, and 8% when not.[10] The *median* rejection rates of the nominal 1% test when Delaware is not treated are 8% when controlling for KW's statutes and 11% when not. The median rejection rates of the nominal 5/10% tests are 19/28% when controlling for KW's statutes and 23/32% when not.

Overrejection is less extreme when Delaware is among the treated states (bottom panels) *and* the number of treated states is relatively high. Still, even when Delaware is treated, median (minimal) rejection rates of nominal 1/5/10% tests are 5/16/24% (3/9/16%) when not using KW controls. With KW controls, overrejection is a little less extreme at low numbers of treated states but otherwise very similar.

Clearly, conventional cluster-robust inference fails spectacularly in this setting. To probe into the source of the problem, one could run the placebo tests for a cross-section rather than a panel. Rather than doing this with real data, the next section will simulate cross-sectional data sets to show that the problem originates in the extremely unequal cluster sizes.

4 Simulated Data

This section shows results using simulated data. It does so to abstract from some details of the real data. In particular, the simulated data demonstrate that the overrejection found in the previous section is not due to an unfortunate choice of variables including lack of suitable controls. The simulations are cross-sectional to reduce computational burden and to demonstrate that unequal cluster sizes, not the panel structure, are the source of the problem.

The data generating process builds on the following scaffolding of $S = 51$ "states" and $N = 5100$ "firms" distributed across these states. In the baseline "balanced" specification, firms are distributed evenly across states, i.e., each state has 100 firms. In the "unbalanced" condition, each state has 51 firms except one state with 2,550 firms; this large state thus has half the sample and will be called "Delaware." In the even "more unbalanced" condition, "Delaware" still has 2,550 firms but the 2,550 remaining firms are not distributed evenly among the remaining 50 states but in a linearly increasing pattern from 2 through 100 observations.

[10] Delaware not being treated excludes the case of 51 treated clusters in the top panels, when the rejection rates of the 1% test drop to 3.7 and 3.5%, respectively.

In each simulation, firms' "outcomes" y_{ijs1}, "industry" groups z_{j1}, and treatment status D_{is1} are assigned randomly and independently from one another. Thus they are in expectation uncorrelated with one another by construction; neither affects the other. Concretely, each firm's outcome variable y_{is1} is drawn as the sum of two normal random variables: one specific to firm i, and another common to all firms in state s. That is, the first component is drawn independently for each firm i, while the second component is perfectly correlated for all firms in state s. The variance of the first, idiosyncratic component is set to nine times the variance of the state-wide component, such that the intra-state correlation of the outcome variable is 0.1. Independently of their state and outcome variable, firms are randomly grouped into 50 equal-sized "industry" groups z_{j1}. "Treatment" status $D_{s1} = 1$ is independently assigned to all firms in randomly chosen S_1 "treated" states and set to $D_{s1} = 0$ for all firms in the remaining $S - S_1$ states. In summary, the data generating process is

$$y_{ijs1} = \mu_{s1} + \eta_{ijs1}, \quad \mu \overset{iid}{\sim} \mathcal{N}(0,1), \quad \eta \overset{iid}{\sim} \mathcal{N}(0,9), \tag{7}$$

and μ, η, D, z are mutually independent.

OLS is then used to estimate

$$y_{ijs1} = \alpha + \beta D_{s1} + \gamma_{j1} z_{j1} + \epsilon_{ijs1} \tag{8}$$

where the the firm-specific intercepts α_i of equation (3) have been replaced by a common intercept α due to the cross-sectional nature of the simulated data (emphasized by the constant time subscript "1"). The state subscript s has been dropped from z to emphasize that the only control variable here is the industry effect, which is orthogonal to states. The t-statistic is then calculated using (5) and (6) and compared to critical values from the t-distribution with 50 degrees of freedom as before. This procedure is run 10,000 times per balance status (balanced, unbalanced, or more unbalanced), number of treated states $S_1 \in \{1, \ldots, 50\}$, and whether "Delaware" is among the treated states.[11] (With cross-sectional data, the cases of S_1 treated states excluding Delaware is equal to the case of $S - S_1$ treated states including Delaware, as will be apparent from the symmetry of the graphs.)

[Figure 2 about here.]

Figure 2 shows the results. In the balanced baseline, tests are not too far from their nominal size, with the well-known exception of overrejection with very few treated or untreated clusters (e.g., Imbens and Kolesár 2016; MacKinnon and Webb 2017). When clusters are unbalanced, however, overrejection is severe regardless of the number of treated clusters. Type 1 error rates of nominal 1/5/10% tests are 6/18/28% even for the most favorable number of treated clusters, and they are higher even for equal numbers of treated and untreated clusters. For example,

[11] Here the highest number of treated states is 50, not 51 as with placebo laws, because the data are cross-sectional, so having 51 treated states would mean that the treatment variable is equal to 1 for all observations.

when the number of treated clusters is 25 but cluster sizes are unbalanced, type 1 error rates of 1/5/10% tests are astonishing 19/42/53%. These results with a binary treatment indicator are comparable to those of Djogbenou, MacKinnon, and Nielsen 2019, fig. 3(a) for the case of a continuous regressor that is correlated within clusters.

Interestingly, overrejection is a little less severe with more unbalanced cluster sizes than with merely unbalanced cluster sizes, i.e., adding heterogeneity to the size of non-"Delaware" clusters actually decreases overrejection. This suggests strongly that the source of the worst problem is the extremely disproportionate size of the "Delaware" cluster, rather than only cluster size heterogeneity per se. This also explains why the overrejection found here is much worse than in the Monte Carlo simulations of MacKinnon and Webb 2017; MacKinnon and Webb 2019a, and in line with that found in the limiting case of Djogbenou, MacKinnon, and Nielsen 2019. Having half the sample in one cluster is too extreme to have been considered in the prior literature (except Djogbenou, MacKinnon, and Nielsen 2019) but an unavoidable feature of firm incorporation patterns.

To generate this pattern of severe overrejection, a necessary feature of the simulated data is the presence of within-state correlation. In unreported simulations where y_{is1} is independently drawn for each firm from a standard normal distribution without the addition of a state-level effect, overrejection is relatively minor. However, this provides little practical reassurance because the very justification for the use of the state-cluster robust variance estimator (5) is that state-level disturbances cannot be ruled out a priori (see Section 2.2).

5 Fixes Recommended in the Literature

There is an active literature in econometric theory attempting to identify and to remedy problems of cluster-robust inference in finite samples, including this paper's problem of unequal cluster sizes (see Cameron and Miller 2015; MacKinnon and Webb 2019b for recent surveys). Some of the approaches developed in that literature are not suitable for the present setting, for example because they require an equal or large number of observations in each cluster (e.g., Donald and Lang 2007; Bester, Conley, and C. B. Hansen 2011; Ibragimov and Müller 2016), because the parameter of interest must be identified within each cluster (e.g., Canay, J. P. Romano, and Shaikh 2017[12]), or because one must at least be able to collect a clearly defined post-treatment indicator from each cluster (Hagemann 2019 (which is not the case

[12]When treatment is assigned at the cluster level and hence not identified within cluster, Canay, J. P. Romano, and Shaikh 2017 implement their method by considering pairs of treated and untreated clusters. In the current setting, however, their postulate of pairs "suggested" by the treatment assignment is not met, such that this implementation seems unappealing.

in the difference-in-difference setting when treatment years vary[13]).[14] There are at least three approaches, however, that could be expected to fix the problem identified in the preceding two sections:

1. While maintaining the overall approach of conventional inference (variance estimate (5), t-statistic (6), critical values from a t-distribution), one could set the degrees-of-freedom of the t-distribution equal to the feasible effective number of clusters of Carter, Schnepel, and Steigerwald 2017 (a method that Carter, Schnepel, and Steigerwald 2017 consider in passing).
2. Still within the framework of conventional inference, one could apply bias and degrees of freedom corrections to the variance estimate (5) so that the distribution of the resulting t-statistic (6) approximates the first two moments of a χ^2 variable, an approach suggested by Young 2016.
3. As an alternative to conventional cluster-robust inference, Cameron, Gelbach, and Miller 2008 propose the cluster wild bootstrap-t, which Djogbenou, MacKinnon, and Nielsen 2019 endorse on the basis of higher-order theory.

Monte Carlo simulations are presented in Young 2016 and Cameron, Gelbach, and Miller 2008 for their respective methods and in Cameron and Miller 2015 and MacKinnon and Webb 2017 for the effective number of clusters and wild bootstrap methods, among others. Based on their Monte Carlo simulations, MacKinnon and Webb 2017 recommend the wild bootstrap for unequal cluster sizes at least when the number of treated clusters is in an intermediate range. In all of these simulations, however, cluster size imbalances are less extreme than for incorporation states. For example, MacKinnon and Webb 2017's largest "wildly different cluster size" is 12% of the observations, whereas Delaware contains over 50%. Only Djogbenou, MacKinnon, and Nielsen 2019, fig. 3 present Monte Carlo evidence with one cluster containing half of all observations and find that various variants of the wild bootstrap fail for inference on a continuous regressor that is correlated within cluster.

To investigate the latter three methods' potential to fix the problem identified in the preceding two sections, this section re-runs the simulations of the preceding section using these three methods. For brevity, exclusive focus is on the "unbalanced"

[13] As in Ibragimov and Müller 2016, fn. 10, one might overcome this problem by considering only years before the first and after the last state adopted the statute in question. Given that at least some adoptions occur many years after others, however, this would either entail very considerable data loss or require dropping very early or late adopters or years. Of course, it might be preferable to restrict the estimates to a narrower window. This leads into broader questions of research design that are beyond the scope of the present paper.

[14] A related technical difference between the first two approaches and those considered below is in the type of asymptotics: the former consider asymptotics $n_s \to \infty$ for fixed S, whereas the latter, including the conventional cluster-robust inference exposited above, consider $S \to \infty$. To the extent $S \to \infty$ asymptotics have been assumed based on the nature of the setting rather than analytical convenience, an appropriate "fix" should retain that assumption.

case where half of the 5,100 firms are located in "Delaware" and the remainder is spread equally among the other 50 states, and specifically on the case where "Delaware" is not among the treated states (which is without loss of generality because, as previously noted, with cross-sectional data the case where "Delaware" is among the treated is equal to the one where it is untreated and the number of treated and untreated states are reversed). The methods are implemented using the `Stata` software modules of C. H. Lee and Steigerwald 2018; Young 2016; Roodman et al. 2019, respectively. The wild bootstrap uses restricted estimates and Rademacher weights in the bootstrap data generation process, since these generally perform better than alternatives in simulations of MacKinnon and Webb 2017; Djogbenou, MacKinnon, and Nielsen 2019; the number of bootstrap replications is set to 999.

[Figure 3 about here.]

Figure 3 shows that none of the three methods solves the present problem.

Starting with the left panel, use of the feasible effective number of clusters for degrees of freedom is too conservative, transforming tests of nominal 1/5/10% size into 0/0/0.5% tests for intermediate numbers of treated clusters. This presumably reflects the fact that Carter, Schnepel, and Steigerwald 2017 developed the feasible effective number of clusters as a lower bound on the effective number of clusters. At the same time, there is still severe overrejection with very few or many treated clusters.

The middle panel shows that Young 2016's approach overrejects for almost all numbers of treated clusters. The overrejection is as severe as with the conventional approach when the fraction of treated clusters is more than half.

The right panel shows that the cluster wild bootstrap performs better than the first two approaches, but still overrejects by a factor of over two for most numbers of treated clusters. Overrejection is less for very small numbers of treated clusters but much worse for high numbers of treated clusters, or more generally, when Delaware's treatment status is not shared by many states. These results are comparable to those of Djogbenou, MacKinnon, and Nielsen 2019, fig. 3(e) for the case of a continuous regressor that is correlated within clusters.

In conclusion, the simulation suggests that all of the available approaches fail when differences in cluster sizes are as extreme as in the study of corporate law changes with firm data.

6 A Permutation Test

This section will suggest a permutation test as a solution to the size problems identified in the previous three sections. The test is an extension to the cluster case of DiCiccio and J. P. Romano 2017 and essentially identical to the RI-t test investigated by MacKinnon and Webb 2019a using simulations. (The only difference between the latter test and the one proposed here is in drawing from the possible permutations without or with replacement, respectively, when their number is large.)

The intuition of the test is a simple inversion of the placebo law logic of the preceding sections. That logic was that a valid test should not find a "significant" result with random placebo laws–a type I error–more frequently than the test's nominal size. While the preceding sections used this logic to criticize conventional tests, it can also be used constructively to formulate a valid test: the null hypothesis should be rejected if and only if the actual test statistic is in the relevant tail of the empirical distribution of equivalent test statistics generated by placebo laws.[15] This section will state this test formally and show that it is exact against the randomization null hypotheses and asymptotically valid against more general null hypotheses.

6.1 Definition of the Test

To state the test formally, let $\mathcal{G}$ be the set of all $S!$ permutations π of $\{1, \ldots, S\}$, let I_s and N_s be the number of firms and firm-years, respectively, in cluster s, and write the observed partially demeaned data as

$$\bigcup_{s \in \{1,\ldots,S\}} \left(\ddot{\mathbf{y}}_s, \ddot{\mathbf{Z}}_s, \mathbf{T}_s, \mathbf{d}_s \right)$$

where for each state s, $\mathbf{y}_s$ is a $N_s \times 1$ vector stacking all the $y_{ij\ldots st}$, $\mathbf{Z}_s$ is a $N_s \times |\mathbf{z}|$ matrix stacking the covariate vectors $\mathbf{z}'_{\mathbf{j}\ldots\mathbf{st}}$, $\mathbf{T}_s$ is a $I_s \times T$ indicator matrix marking years in which the I_s firms are in the sample, and $\mathbf{d}_s$ is a $T \times 1$ vector indicating treatment status for state s in year t. Also write the t-statistic in (6) as an explicit function of the partially demeaned data

$$\hat{t} = \hat{t}\left(\bigcup_{s \in \{1,\ldots,S\}} \left(\ddot{\mathbf{y}}_s, \ddot{\mathbf{Z}}_s, \mathbf{T}_s, \mathbf{d}_s \right) \right). \tag{9}$$

The permutation distribution of the t-statistic is then

$$\hat{R}(\tau) = \frac{1}{S!} \sum_{\pi \in \mathcal{G}} \mathbb{1}_{\tau > \hat{t}\left(\bigcup_{s \in \{1,\ldots,S\}} \left(\ddot{\mathbf{y}}_s, \ddot{\mathbf{Z}}_s, \mathbf{T}_s, \mathbf{d}_{\pi(s)} \right) \right)} \tag{10}$$

where $\mathbb{1}$ is the indicator function. Note that the permutation distribution is conditional on the observed data, and in particular on the number of treated states and the set of treatment years, which are automatically held constant in all permutations by the definition of the test.

Similarly, a (two-sided) permutation p-value for a given t-statistic t can be

[15] By contrast, Hagemann 2019 permutes cluster-specific intercept estimates, not placebo laws (i.e., independent variables) themselves.

calculated as[16]

$$\hat{p}\left(t\right)=\frac{1}{S!}\sum_{\pi\in\mathcal{G}}\mathbb{1}_{|t|\leq\left|\hat{t}\left(\bigcup\limits_{s\in\{1,\ldots,S\}}\left(\ddot{\mathbf{y}}_{s},\ddot{\mathbf{Z}}_{s},\mathbf{T}_{s},\mathbf{d}_{\pi(s)}\right)\right)\right|}. \tag{11}$$

Calculating the t-statistic (6) for all $S!$ permutations would be prohibitively computationally costly when $S=51$, as here. When many clusters have identical treatment patterns (in particular, when many clusters never receive treatment), the computational burden can be substantially reduced by considering only a single permutation of that subset of clusters for every permutation of the others. Nevertheless, the computational burden will still be enormous unless the vast majority of clusters received identical treatment. It is easier to consider a stochastic approximation to the permutation p-value by randomly drawing with replacement[17] B permutations from $\mathcal{G}$ and calculating

$$\tilde{p}\left(t\right)=\frac{1}{B}\left(1+\sum_{b=1}^{B-1}\mathbb{1}_{|t|\leq\left|\hat{t}\left(\bigcup\limits_{s\in\{1,\ldots,S\}}\left(\ddot{\mathbf{y}}_{s},\ddot{\mathbf{Z}}_{s},\mathbf{T}_{s},\mathbf{d}_{\pi_{b(s)}}\right)\right)\right|}\right). \tag{12}$$

Tests of the desired size against a null to be stated precisely below can then be performed by reference to $\tilde{p}$.

To understand the mechanics of the test, note that $\ddot{\mathbf{y}}_s$ and $\ddot{\mathbf{Z}}_s$ each have N_s rows and are demeaned, whereas $\mathbf{d}_s$ has T rows and is not demeaned. This means that $\ddot{\mathbf{y}}_s, \ddot{\mathbf{Z}}_s$ can be used without further transformation in regression algebra to estimate (4) and (5), whereas $\mathbf{d}_s$ cannot. In other words, to compute (6), it is first necessary to construct from $\mathbf{d}_s$ and $\mathbf{T}_s$ a $N_s \times 1$ vector of firm-year specific deviations from the firm-specific treatment mean, which depends of course on $\mathbf{d}_s$ (i.e., when the treatment was applied, if at all). The reason to write the data this way is that the panel is doubly unbalanced. First, N_s differs across states s, such that it would not be possible to switch a vector or matrix of row-dimension N_s from one state to another. Second, and relatedly, firms are in the sample for different years, such that the treatment means may differ even for firms within the same state and the information in $\mathbf{T}_s$ is required to calculate them from $\mathbf{d}_s$. To discuss the mechanics and the validity of the test, it is therefore appropriate to focus on vectors $\mathbf{d}_s$ that are of the same dimension for all states, while making explicit the dependence of the resulting statistic on $\mathbf{T}_s$. At the same time, focusing on the partially demeaned data makes clear that the test need not make any assumptions on the individual fixed effects α_i and their relation to $\mathbf{d}_s$.

[16] This formulation and resulting test is slightly conservative in its treatment of ties. To be perfectly exact, the summand might have to be in $(0, 1)$ for ties (see, e.g., Lehmann and J. P. Romano 2005, ch. 15.2.1). Given a large number of permutations and continuous variables, however, ties will be vanishingly rare and can be safely ignored.

[17] Random draws without replacement are also valid but harder to implement.

6.2 Validity of the Test

Under the randomization null hypothesis defined and discussed in 6.2.1, the permutation p-value $\hat{p}$ (11) is exact even in finite samples. If the randomization hypothesis does not hold (6.2.2) $\hat{p}$ is still likely to be correct asymptotically under the less restrictive null hypothesis of zero average treatment effect ($\beta = 0$); in any event, Monte Carlo simulation evidence suggests coverage is much better than that of the other tests reviewed in this paper. When B is large (say, 100,000 draws), $\tilde{p}$ will be exceedingly close to $\hat{p}$, so the discussion will extend to an implementation using $\tilde{p}$.

6.2.1 Under the Randomization Hypothesis

Under the randomization hypothesis null, the permutation p-value $\hat{p}$ (11) is exact even in finite samples (e.g., Lehmann and J. P. Romano 2005, Theorem 15.2.2). The randomization hypothesis is that the distribution of the data is invariant under any permutation $\pi \in \mathcal{G}$ under the null. Abusing notation by maintaining the same symbols for random variables as for their realization, the randomization hypothesis can be formally stated as

$$\bigcup_{s\in\{1,\dots,S\}} \left(\ddot{\mathbf{y}}_s, \ddot{\mathbf{Z}}_s, \mathbf{T}_s, \mathbf{d}_s\right) \overset{d}{=} \bigcup_{s\in\{1,\dots,S\}} \left(\ddot{\mathbf{y}}_s, \ddot{\mathbf{Z}}_s, \mathbf{T}_s, \mathbf{d}_{\pi(s)}\right) \quad \forall \pi \in \mathcal{G}. \tag{13}$$

In other words, any permutation of the data could have been sampled with equal probability as the actual sample. This requires $\mathbf{d}$ to be independent of $\left(\ddot{\mathbf{y}}, \ddot{\mathbf{Z}}, \mathbf{T}\right)$ at least conditionally on the joint realization $\bigcup_{s\in\{1,\dots,S\}} \left(\ddot{\mathbf{y}}_s, \ddot{\mathbf{Z}}_s, \mathbf{T}_s, \mathbf{d}_s\right)$.

Implicit in the randomization hypothesis is the assumption that the population of U.S. states is drawn from a superpopulation of possible states. This assumption is also implicit in the conventional approach to cluster-robust inference with its key assumption $S \to \infty$ (cf. Abadie et al. 2017).

However, the randomization hypothesis is stronger than the assumptions required for conventional inference in three respects.

First, treatment $\mathbf{d}$ must be conditionally independent of $\ddot{\mathbf{y}}$ or equivalently–given the conditioning on the model (3)–of $\ddot{\boldsymbol{\epsilon}}$. By contrast, the strict exogeneity assumption required for consistency of the FE estimator (and thus conventional inference) is merely a zero conditional mean assumption, $\mathbf{E}(\ddot{\epsilon}_s|\mathbf{d}_s) = \mathbf{0}\,\forall\, s$. Both assumptions rule out reverse causation where higher or lower realizations of $\ddot{\boldsymbol{\epsilon}}$ in some period trigger adoption of the treatment. But only the stronger independence assumption rules out heteroskedasticity, in particular the possibility that treatment induces a greater (or lesser) dispersion of outcomes. This assumption will be met by a sharp null hypothesis of no treatment effect whatsoever (as in the popular Fisher exact test), but not by a null of zero *average* treatment effect. (Note, however, that treatment effect heterogeneity across clusters might also bias the FE estimator (Gibbons, Serrato, and Urbancic 2018).)

Second, treatment $\mathbf{d}$ must be conditionally independent of the control variables $\ddot{\mathbf{Z}}$. By contrast, conventional inference makes no assumption on the relation between

treatment and $\ddot{\mathbf{Z}}$. Independence is not an innocuous assumption since, e.g., $\ddot{\mathbf{Z}}$ includes industry-year fixed effects and states may be more prone to adopt certain statutes to protect "their" industries at certain times. Then again, Karpoff and Wittry 2018 show that adoption of many anti-takeover statutes was triggered by a single firm and thus arguably random for other firms.

Finally and most subtly, the randomization hypothesis asserts that $\mathbf{d}$ is conditionally independent of the composition of the cluster, namely the number of firms in the cluster and the years that they appear in the sample, as recorded in $\mathbf{T}$. The latter is less of a concern in the sense that if there were dependence between treatment and firm years in the sample (in particular, survivorship or selection bias), then the FE estimator might not even be consistent, and inference would be a secondary concern. Turning to number of firms in the cluster, there is some reason to think it might not be unrelated to the distribution of $\mathbf{d}$. R. Romano 1993 and others have argued that Delaware is much less likely to adopt some bad statutes and more likely to adopt some good statutes for political economy reasons, given the importance of its incorporation business to its economy and its budget. This would argue for considering only permutations π that maintain Delaware's treatment status, as is done in the simulations below in one of the two versions of the test. MacKinnon and Webb 2019a, section 3.2 also argue for conditioning the permutation on cluster sizes, but their argument is based on the different perspective of maintaining test size conditional on the observed joint distribution of treatment assignment and cluster sizes (which has well-known mechanical consequences for the distribution of the resulting test statistic).[18]

In summary, the randomization hypothesis (13) is arguably as plausible as the sharp null hypothesis of the Fisher exact test, at least if one conditions on Delaware's treatment status and is willing to assume that treatment assignment was unrelated to $\ddot{\mathbf{Z}}$. If the randomization hypothesis does hold, then the permutation p-values $\hat{p}$ (11) and associated test specified above are exact even in finite samples.

6.2.2 Under the Null of Zero Average Treatment Effect

If the randomization hypothesis does not hold, then the permutation p-values $\hat{p}$ (11) for the t-statistic (6) and associated test may still be asymptotically valid against the less strict null hypothesis of zero average treatment effect ($\beta = 0$). DiCiccio and J. P. Romano 2017, Theorem 3.3 prove this for permutation of a subset of regressors in a simple cross-sectional regression. While their proof cannot be easily extended to the unbalanced panel cluster setting, they also review other settings

[18] In this context, MacKinnon and Webb 2019a also discuss the further question how to deal with implementation years in the panel context. After all, one could also permute implementation years, within and/or across states. It seems preferable to hold the set of implementation years fixed, however, for similar reasons as the number of treated states. Unconditionally, errors might have unequal variance over time. Conditionally, the distribution of treatment years will influence the distribution of the test statistic through its effect on the implied cluster weights. In any event, researchers should assess and discuss the sensitivity of their results to this choice.

where studentizing the test statistic makes a permutation test asymptotically valid. Since the t-statistic (6) is studentized, it is reasonable to conjecture that it will be asymptotically valid.

In any event, finite sample performance is much more important than asymptotic validity. After all, the whole point of this paper thus far has been to show that conventional inference spectacularly overrejects in finite samples in spite of asymptotic consistency of the conventional variance estimator. To investigate the finite sample performance of the permutation test, we can resort to more Monte Carlo simulations. To gain insight about the behavior of the permutation test when the randomization hypothesis does not hold, the data generating process (7) from section 4, for which the randomization hypothesis does hold, needs to be modified. Of the infinite possible modifications, only one form of heteroskedasticity is simulated here because of the simulations' high computational cost (the usual cost of simulations multiplied by B); further simulations should be tailored to concrete applications.

Here another standard normal variable ξ constant within state is added to the data generating process only for treated observations, such that the error variance is about 10% higher and the intra-state correlation about double in treated relative to untreated states. The modified data generating process is now

$$y_{ijs1} = \mu_{s1} + \eta_{ijs1} + D\xi_{s1}, \quad \mu, \xi \overset{iid}{\sim} \mathcal{N}(0,1), \quad \eta \overset{iid}{\sim} \mathcal{N}(0,9), \tag{14}$$

with μ, η, ξ, D, z mutually independent.

As before, OLS is then used to estimate (8) and the t statistic calculated using (5) and (6). Unlike before, however, the critical values for each test are now taken from the stochastic approximation to the randomization distribution (10). For 1 or 50 treated clusters, the full randomization distribution is derived based on all possible permutations of the treatment indicator (49 or 50, depending on whether "Delaware"'s treatment status is preserved). For other numbers of treated clusters, the stochastic approximation to the randomization distribution with $B = 1,000$ is used. ($B = 1,000$ is too low for the precision desired in an actual empirical test but good enough for probing the properties of the test.) Delaware is never treated, but this is without loss of generality because with cross-sectional data, the cases of S_1 treated clusters including Delaware are equivalent to $51 - S_1$ treated clusters excluding Delaware, and vice versa. As before, the simulation is run 10,000 times per number of treated clusters.

[Figure 4 about here.]

Figure 4 shows the result of the simulations using heteroskedastic data. It shows separate sets of results for the (stochastic approximation) of the permutation test (top left panel), the variant of the permutation test that only uses permutations π maintaining Delaware's treatment status (top right), and, for comparison, the conventional test using the t-distribution with $S - 1$ degrees of freedom (bottom left) and the cluster wild bootstrap (bottom right). The case of 50 treated clusters

is omitted from the top right panel because no permutation preserves Delaware as untreated when only Delaware is untreated in the realized data.

Clearly, neither variant of the permutation test is perfect. That said, both variants are much truer to nominal size than the cluster wild bootstrap and especially the conventional test, which grotesquely overrejects as before. The basic permutation test is close to nominal size for $\alpha = 1\%$ for any number of treated clusters except at the extremes; for $\alpha = 5\%$ and $\alpha = 10\%$, it is close to nominal size for intermediate numbers of treated clusters and relatively close on average (means 6.6/13.8%, medians 4.4/8.7%) but it underrejects for lower and overrejects for higher numbers of treated clusters. The variant of the permutation test that preserves "Delaware"'s treatment status is true to size in the lower third of treated clusters and close to size on average (means 0.8/3.8/7.8%, medians 0.7/3.8/7.6%) but it gets increasingly excessively conservative as the number of treated clusters increases. Still, both variants' performance is excellent compared to the conventional test: the conventional test's least bad performances of 5.3/15.3/23.7% rejection rates for nominal 1/5/10% size at some numbers of treated clusters are as bad as the *worst* performances of the second permutation test (6.6/14.5/19.3%), and not materially better than the first permutation test's worst performances.

The cluster wild bootstrap is reasonably close to size for 5-12 treated clusters but strongly underrejects for fewer and increasingly overrejects for more treated clusters, and its average rejection rates are far above nominal size (means 5.9/16.1/25.7%, medians 2.1/11.2/21.3%); this remains true even when focusing only on the middle range of 10-40 treated clusters. The permutation test preserving Delaware's treatment status dominates the cluster wild bootstrap in terms of deviation from nominal size (measured as the absolute value of the logarithm of the ratio of rejection rate to size): the permutation test's deviation is smaller than the wild bootstrap's for all nominal test sizes and numbers of treated clusters except for the 5% test for five treated clusters (where both tests have almost exactly the right size).

7 Conclusion

Using Monte Carlo simulations, this paper demonstrates severe problems with conventional inference when using state corporate laws for identification of corporate governance effects in firm-level data, in particular the popular difference-in-difference panel approach. The paper also shows that various fixes proposed in the literature including the cluster wild bootstrap cannot deal with the extreme imbalance in incorporation state cluster sizes. The paper proposes a permutation test to address this problem along the lines of DiCiccio and J. P. Romano 2017; MacKinnon and Webb 2019a. The permutation test is exact under the randomization hypothesis, and shows promising performance superior to alternative tests in Monte Carlo simulations even when the randomization hypothesis does not hold.

Whether or not the permutation test proposed here will ultimately be adopted, researchers need to do something to address the severe inferential challenge posed by unequal cluster sizes. Importantly, while this paper has focused on demonstrat-

ing the worst problem originating from Delaware's dominant size, simply omitting Delaware firms from the sample will not solve all issues. Even without Delaware, incorporation cluster sizes are very unequal, which can be expected to trigger lesser but still sizeable inference problems as reviewed by, e.g., MacKinnon and Webb 2017.

Beyond the specifics of the tests, this paper can also be read as another cautionary tale about trying to find relatively small effects in noisy data using complex methods such as high-dimensional fixed effect models (cf. Young 2019 on instrumental variables). The high number of firm-year observations may mislead one into thinking that even small effects should be detectable. Once it is realized that the number of clusters is the relevant degrees of freedom for inference, however, it becomes clear that power will often be an issue. This is especially so because, as Carter, Schnepel, and Steigerwald 2017 have shown, the rate of convergence of the variance estimator is governed by the *effective* number of clusters, which with incorporation clusters is generally in the low single digits. The specifics will depend on the hypothesized effect size, the noisiness of the dependent variable, the distribution of the treatment assignment, and the ability to control for known predictors. Fortunately, modern computing power offers the ability to perform custom-made power calculations even for complex problems with relative ease. Researchers can and should also check the performance of their statistical tests specifically under the conditions that they are studying using methods such as those discussed in this paper.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge (2017). "When Should You Adjust Standard Errors for Clustering?"

Angrist, Joshua D. and Jörn-Steffen Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Bartlett, Robert and Frank Partnoy (2019). "The Misuse of Tobin's Q".

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004). "How Much Should We Trust Differences-In-Differences Estimates?" In: *The Quarterly Journal of Economics* 119.1, pp. 249–275.

Bester, C. Alan, Timothy G. Conley, and Christian B. Hansen (2011). "Inference with dependent data using cluster covariance estimators". In: *Journal of Econometrics* 165.2, pp. 137–151.

Bharath, Sreedhar T and Michael Hertzel (2019). "External Governance and Debt Structure". In: *The Review of Financial Studies* 32.9, pp. 3335–3365.

Cain, Matthew D., Stephen B. McKeon, and Steven Davidoff Solomon (2017). "Do takeover laws matter? Evidence from five decades of hostile takeovers". In: *Journal of Financial Economics* 124.3, pp. 464–485.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2008). "Bootstrap-Based Improvements for Inference with Clustered Errors". In: *The Review of Economics and Statistics* 90.3, pp. 414–427.

Cameron, A. Colin and Douglas L. Miller (2015). "A Practitioner's Guide to Cluster-Robust Inference". In: *The Journal of Human Resources* 50.2, pp. 317–372.

Canay, Ivan A., Joseph P. Romano, and Azeem M. Shaikh (2017). "Randomization Tests Under an Approximate Symmetry Assumption". In: *Econometrica* 85.3, pp. 1013–1030.

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald (2017). "Asymptotic Behavior of a t-Test Robust to Cluster Heterogeneity". In: *The Review of Economics and Statistics* 99.4, pp. 698–709.

Catan, Emiliano M. and Marcel Kahan (2016). "The Law and Finance of Antitakeover Statutes". In: *Stanford Law Review* 68, pp. 629–680.

Coates IV, John C. (2000). "Takeover Defenses in the Shadow of the Pill: A Critique of the Scientific Evidence". In: *Texas Law Review* 79.2, pp. 271–382.

Cohen, Alma and Charles C.Y. Wang (2013). "How do staggered boards affect shareholder value? Evidence from a natural experiment". In: *Journal of Financial Economics* 110.3, pp. 627–641.

Correia, Sergio (2016). *Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator*. Tech. rep. Working Paper.

Cremers, K.J. Martijn, Scott B. Guernsey, and Simone M. Sepe (2019). "Stakeholder Orientation and Firm Value".

Demiroglu, Cem, Cansu Iskenderoglu, and Oguzhan Ozbas (2019). "Managerial Discretion and Efficiency of Internal Capital Markets".

DiCiccio, Cyrus J. and Joseph P. Romano (2017). "Robust Permutation Tests For Correlation And Regression Coefficients". In: *Journal of the American Statistical Association* 112.519, pp. 1211–1220.

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

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

Gibbons, Charles E., Juan Carlos Suárez Serrato, and Michael B. Urbancic (2018). "Broken or Fixed Effects?" In: *Journal of Econometric Methods* 8 (1), pp. 1–12.

Gutiérrez Urtiaga, María and Antonio B. Vazquez (2019). "Boards of Directors' Legal Incentives and Firm Outcomes".

Hagemann, Andreas (2019). "Placebo inference on treatment effects when the number of clusters is small". In: *Journal of Econometrics* 213 (1), pp. 190–209.

Hansen, Bruce E. and Seojeong Lee (2019). "Asymptotic theory for clustered samples". In: *Journal of Econometrics* 210.2, pp. 268–290.

He, Zhaozhao and David A. Hirshleifer (2019). "The Exploratory Mindset and Corporate Innovation".

Ibragimov, Rustam and Ulrich K. Müller (2016). "Inference with Few Heterogeneous Clusters". In: *The Review of Economics and Statistics* 98.1, pp. 83–96.

Imbens, Guido W. and Michal Kolesár (2016). "Robust Standard Errors in Small Samples: Some Practical Advice". In: *The Review of Economics and Statistics* 98.4, pp. 701–712.

Karpoff, Jonathan M. and Michael D. Wittry (2018). "Institutional and Legal Context in Natural Experiments: The Case of State Antitakeover Laws". In: *The Journal of Finance* 73.2, pp. 657–714.

Lee, Chang Hyung and Douglas G. Steigerwald (2018). "Inference for Clustered Data". In: *The Stata Journal* 18.2, pp. 447–460.

Lehmann, Erich L. and Joseph P. Romano (2005). *Testing Statistical Hypotheses*. Springer-Verlag New York.

Liang, Kung-Yee and Scott L. Zeger (1986). "Longitudinal data analysis using generalized linear models". In: *Biometrika* 73.1, pp. 13–22.

MacKinnon, James G. and Matthew D. Webb (2017). "Wild Bootstrap Inference for Wildly Different Cluster Sizes". In: *Journal of Applied Econometrics* 32.2, pp. 233–254.

— (2019a). "Randomization Inference For Difference-in-differences With Few Treated Clusters". In: *Journal of Econometrics*.

— (2019b). *When and How to Deal with Clustered Errors in Regression Models*. Working Paper 1421. Kingston, Ontario: Queen's Economics Department.

Petersen, Mitchell A. (2009). "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches". In: *Review of Financial Studies* 22.1, pp. 435–480.

Romano, Roberta (1993). *The Genius of American Corporate Law*. The AEI Press.

Roodman, David, Morten Ørregaard Nielsen, James G. MacKinnon, and Matthew D. Webb (2019). "Fast and wild: Bootstrap inference in Stata using boottest". In: *The Stata Journal* 19.1, pp. 4–60.

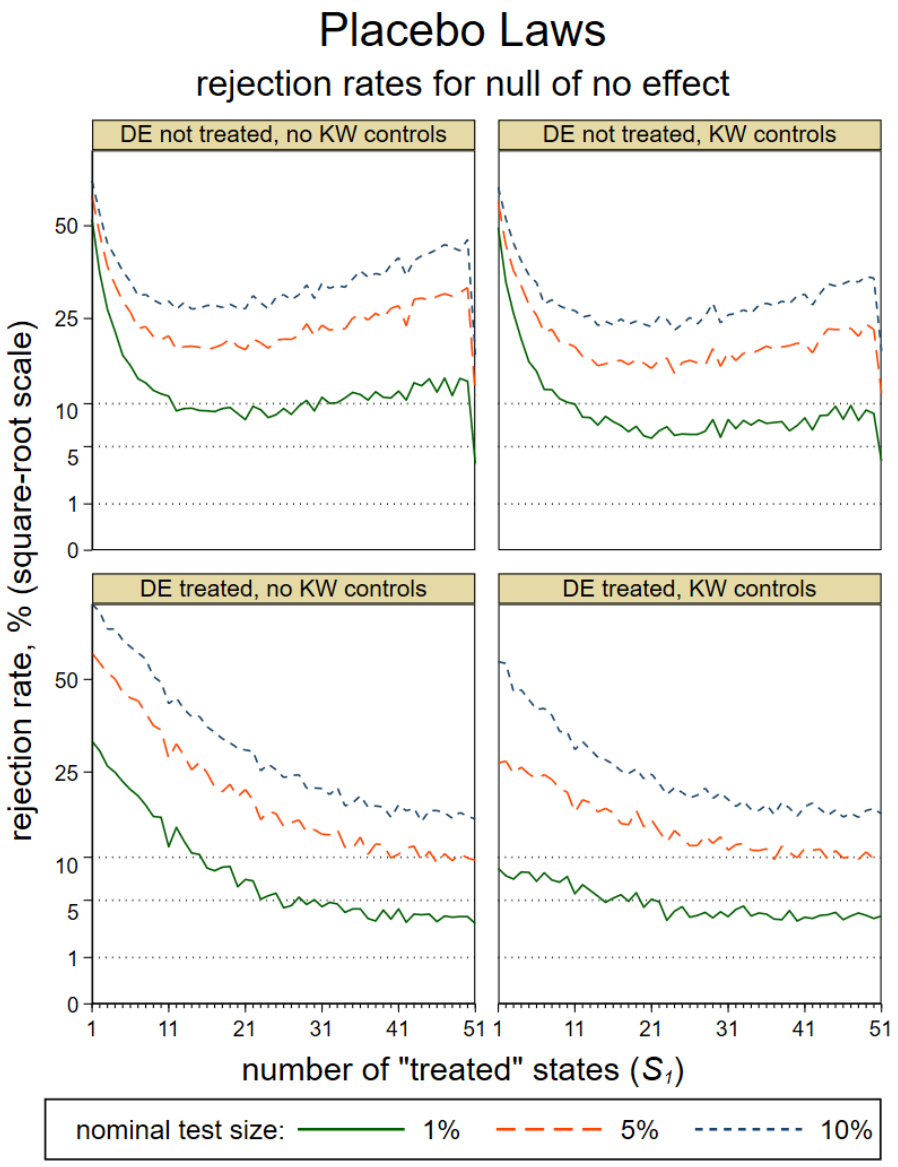
White, Halbert (1984). *Asymptotic Theory for Econometricians*. Academic Press.

Wooldridge, Jeffrey M. (2010). *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. MIT Press.

Young, Alwyn (2016). "Improved, Nearly Exact, Statistical Inference with Robust and Clustered Covariance Matrices using Effective Degrees of Freedom Corrections".

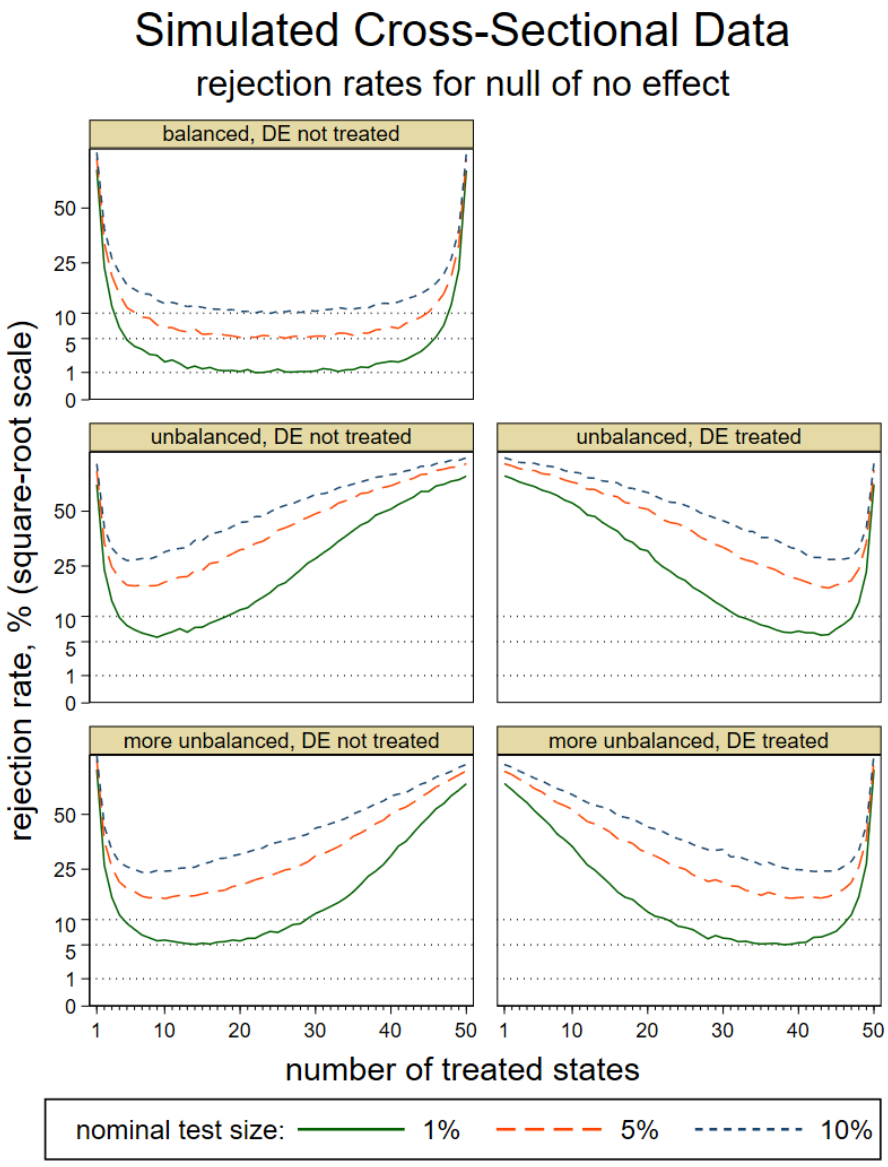
— (2019). "Consistency Without Inference: Instrumental Variables in Practical Application".

Figure 1:



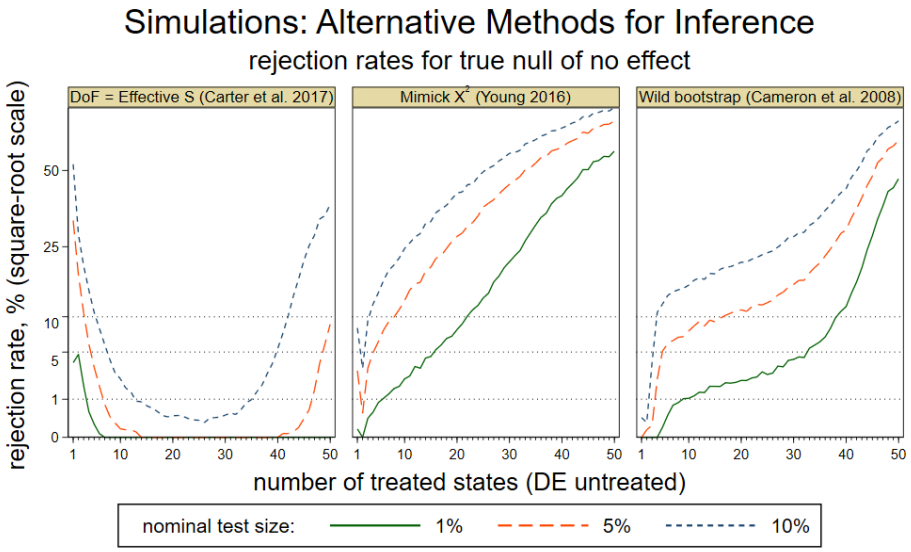
Rejection rates of conventional tests (using standard errors clustered by state of incorporation) after estimating equation (4) with D_{it}=1 assigned to S_1 random states starting in random years drawn independently for each "treated" state from a discrete uniform distribution over all sample years. In the top panels, DE is always among "treated" states; in the bottom panels, DE is never "treated" (except for S_1=51). The sample is all publicly traded U.S. firms 1983-2015. The dependent variable y_{it} is Tobin\'s q. Covariates z_{it} are fixed effects for industry-year (using Fama-French 49 industries) (all panels) and dummies for 5 2nd-gen. anti-takeover statutes from Karpoff and Wittry (2018) (right panels). Data on q were constructed from CRSP-Compustat. Data on SIC industry codes is from Compustat, and the scheme for conversion to Fama-French 49 industries is from Ken French\'s website. Data on historic state of incorporation is from the SEC EDGAR database for years starting in 1994 where available and otherwise from CRSP-Compustat header, and backfilled for prior years. Rates calculated from 2,000 runs per number of "treated" states, DE "treatment" status, and KW controls.

Figure 2:



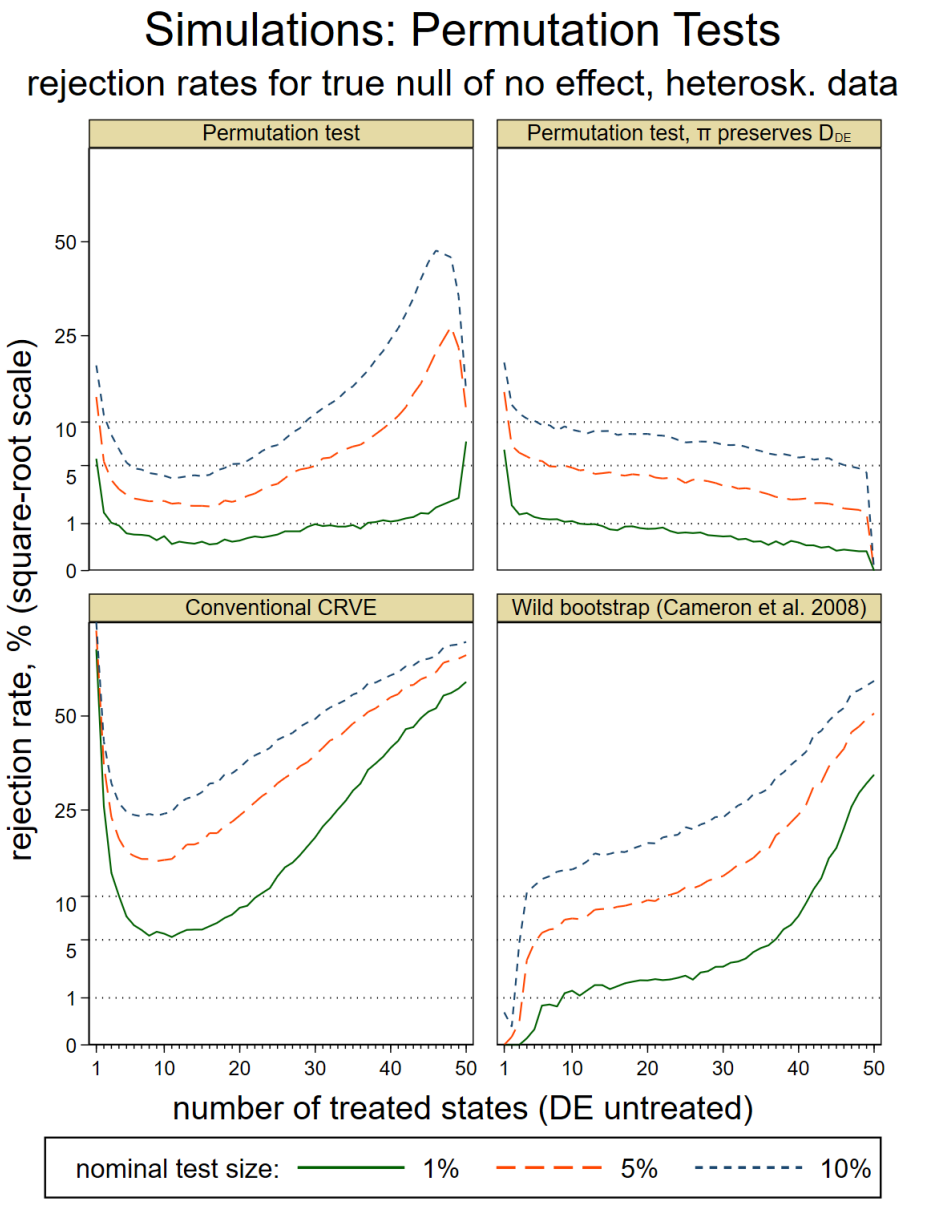
Rejection rates of conventional tests (using standard errors clustered by state of incorporation) after estimating equation (2) on simulated cross-sectional data (t=1 for all obs.). The data generating process creates S=51 states and N=5,100 firms distributed across states as follows:
(1) balanced: 100 observations per state,
(2) unbalanced: 51 observations per state except "Delaware" with 2,550 observations, or
(3) more unbalanced: 2,550 observations distributed uniformly between 2 and 100 in 50 states, with the remaining 2,550 in "Delaware."
The value of each firm's outcome variable y_{is} is the sum of two standard normal variables: one specific to firm i, and another common to all firms in state s. The respective number of "treated" states is randomly assigned D_s=1 and the remainder D_s=0, "Delaware" being (1) never drawn on the lef and (2) always drawn on the right. Firms are randomly grouped into 50 equal-sized "industry" groups (z_{is}). Neither industry nor treatment affects y_{is}. Rates calculated from 10,000 runs per number of treated states, "Delaware" treatment status, and balance.

Figure 3:



Rejection rates of alternative tests as per each plot's header after estimating equation (2) on simulated cross-sectional data (t=1 for all obs.) The data generating process creates 51 observations for each of S=51 states except "Delaware" for which it creates 2,550 observations. Each firm's outcome variable y_{ist} is the sum of two standard normal variables: one specific to firm i, and another common to all firms in state s. Firms are randomly grouped into 50 equal-sized "industry" groups (z_i). The respective number of "treated" states is randomly assigned D_{st}=1 and the remainder D_{st}=0, "Delaware" being never drawn. Neither industry nor treatment affects y_{ist}.
Rates calculated from 10,000 runs per method and number of treated states.

Figure 4:



Rejection rates of four tests as per plot header after estimating equation (8) on cross-sectional data (t=1 for all obs.) simulated using (14). The data generating process (14) creates 51 observations for each of S=51 states except "Delaware" for which it creates 2,550 observations. Each firm's outcome variable y_{sr} is the sum of 2-3 standard normal variables: one specific to firm i, another common to all firms in state s, and a third common to all firms in treated state s. Firms are randomly grouped into 50 equal-sized "industry" groups (z_{ir}). The respective number of "treated" states is randomly assigned D_{sr}=1 and the remainder D_{sr}=0, "Delaware" being never drawn. Neither industry nor treatment affects y_{sr}. Rates calculated from 10,000 runs per method and number of treated states.