Online Appendix for: Wild Bootstrap Inference for Wildly Different Cluster Sizes

James G. MacKinnon
jgm@econ.queensu.ca

Matthew D. Webb
matt.webb@carleton.ca

September 18, 2019

Abstract

This online appendix provides supplemental material for our paper, "Wild bootstrap inference for wildly different cluster sizes," (*Journal of Applied Econometrics* 32, 2017, 233–254). This material principally consists of tables and figures, but there are also some analytical results on why the wild bootstrap fails with few treated clusters, on the consequences of treated clusters being of wildly different sizes, and on the effects of aggregation on test power.

[Added 2019-08] The published paper and the original version of this appendix were affected by an unfortunate algebraic error in programs used for some of the simulations. Programs that were supposed to generate data with an intra-cluster correlation of ρ actually generated data with an intra-cluster correlation of ρ^2. All figures and tables that have been corrected are so indicated. The final section contains corrected versions of figures that appeared in the paper.

A.1 Introduction

Section A.2 contains extended versions of Tables I and II from Section 4, along with two additional tables. Section A.3 presents simulation results for pure treatment effect models, where every observation in a cluster is treated if any observation is treated. Section A.4 briefly discusses the unrestricted wild cluster bootstrap and presents evidence on how well it performs for the placebo laws experiments. Section A.5 expands on the discussion in Section 6 and presents two additional figures which provide more intuition about why the wild bootstrap fails when the number of treated clusters is small.

Section A.6 briefly explains why inference based on CRVE t statistics and both variants of the wild cluster bootstrap are valid under standard asymptotics. Section A.7 provides evidence about the performance of the wild bootstrap when there are few clusters and cluster sizes are unbalanced. Section A.8 discusses the use of aggregate data and presents results for placebo laws experiments using 1071 observations at the state-year level. Section A.9 discusses the power loss associated with aggregating micro data to the cluster-period level, creating perfectly balanced clusters. Finally, Section A.10 contains corrected versions of some of the figures from the paper itself.

A.2 Tables for Continuous Regressor Experiments

This section provides additional results for Section 4. Rejection frequencies are reported for $\rho_\epsilon = 0.0, 0.1, \ldots, 0.9$ instead of just four values of ρ_ϵ, and they are reported to five decimal places instead of four. Tables A.1 and A.2 contain results for 50 equal-sized and state-sized clusters, respectively, and Tables A.3 and A.4 contain results for 100 equal-sized and state-sized clusters. For tests with rejection frequencies close to the nominal level of 0.05, the standard error of these estimates is 0.000345.

`[Added 2019-08]` Tables I and II in the paper and the four tables in the original version of this appendix were affected by a programming error. The intended values of ρ_ϵ and ρ_x were inadvertently replaced by their squares. Tables A.1 through A.4 contain corrected results. Note, however, that the results in the original tables would be correct if the values of ρ_ϵ and ρ_x on the vertical and horizontal axes were squared.

Table A.1: Rejection Frequencies with 50 Equal-Sized Clusters

				$\boldsymbol{\rho_x}$			
$\boldsymbol{\rho_\epsilon}$		**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
0.0	$\mathbf{t(G-1)}$	0.04978	0.05109	0.05280	0.05653	0.06004	0.06568
	$\mathbf{t(G^*-1)}$	0.04948	0.04961	0.04911	0.05001	0.05096	0.05357
	bootstrap	0.04946	0.05003	0.04989	0.05027	0.04999	0.04980
0.1	$\mathbf{t(G-1)}$	0.05057	0.05211	0.05427	0.05723	0.06103	0.06655
	$\mathbf{t(G^*-1)}$	0.05027	0.05031	0.04982	0.05022	0.05121	0.05439
	bootstrap	0.05032	0.05014	0.05023	0.04995	0.04988	0.05043
0.2	$\mathbf{t(G-1)}$	0.05020	0.05235	0.05606	0.06062	0.06388	0.06590
	$\mathbf{t(G^*-1)}$	0.04987	0.04965	0.05033	0.05188	0.05320	0.05390
	bootstrap	0.04967	0.04978	0.05008	0.05069	0.05046	0.04980
0.3	$\mathbf{t(G-1)}$	0.05060	0.05292	0.05735	0.06124	0.06418	0.06581
	$\mathbf{t(G^*-1)}$	0.05016	0.04865	0.05022	0.05164	0.05302	0.05346
	bootstrap	0.04991	0.04941	0.05045	0.05003	0.05006	0.04955
0.4	$\mathbf{t(G-1)}$	0.05102	0.05400	0.05797	0.06205	0.06467	0.06642
	$\mathbf{t(G^*-1)}$	0.05041	0.04816	0.04932	0.05166	0.05301	0.05395
	bootstrap	0.05014	0.04976	0.04980	0.05017	0.05003	0.05004
0.5	$\mathbf{t(G-1)}$	0.05061	0.05534	0.05871	0.06251	0.06506	0.06558
	$\mathbf{t(G^*-1)}$	0.04961	0.04806	0.04915	0.05184	0.05334	0.05360
	bootstrap	0.05014	0.05084	0.05007	0.05009	0.05004	0.04963
0.6	$\mathbf{t(G-1)}$	0.05087	0.05518	0.05965	0.06293	0.06535	0.06645
	$\mathbf{t(G^*-1)}$	0.04914	0.04677	0.04916	0.05158	0.05344	0.05412
	bootstrap	0.05004	0.05063	0.05019	0.04999	0.05012	0.05024
0.7	$\mathbf{t(G-1)}$	0.05106	0.05469	0.05936	0.06274	0.06486	0.06606
	$\mathbf{t(G^*-1)}$	0.04825	0.04509	0.04824	0.05112	0.05285	0.05379
	bootstrap	0.05016	0.04989	0.04961	0.04984	0.04953	0.04961
0.8	$\mathbf{t(G-1)}$	0.05101	0.05498	0.05982	0.06443	0.06479	0.06549
	$\mathbf{t(G^*-1)}$	0.04634	0.04446	0.04861	0.05235	0.05237	0.05326
	bootstrap	0.05009	0.04988	0.05011	0.05077	0.04948	0.04924
0.9	$\mathbf{t(G-1)}$	0.05030	0.05525	0.05996	0.06353	0.06555	0.06620
	$\mathbf{t(G^*-1)}$	0.04322	0.04411	0.04840	0.05161	0.05349	0.05414
	bootstrap	0.04984	0.05018	0.05046	0.05020	0.05046	0.05041

Notes: Rejection frequencies are at the .05 level and are based on 400,000 replications. There are 50 equal-sized clusters with a total of 2000 observations. The effective number of clusters is $G^*(\hat{\rho})$. The wild bootstrap uses the Rademacher distribution with 399 bootstraps. `[corrected 2019-08]`

Table A.2: Rejection Frequencies with 50 State-Sized Clusters

				$\boldsymbol{\rho_x}$			
$\boldsymbol{\rho_\epsilon}$		**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
0.0	$\mathbf{t(G-1)}$	0.05784	0.06019	0.06305	0.06710	0.07273	0.08191
	$\mathbf{t(G^*-1)}$	0.05035	0.03486	0.02875	0.02652	0.02567	0.02523
	bootstrap	0.04905	0.05001	0.05113	0.05080	0.05092	0.05060
0.1	$\mathbf{t(G-1)}$	0.05903	0.06404	0.06996	0.07792	0.08497	0.09338
	$\mathbf{t(G^*-1)}$	0.05128	0.03559	0.03034	0.02881	0.02942	0.02820
	bootstrap	0.05016	0.05092	0.05085	0.05176	0.05176	0.05124
0.2	$\mathbf{t(G-1)}$	0.05820	0.06962	0.07826	0.08665	0.09597	0.10113
	$\mathbf{t(G^*-1)}$	0.05063	0.03506	0.03083	0.03055	0.03204	0.02938
	bootstrap	0.04935	0.05150	0.05142	0.05162	0.05242	0.05106
0.3	$\mathbf{t(G-1)}$	0.05892	0.07377	0.08341	0.09179	0.09926	0.10434
	$\mathbf{t(G^*-1)}$	0.05077	0.03280	0.02977	0.02964	0.03060	0.02944
	bootstrap	0.05006	0.05197	0.05158	0.05177	0.05164	0.05182
0.4	$\mathbf{t(G-1)}$	0.05907	0.07510	0.08549	0.09506	0.10086	0.10506
	$\mathbf{t(G^*-1)}$	0.05089	0.02946	0.02770	0.02874	0.03003	0.02930
	bootstrap	0.05006	0.05133	0.05145	0.05236	0.05181	0.05125
0.5	$\mathbf{t(G-1)}$	0.05937	0.07712	0.08704	0.09589	0.10226	0.10619
	$\mathbf{t(G^*-1)}$	0.05029	0.02705	0.02645	0.02770	0.02954	0.02904
	bootstrap	0.05057	0.05216	0.05189	0.05196	0.05204	0.05164
0.6	$\mathbf{t(G-1)}$	0.05854	0.07778	0.08847	0.09750	0.10316	0.10606
	$\mathbf{t(G^*-1)}$	0.04843	0.02476	0.02579	0.02761	0.02932	0.02870
	bootstrap	0.05006	0.05203	0.05259	0.05219	0.05216	0.05110
0.7	$\mathbf{t(G-1)}$	0.05876	0.07858	0.08983	0.09650	0.10387	0.10761
	$\mathbf{t(G^*-1)}$	0.04651	0.02313	0.02494	0.02697	0.02871	0.02872
	bootstrap	0.05098	0.05188	0.05267	0.05187	0.05220	0.05172
0.8	$\mathbf{t(G-1)}$	0.05780	0.07869	0.09019	0.09789	0.10440	0.10776
	$\mathbf{t(G^*-1)}$	0.04285	0.02200	0.02433	0.02658	0.02854	0.02882
	bootstrap	0.05096	0.05232	0.05277	0.05240	0.05187	0.05198
0.9	$\mathbf{t(G-1)}$	0.05668	0.07866	0.08857	0.09799	0.10474	0.10718
	$\mathbf{t(G^*-1)}$	0.03754	0.02103	0.02325	0.02612	0.02868	0.02826
	bootstrap	0.05133	0.05222	0.05178	0.05192	0.05219	0.05113

Notes: See notes to Table A.1. There are 50 clusters proportional to US state populations with a total of 2000 observations. `[Corrected 2019-08]`

Table A.3: Rejection Frequencies with 100 Equal-Sized Clusters

				$\boldsymbol{\rho_x}$			
$\boldsymbol{\rho_\epsilon}$		**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
0.0	$\mathbf{t(G-1)}$	0.05029	0.05099	0.05184	0.05392	0.05581	0.05770
	$\mathbf{t(G^*-1)}$	0.05002	0.05009	0.04975	0.05067	0.05132	0.05230
	bootstrap	0.04990	0.05008	0.05020	0.05065	0.05032	0.04952
0.1	$\mathbf{t(G-1)}$	0.05003	0.05110	0.05233	0.05393	0.05583	0.05893
	$\mathbf{t(G^*-1)}$	0.04977	0.05007	0.05017	0.05051	0.05123	0.05295
	bootstrap	0.04947	0.05006	0.05012	0.05019	0.05015	0.05038
0.2	$\mathbf{t(G-1)}$	0.05057	0.05064	0.05315	0.05413	0.05577	0.05851
	$\mathbf{t(G^*-1)}$	0.05026	0.04934	0.05055	0.05031	0.05098	0.05269
	bootstrap	0.05004	0.04928	0.05020	0.04936	0.04928	0.05047
0.3	$\mathbf{t(G-1)}$	0.05068	0.05202	0.05366	0.05590	0.05762	0.05921
	$\mathbf{t(G^*-1)}$	0.05028	0.05031	0.05052	0.05160	0.05246	0.05330
	bootstrap	0.05017	0.05035	0.05008	0.05047	0.05051	0.05051
0.4	$\mathbf{t(G-1)}$	0.05104	0.05140	0.05394	0.05651	0.05717	0.05821
	$\mathbf{t(G^*-1)}$	0.05057	0.04914	0.05018	0.05155	0.05189	0.05207
	bootstrap	0.05062	0.04918	0.04997	0.05042	0.04984	0.04968
0.5	$\mathbf{t(G-1)}$	0.05053	0.05201	0.05421	0.05676	0.05769	0.05814
	$\mathbf{t(G^*-1)}$	0.04986	0.04905	0.04996	0.05184	0.05231	0.05228
	bootstrap	0.05006	0.04939	0.04959	0.05046	0.05013	0.04984
0.6	$\mathbf{t(G-1)}$	0.05095	0.05276	0.05426	0.05650	0.05786	0.05844
	$\mathbf{t(G^*-1)}$	0.05003	0.04917	0.04955	0.05134	0.05239	0.05264
	bootstrap	0.05039	0.05004	0.04944	0.04988	0.04999	0.04997
0.7	$\mathbf{t(G-1)}$	0.05106	0.05241	0.05420	0.05708	0.05790	0.05783
	$\mathbf{t(G^*-1)}$	0.04936	0.04820	0.04907	0.05167	0.05220	0.05220
	bootstrap	0.05039	0.04974	0.04919	0.05021	0.04991	0.04951
0.8	$\mathbf{t(G-1)}$	0.05044	0.05309	0.05565	0.05730	0.05795	0.05888
	$\mathbf{t(G^*-1)}$	0.04802	0.04836	0.05032	0.05162	0.05220	0.05300
	bootstrap	0.04974	0.05032	0.05060	0.05049	0.05009	0.05027
0.9	$\mathbf{t(G-1)}$	0.05119	0.05291	0.05622	0.05664	0.05803	0.05797
	$\mathbf{t(G^*-1)}$	0.04763	0.04795	0.05057	0.05087	0.05231	0.05213
	bootstrap	0.05063	0.05015	0.05091	0.04953	0.05018	0.04981

Notes: See notes to Table A.1. There are 100 equal-sized clusters with a total of 2000 observations. `[Corrected 2019-08]`

Table A.4: Rejection Frequencies with 100 State-Sized Clusters

				$\boldsymbol{\rho_x}$			
$\boldsymbol{\rho_\epsilon}$		**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
0.0	$\mathbf{t(G-1)}$	0.05497	0.05614	0.05743	0.05943	0.06355	0.06864
	$\mathbf{t(G^*-1)}$	0.05101	0.03951	0.03370	0.03221	0.03190	0.03116
	bootstrap	0.05007	0.05076	0.05042	0.05005	0.05043	0.05083
0.1	$\mathbf{t(G-1)}$	0.05469	0.05695	0.06036	0.06431	0.06812	0.07321
	$\mathbf{t(G^*-1)}$	0.05071	0.03948	0.03563	0.03385	0.03375	0.03328
	bootstrap	0.04999	0.05028	0.05061	0.05066	0.05012	0.05008
0.2	$\mathbf{t(G-1)}$	0.05483	0.06022	0.06413	0.06948	0.07437	0.07857
	$\mathbf{t(G^*-1)}$	0.05088	0.04025	0.03620	0.03553	0.03603	0.03506
	bootstrap	0.05001	0.05080	0.05024	0.05009	0.05048	0.05053
0.3	$\mathbf{t(G-1)}$	0.05532	0.06187	0.06833	0.07327	0.07795	0.08153
	$\mathbf{t(G^*-1)}$	0.05119	0.03927	0.03656	0.03615	0.03674	0.03579
	bootstrap	0.05014	0.05035	0.05109	0.05073	0.05082	0.05059
0.4	$\mathbf{t(G-1)}$	0.05428	0.06359	0.06972	0.07556	0.07909	0.08266
	$\mathbf{t(G^*-1)}$	0.04983	0.03723	0.03580	0.03608	0.03642	0.03537
	bootstrap	0.04957	0.05038	0.05107	0.05097	0.05040	0.05035
0.5	$\mathbf{t(G-1)}$	0.05505	0.06413	0.07018	0.07640	0.07987	0.08279
	$\mathbf{t(G^*-1)}$	0.05018	0.03545	0.03411	0.03525	0.03595	0.03581
	bootstrap	0.05018	0.05046	0.05042	0.05124	0.05033	0.05013
0.6	$\mathbf{t(G-1)}$	0.05560	0.06512	0.07079	0.07676	0.08090	0.08438
	$\mathbf{t(G^*-1)}$	0.04981	0.03385	0.03269	0.03426	0.03589	0.03628
	bootstrap	0.05050	0.05042	0.04970	0.05034	0.05055	0.05105
0.7	$\mathbf{t(G-1)}$	0.05561	0.06444	0.07177	0.07702	0.08106	0.08368
	$\mathbf{t(G^*-1)}$	0.04898	0.03152	0.03204	0.03359	0.03531	0.03537
	bootstrap	0.05084	0.04993	0.05024	0.05014	0.05070	0.05059
0.8	$\mathbf{t(G-1)}$	0.05524	0.06568	0.07252	0.07775	0.08164	0.08393
	$\mathbf{t(G^*-1)}$	0.04717	0.03077	0.03155	0.03323	0.03460	0.03540
	bootstrap	0.05051	0.05018	0.05052	0.05076	0.05073	0.05042
0.9	$\mathbf{t(G-1)}$	0.05468	0.06609	0.07304	0.07749	0.08161	0.08442
	$\mathbf{t(G^*-1)}$	0.04433	0.02906	0.03108	0.03304	0.03433	0.03587
	bootstrap	0.05112	0.05047	0.05085	0.05099	0.05023	0.05093

Notes: See notes to Table A.1. There are 100 clusters proportional to US state populations with a total of 2000 observations. [Corrected 2019-08]

A.3 Treatment Effects

Section 5 deals with treatment effects in the context of a difference-in-differences regression, where certain clusters are treated during certain time periods. Here we consider pure treatment effects with no time dimension, where the test regressor is an indicator variable that equals 1 for some proportion P of the clusters. Thus, for each cluster, either all observations are treated or all are not treated.

In Figures A.1, A.2, and A.3, we report results for 50 clusters with 2000 observations, $\rho_\epsilon = 0.05$, and G_1 that varies between 1 and 49.[1] The treatments are applied to equal-sized clusters and to state-sized clusters both from smallest to largest and from largest to smallest. The simulations use $400,000$ replications.

Figure A.1 shows results for tests based on CRVE standard errors and $t(49)$ critical values. As in the DiD case, there is very severe overrejection when either G_1 or $G-G_1$ is very small. Note that, because rejection frequencies vary so much, a square root transformation is applied to the vertical axis.

In Figure A.1, overrejection is quite modest when G_1 and $G - G_1$ are both far from 0. With equal-sized clusters, rejection frequencies are very close to 0.05 for G_1 between 17 and 33. With state-sized clusters, they are somewhat higher, never falling below 0.0616. The graph for equal-sized clusters is symmetric around $G_1 = 25$, while the ones for state-sized clusters are somewhat asymmetric. Overrejection is substantially more severe when only a few small clusters are either treated or not treated than when only a few large clusters are in that situation.

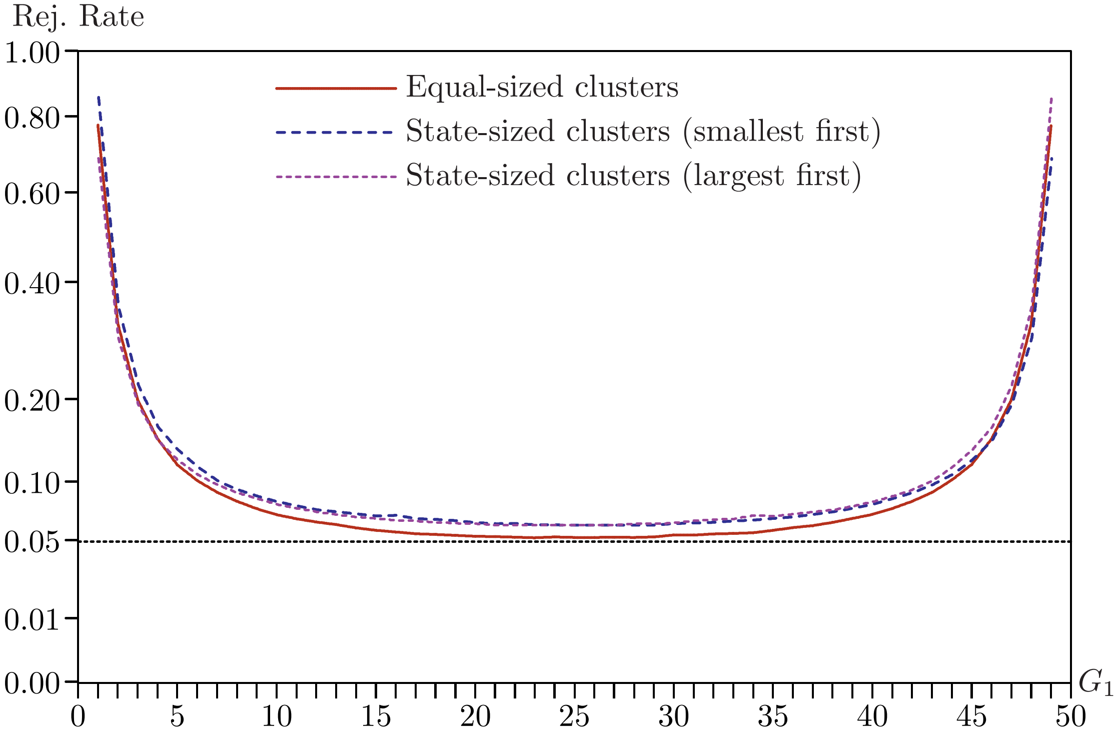
Figure A.2 shows results for tests based on $t(G^* - 1)$ critical values, where G^* is based on $\hat{\rho}$. There is extreme underrejection when $G_1 = 1$ and $G_1 = 49$. Rejection frequencies are very sensitive to the degrees of freedom parameter; using critical values based on $t(G^*)$ instead of $t(G^* - 1)$ leads to moderately severe overrejection. Away from the extremes, the tests can either underreject or overreject, although they always overreject for equal-sized clusters when $3 \leq G_1 \leq 47$. The rejection frequencies appear to be symmetric around $G_1 = 25$ for equal-sized clusters, but quite asymmetric for state-sized ones. In the latter case, there can be either noticeable overrejection or severe underrejection.

Figure A.3 shows results for wild bootstrap tests based on simulations with 399 bootstraps. In all cases, there is severe underrejection when either G_1 or $G - G_1$ is very small. In the most extreme cases, there are no rejections at all. For equal-sized clusters, there is modest overrejection when the number of treated or untreated clusters is between 4 and 6, but the wild bootstrap tests work extremely well for G_1 between about 7 and 43.

For state-sized clusters, the pattern is a bit more complicated. When the states are treated from smallest to largest, the bootstrap tests always underreject severely when G_1 is very close to 0 or 50, and they overreject quite severely when G_1 is between 44 and 48. When the states are treated from largest to smallest, the opposite problem occurs, with serious overrejection when G_1 is between 2 and 6, and severe underrejection when G_1 is very close to 0 or 50. In both cases, the peak overrejection occurs when there are 3 treated (or non-treated) clusters that together account for 26.6% of the observations.

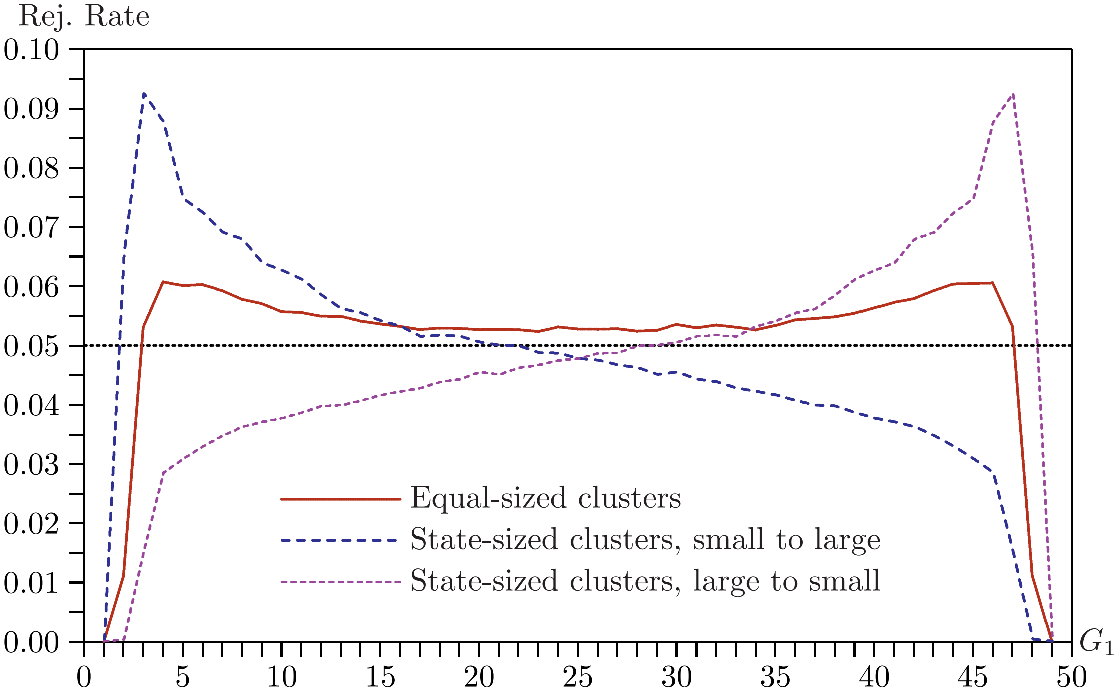
[1] In earlier versions of the paper, we set $\rho_\epsilon = 0.5$, a number that is unrealistically high. That is why Figures A.2 and A.3 look noticeably different from previous versions of the same figures.

Figure A.1: Rejection rates for pure treatment case, $t(G-1)$



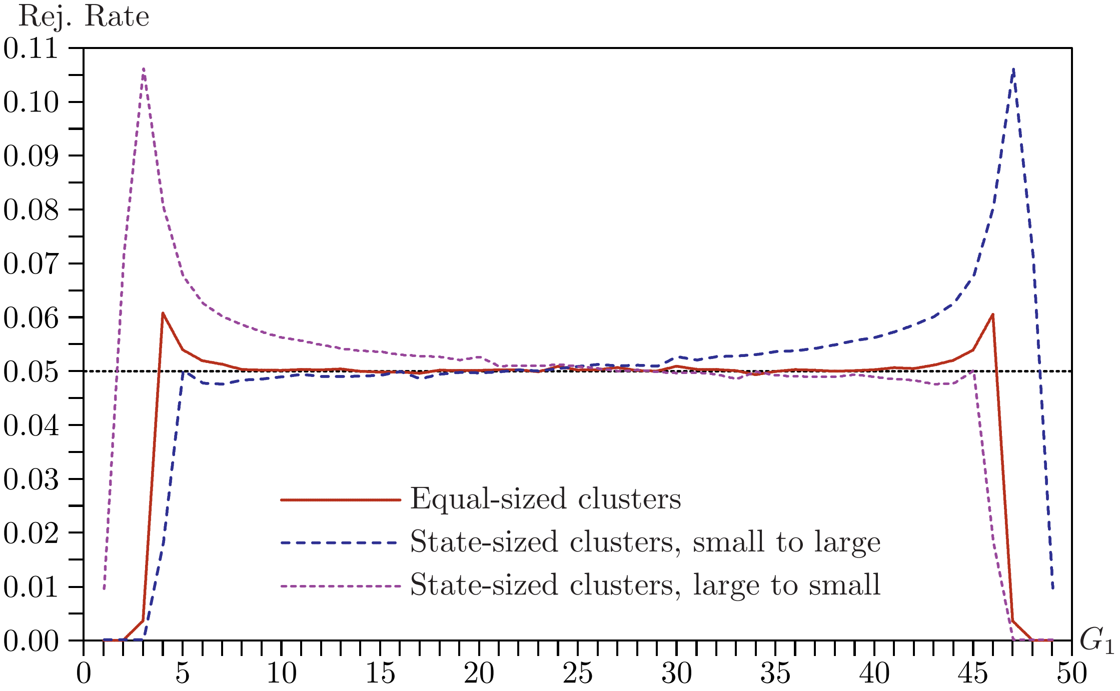
[Corrected 2019-08]

Figure A.2: Rejection rates for pure treatment case, $t(G^*-1)$



[Corrected 2019-08]

Figure A.3: Rejection rates for pure treatment case, wild bootstrap



[Corrected 2019-08]

Figures A.2 and A.3 are the only ones that changed substantially when the value of ρ_ϵ was increased from 0.0025 (due to a programming error) to the intended value of 0.05. The overrejection by the WCR bootstrap when a few large clusters are treated became much more severe, as did the overrejection by tests based on $t(G^* - 1)$ when a few small clusters are treated. Why the value of ρ_ϵ matters so much more for the pure treatment case than for the DiD case (Figures 2 and 3 changed much less when ρ_ϵ was increased, although in qualitatively the same ways) is unknown.

A.4 The Unrestricted Wild Cluster Bootstrap

The wild cluster bootstrap is discussed in Section 2. The bootstrap DGP (3) uses restricted residuals, because it is generally best to base bootstrap DGPs on restricted estimates; see Davidson and MacKinnon (1999). However, it is also perfectly valid to use the unrestricted residuals $\hat{\epsilon}_{ig}$ instead of the restricted ones $\tilde{\epsilon}_{ig}$ in the bootstrap DGP. In addition, it would be possible to use the unrestricted estimates $\hat{\boldsymbol{\beta}}$ rather than the restricted estimates $\tilde{\boldsymbol{\beta}}$, provided the bootstrap test statistic were modified so as to test the null hypothesis that $\beta_k = \hat{\beta}_k$ rather than $\beta_k = 0$. The two variants of the unrestricted wild cluster bootstrap would yield identical results in all our experiments, because the $\boldsymbol{X}$ matrix does not depend on $\boldsymbol{\beta}$. We will refer to the wild cluster bootstrap based on restricted residuals as WCR and to the one based on unrestricted residuals as WCU.

Figure A.4 shows rejection frequencies for the WCR and WCU bootstraps for the placebo laws experiments. As in Figure A.1, a square root transformation has been applied to the vertical axis. For clarity, the vertical axis only extends to 0.30, which means that results for

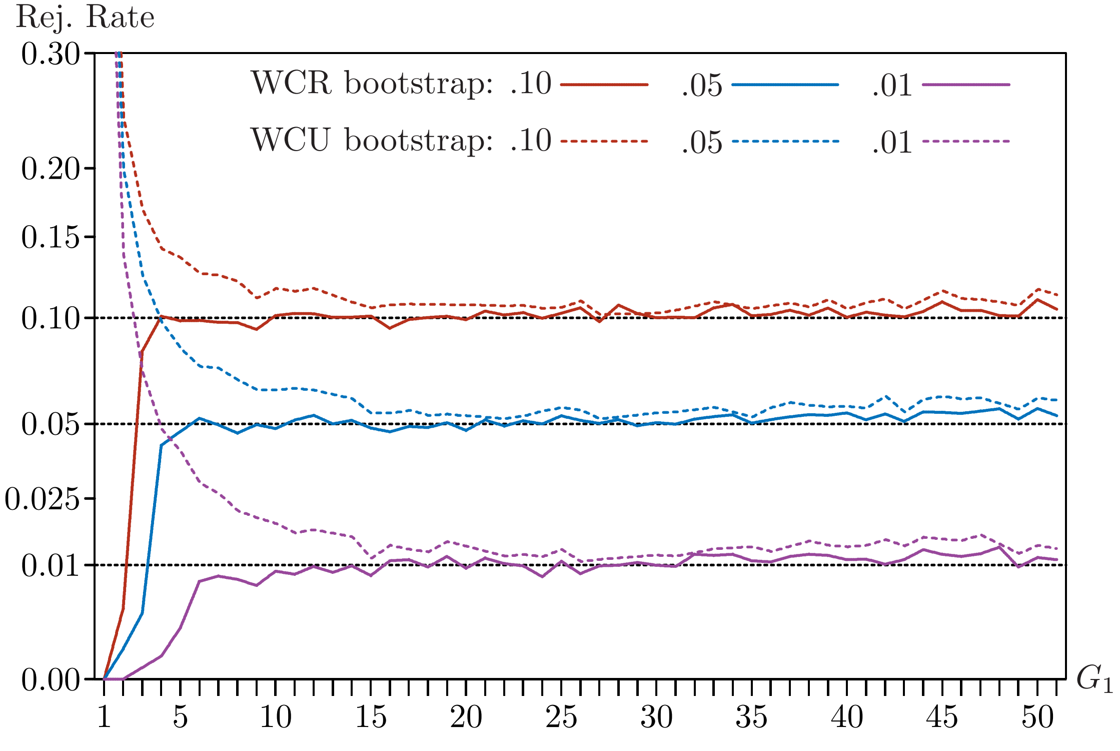
WCU with just one treated state cannot be shown. The rejection frequencies at the .10, .05, and .01 levels for this case are 0.779, 0.731, and 0.647, respectively. The results for WCR are the same as the ones in Figure 9. For very small values of G_1, WCR underrejects very severely, and WCU overrejects very severely. For somewhat larger values, WCR works well, but WCU overrejects noticeably. For values of G_1 greater than about 15, both procedures work quite well, although not perfectly. They both have a slight tendency to overreject in this region, with WCU always rejecting a bit more often than WCR, especially for very large values of G_1. Overall, WCR is clearly the procedure of choice.

The reason why WCU overrejects, often severely, for small values of G_1 was discussed in Section 6. Essentially, the problem is that the bootstrap DGP uses unrestricted residuals for the treated cluster(s) which sum to zero when $G_1 = 1$ and sum to numbers that are too small when $G_1 > 1$. This causes the variance of the bootstrap test statistics to be too small. Exactly the same problem causes the CRVE to underestimate the variance of the coefficient on the treatment dummy.

In Section 6, we discussed why the sum of the unrestricted residuals for just one treated cluster is equal to zero, but we did not formally discuss the case in which $G_1 > 1$. Consider once again the pure treatment model given by equation (8), where the first cluster and at least one other cluster is treated. The sum of the residuals for the first cluster is $\boldsymbol{\iota}'_{N_1}\hat{\boldsymbol{\epsilon}}_1$, and we want to find the expectation of this quantity.

By a well-known result for OLS regression, $\hat{\boldsymbol{\epsilon}} = \big(\mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\big)\boldsymbol{\epsilon}$. In the case of the

Figure A.4: Placebo laws: Rejection rates for two wild bootstrap methods



pure treatment model,

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} \boldsymbol{\iota}'\boldsymbol{\iota} & \boldsymbol{\iota}'\boldsymbol{d} \\ \boldsymbol{d}'\boldsymbol{\iota} & \boldsymbol{d}'\boldsymbol{d} \end{bmatrix} = \begin{bmatrix} N & M_1 \\ M_1 & M_1 \end{bmatrix},$$

and

$$\begin{bmatrix} \boldsymbol{\iota}'_{N_1} \\ \mathbf{0} \end{bmatrix}' \begin{bmatrix} \boldsymbol{\iota} & \boldsymbol{d} \end{bmatrix} = \begin{bmatrix} N_1 & N_1 \end{bmatrix},$$

so that

$$\boldsymbol{\iota}'_{N_1}\hat{\boldsymbol{\epsilon}}_1 = \boldsymbol{\iota}'_{N_1}\boldsymbol{\epsilon} - \begin{bmatrix} N_1 & N_1 \end{bmatrix} \begin{bmatrix} N & M_1 \\ M_1 & M_1 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\iota} & \boldsymbol{d} \end{bmatrix}' \boldsymbol{\epsilon}.$$

Recall that N_1 is the number of observations in cluster 1, and M_1 is the total number of observations in all the treated clusters.

In order to take the expectation of $\boldsymbol{\iota}'_{N_1}\hat{\boldsymbol{\epsilon}}_1\hat{\boldsymbol{\epsilon}}'_1\boldsymbol{\iota}_{N_1}$, we need to make an assumption about how the vector $\boldsymbol{\epsilon}$ is distributed. The simplest assumption is that $\boldsymbol{\epsilon}$ has mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$. In that case, it is easy to see that

$$\begin{aligned} \mathrm{Var}(\boldsymbol{\iota}'_{N_1}\hat{\boldsymbol{\epsilon}}_1) &= \sigma^2\boldsymbol{\iota}'_{N_1}\Big(\mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\Big)\boldsymbol{\iota}_{N_1} \\ &= N_1\sigma^2 - \begin{bmatrix} N_1 & N_1 \end{bmatrix} \begin{bmatrix} N & M_1 \\ M_1 & M_1 \end{bmatrix} \begin{bmatrix} N_1 \\ N_1 \end{bmatrix} \sigma^2. \end{aligned} \tag{A.1}$$

It is not difficult to verify that equation (A.1) can be rewritten as

$$\mathrm{Var}(\boldsymbol{\iota}'_{N_1}\hat{\boldsymbol{\epsilon}}_1) = \sigma^2\frac{N_1(M_1 - N_1)}{M_1}. \tag{A.2}$$

Thus the factor by which the variance of $\boldsymbol{\iota}'_{N_1}\hat{\boldsymbol{\epsilon}}_1$, the sum of the residuals for the first treated cluster, is shrunk relative to the variance of $\boldsymbol{\iota}'_{N_1}\boldsymbol{\epsilon}_1$ is simply $(M_1 - N_1)/M_1$. This factor is the ratio of the number of treated observations in clusters other than first one to the total number of treated observations. This result is consistent with what we saw in Section 6, namely, that when only one cluster is treated, the variance of $\boldsymbol{\iota}'_{N_1}\hat{\boldsymbol{\epsilon}}_1$ is zero.

The result (A.2) depends on the assumption of IID errors, which may not seem entirely reasonable in this case. However, any other assumption would lead to a much more complicated analysis, and it seems very unlikely that such an analysis would reverse the key implication of equation (A.2). This implication is that, as the fraction of treated observations that belong to a single cluster increases, the variance of the sum of the residuals for that cluster falls relative to the variance of the sum of the corresponding error terms. This strongly suggests that inference based on CRVE t statistics will be less reliable when the sizes of the treated clusters varies substantially than when all the treated clusters are approximately the same size. Section A.7 provides some evidence on this point.

A.5 Why the Wild Bootstrap Fails for Few Treated Clusters

In Section 6, we discussed why the wild bootstrap fails when the number of treated clusters is small. In particular, we considered the case in which $G_1 = 1$ and the bootstrap DGP uses the Rademacher distribution. We showed that, when the test statistic t_2 is large in absolute value, the distribution of the restricted wild cluster bootstrap t statistics tends to be bimodal, with half the realizations distributed around t_2 and the other half distributed around $-t_2$. This was illustrated in Figure 4. As Section 6 of the paper explains, and as Figures 6 and 7 illustrate, the WCU and WCR bootstraps fail in very different ways when G_1 is small. When $G_1 = 1$ and all the clusters are of equal size, the distribution of the WCU bootstrap statistics seems to be very close to $t(G-1)$. This happens because the first term in expression (17) is identically zero. The numerator of the bootstrap statistic is therefore proportional to

$$\sum_{g=2}^{G} v_g^{*j} \left(\sum_{i=1}^{N_g} \hat{\epsilon}_{ig} \right). \tag{A.3}$$

This is a summation of $G-1$ random variables, each of which has expectation zero. There is just one random variable for each of the untreated clusters, because each of the summations inside the parentheses is multiplied by a single auxiliary random variable v_g^{*j}.

As can be seen from equation (13), the denominator of the bootstrap statistic is in this case proportional to the square root of

$$\sum_{g=2}^{G} \left(\sum_{i=1}^{N_g} \hat{\epsilon}_{ig}^{*} \right)^2, \tag{A.4}$$

where the $\hat{\epsilon}_{ig}^{*}$ are the OLS residuals for the bootstrap data. These residuals (for the $G-1$ untreated clusters only) must be approximately equal to the corresponding bootstrap error terms, $v_g^{*j}\hat{\epsilon}_{ig}$. The summations of the $\hat{\epsilon}_{ig}^{*}$ over each cluster are not exactly normally distributed. Nevertheless, in our experiments, whatever distribution they do have when all the N_g are equal seems to be close enough that $t(G-1)$ is a very good approximation to the distribution of the WCU bootstrap test statistics according to Kolmogorov-Smirnov tests. When cluster sizes are unbalanced, however, each of the $G-1$ summations in expression (A.4) has a different variance, and so we would not expect any t distribution to provide a good approximation. That is exactly what we find when the N_g vary substantially.

This explains why the rejection frequencies for WCU bootstrap tests and for tests based on the $t(G-1)$ distribution are so close when $G_1 = 1$ in many of our experiments. For example, for the placebo laws experiments with aggregate data, the former rejects 71.8% of the time at the .05 level, and the latter rejects 72.0% of the time. As expected, the correspondence is not quite as good with micro data. The WCU bootstrap tests reject 73.1% of the time, and tests based on the $t(G-1)$ distribution reject 75.0% of the time.

Of course, as G_1 increases, both variants of the wild cluster bootstrap improve rapidly. Figures A.5 and A.6 are similar to Figure 4, with 2000 observations and 50 equal-sized clusters. They show the densities of t_4 when $G_1 = 3$ and $G_1 = 5$, respectively, along with the densities of the WCR and WCU bootstrap statistics for the particular samples for which the realized test statistic $\hat{t}_4$ is the 0.975 quantile of the t_4 distribution.[2]

In Figure A.5, neither bootstrap density provides a good approximation, but both are appreciably better than those for $G_1 = 1$. Interestingly, due presumably to the use of the Rademacher distribution, the WCR bootstrap density is now trimodal. It underrejects

[2] Figures A.5 and A.6, like Figures 4 and 5, are based on simulations that mistakenly set $\rho_\epsilon = 0.0025$. The value of ρ_ϵ has no effect on the key features of these figures.

Figure A.5: Densities of actual and bootstrap t statistics when $G_1 = 3$

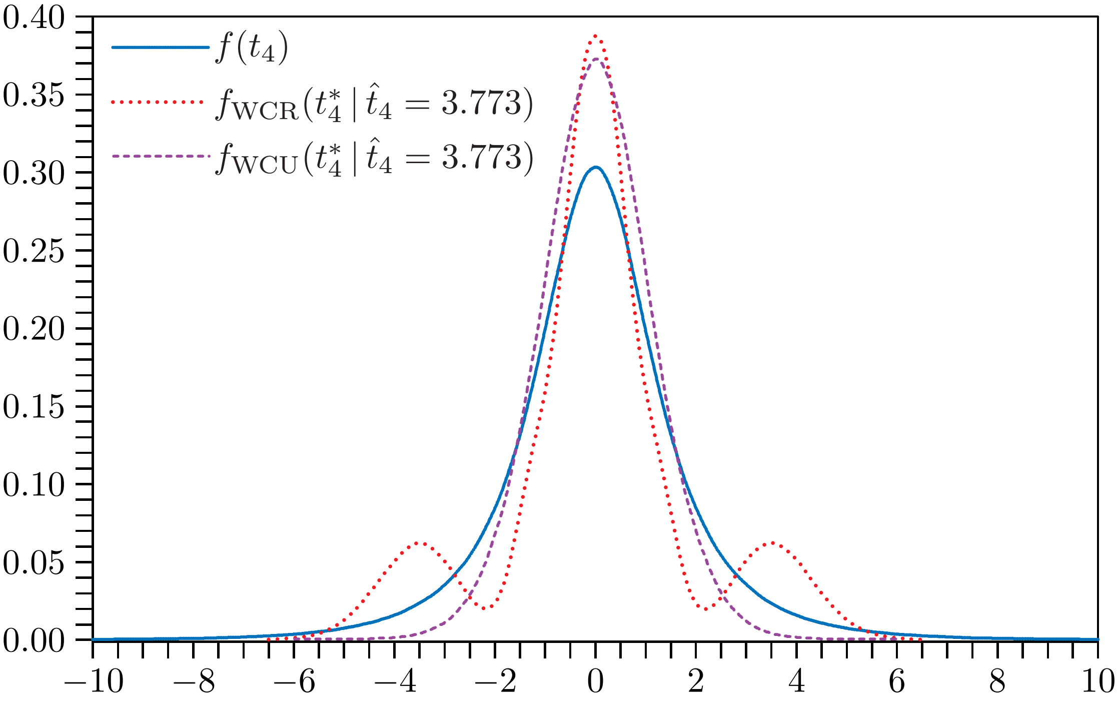
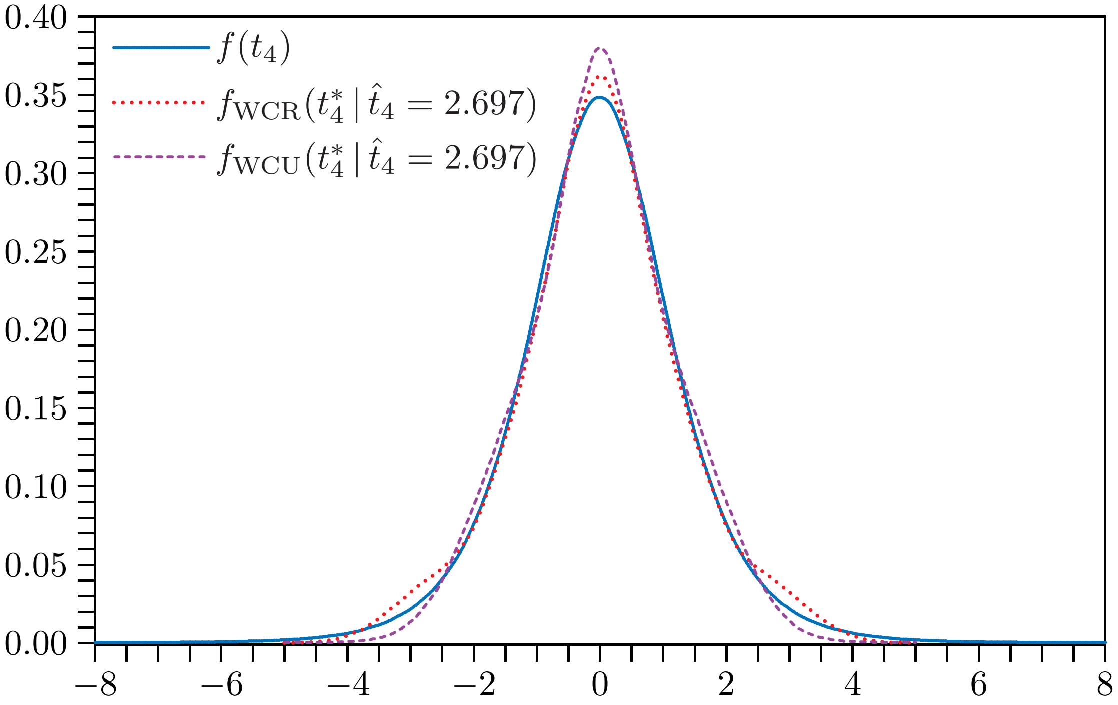


Figure A.6: Densities of actual and bootstrap t statistics when $G_1 = 5$



because the two outer modes give it too much mass for $|t_4^*| > |t_4|$, and those modes move away from the origin as $|t_4|$ increases. The outer modes are associated with realizations of the Rademacher random variables that are either 1 or -1 for all three treated clusters. The WCU bootstrap density has much thinner tails than the actual density, which explains why tests based on it overreject severely.

In Figure A.6, both bootstrap densities perform very much better than in Figure A.5. Moving from $G_1 = 3$ to $G_1 = 5$ evidently makes a big difference. The WCR bootstrap density is no longer trimodal, although it does have two bulges which presumably arise for the same reason. The WCU bootstrap density still has thinner tails than the actual density, but to a lesser extent.

A.6 Standard Asymptotics

In Section 6, we showed why inference based on CRVE t statistics fails asymptotically when G_1 is fixed and why the bootstrap fails to solve the problem. In this section, we briefly discuss why these issues do not occur under a standard asymptotic construction. Instead of holding G_1 fixed as $N \to \infty$, we make the more conventional assumption that $\phi = G_1/G$ tends to a constant that is strictly between 0 and 1. The simplest way to do this would be to consider only samples that are integer multiples of the original sample size, with regressor matrices that simply repeat the original one, so that ϕ and $\bar{d}$ are the same for every sample. Alternatively, we could allow both ϕ and $\bar{d}$ to vary somewhat as N increases, provided that $\bar{d} \to d_\infty$ and $\phi \to \phi_\infty$ as $N \to \infty$, with $0 < d_\infty < 1$ and $0 < \phi_\infty < 1$. We continue to assume, for simplicity, that G/N tends to a constant as $N \to \infty$.

Under the null hypothesis, the CRVE statistic is given by equation (12), which we rewrite here for convenience:

$$t_2 = \frac{c(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\epsilon}}{\Big(\sum_{g=1}^{G}(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)'\hat{\boldsymbol{\epsilon}}_g\hat{\boldsymbol{\epsilon}}_g'(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)\Big)^{1/2}}. \tag{A.5}$$

Ignoring the scalar factor c, the numerator of this test statistic is

$$(1-\bar{d})\sum_{g=1}^{G_1}\sum_{i=1}^{N_g}\epsilon_{ig} - \bar{d}\sum_{g=G_1+1}^{G}\sum_{i=1}^{N_g}\epsilon_{ig}. \tag{A.6}$$

Since $\bar{d}$ tends to a constant between 0 and 1, and the ϵ_{ig} have mean zero, both terms here are evidently $O_p(N^{1/2})$, and so is their weighted sum.

The square of the denominator of (A.5) is given by expression (13), which is

$$(1-\bar{d})^2\sum_{g=1}^{G_1}\bigg(\sum_{i=1}^{N_g}\hat{\epsilon}_{ig}\bigg)^2 + \bar{d}^2\sum_{g=G_1+1}^{G}\bigg(\sum_{i=1}^{N_g}\hat{\epsilon}_{ig}\bigg)^2. \tag{A.7}$$

Both terms here are evidently $O_p(N)$, and so is their weighted sum. By the results of CSS, N^{-1} times expression (A.7) consistently estimates the variance of $N^{-1/2}$ times expression (A.6). Thus, provided we can apply a central limit theorem to the latter, we obtain the standard result that $t_2 \overset{a}{\sim} \mathrm{N}(0,1)$.

For the bootstrap to be asymptotically valid in this case, the distribution of the bootstrap test statistic t_2^{*j}, which is $\hat{\beta}_2^{*j}$ divided by the bootstrap CRVE standard error, must be asymptotically standard normal. This test statistic was given in (16) and is repeated here for convenience:

$$t_2^{*j} = \frac{c(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\epsilon}^{*j}}{\left(\sum_{g=1}^{G}(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)'\hat{\boldsymbol{\epsilon}}_g^{*j}(\hat{\boldsymbol{\epsilon}}_g^{*j})'(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)\right)^{1/2}}. \tag{A.8}$$

Here $\boldsymbol{\epsilon}^{*j}$ is the vector of bootstrap error terms for bootstrap sample j, and $\hat{\boldsymbol{\epsilon}}_g^{*j}$ is the subvector of the bootstrap residual vector $\hat{\boldsymbol{\epsilon}}^{*j}$ corresponding to cluster g. These bootstrap residual vectors are obtained by regressing the $\boldsymbol{y}^{*j}$ on $\boldsymbol{\iota}$ and $\boldsymbol{d}$.

The arguments needed to show that t_2^{*j} is asymptotically standard normal are very similar to the ones for t_2 itself. Under the null hypothesis, the residual vectors $\tilde{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{\epsilon}}$ both converge to the error vector $\boldsymbol{\epsilon}$ as $N \to \infty$. Since the bootstrap DGP (assume the Rademacher distribution for simplicity) merely changes the signs of all the residuals for each cluster with probability one-half, the bootstrap error vectors $\boldsymbol{\epsilon}^{*j}$ must asymptotically follow exactly the same distribution as $\boldsymbol{\epsilon}$ if that distribution is symmetric.[3] Similarly, the bootstrap residual vectors $\hat{\boldsymbol{\epsilon}}^{*j}$ must follow the same distributions asymptotically as the actual residual vectors $\hat{\boldsymbol{\epsilon}}$. Thus everything that is true asymptotically for the actual t statistics (A.5) is also true for the bootstrap t statistics (A.8). We conclude that, under the standard asymptotic construction of this section, t_2 and the t_2^{*j} both follow the standard normal distribution asymptotically.

The crucial feature of the standard asymptotic construction is that $\bar{d} \to d_\infty$ with $0 < d_\infty < 1$ as $N \to \infty$. If instead $\bar{d} \to 0$, as it does when G_1 is fixed, or $\bar{d} \to 1$, as it does when $G_0 = G - G_1$ is fixed, then the two terms in each of (A.6) and (A.7), and their bootstrap analogs, are of different orders. When G_1 is fixed, the first term in expression (A.6) is the product of two factors, each of which is $O_p(1)$, while the second term is $O_p(N^{-1})O_p(N^{1/2}) = O_p(N^{-1/2})$. Similarly, the first term in expression (A.7) is $O_p(1)$, while the second term is $O_p(N^{-2})O_p(N) = O_p(N^{-1})$. Thus standard asymptotic arguments break down. Indeed, as was shown in Section 6, $\hat{\beta}_2$ is actually inconsistent.

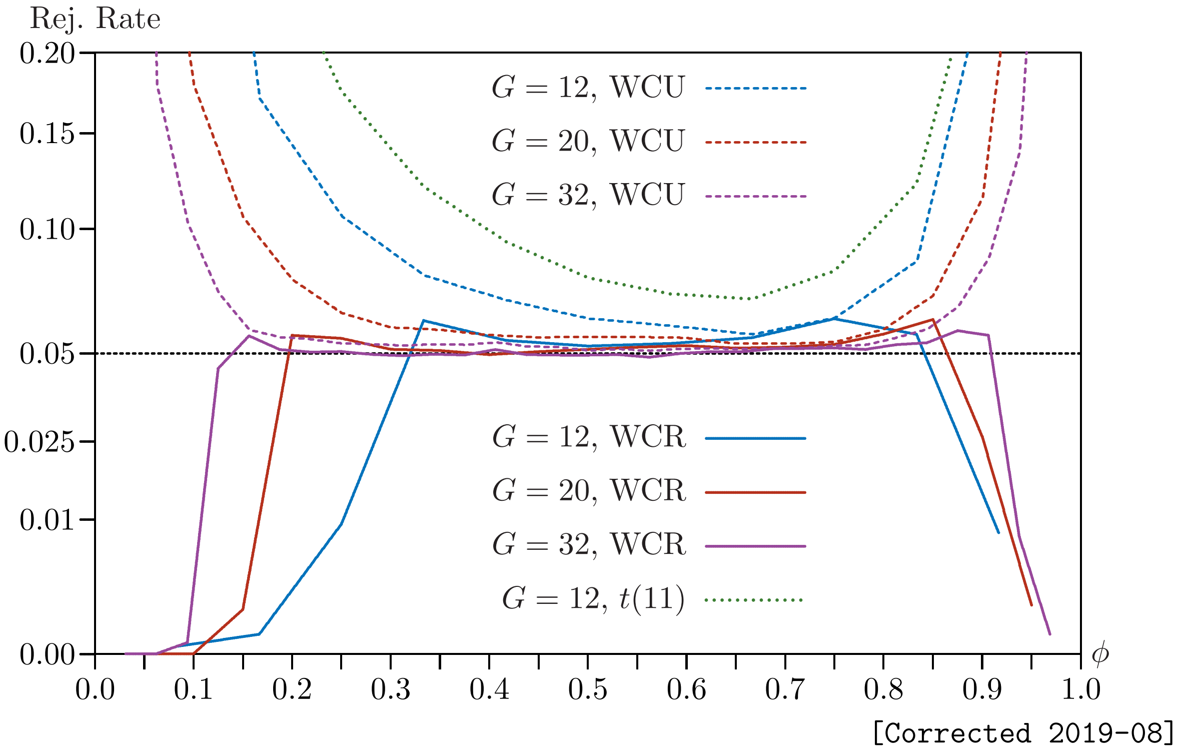
A.7 Few Unbalanced Clusters

In all but two of the experiments reported in the paper, the number of clusters G is at least 50, because in many of them cluster sizes are proportional to state populations. In applied work, however, the number of clusters is often substantially less than 50. In this section, we present some results for smaller values of G and for cluster sizes that vary in different ways. The results we report are just a few of the many we obtained, but they are fairly representative.

The model is the DiD regression given in equation (7). Half the observations in treated clusters are treated, and $\rho_\epsilon = 0.05$. Experiments with 400,000 replications and 399 bootstrap samples are performed for five values of G: 12, 16, 20, 25, and 32. There are $N = 50G$ observations. In order to allow for unbalanced cluster sizes, N_g is determined by a parameter

[3] Even if the ϵ_{ig} were not symmetrically distributed, the wild cluster bootstrap using the Rademacher distribution would still be asymptotically valid, but the argument would be a bit more complicated.

Figure A.7: Rejection frequencies for bootstrap tests when G is small and $\gamma = 3$



$\gamma \geq 0$, as follows:

$$N_g = \left[N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right], \quad g = 1, \ldots, G-1, \tag{A.9}$$

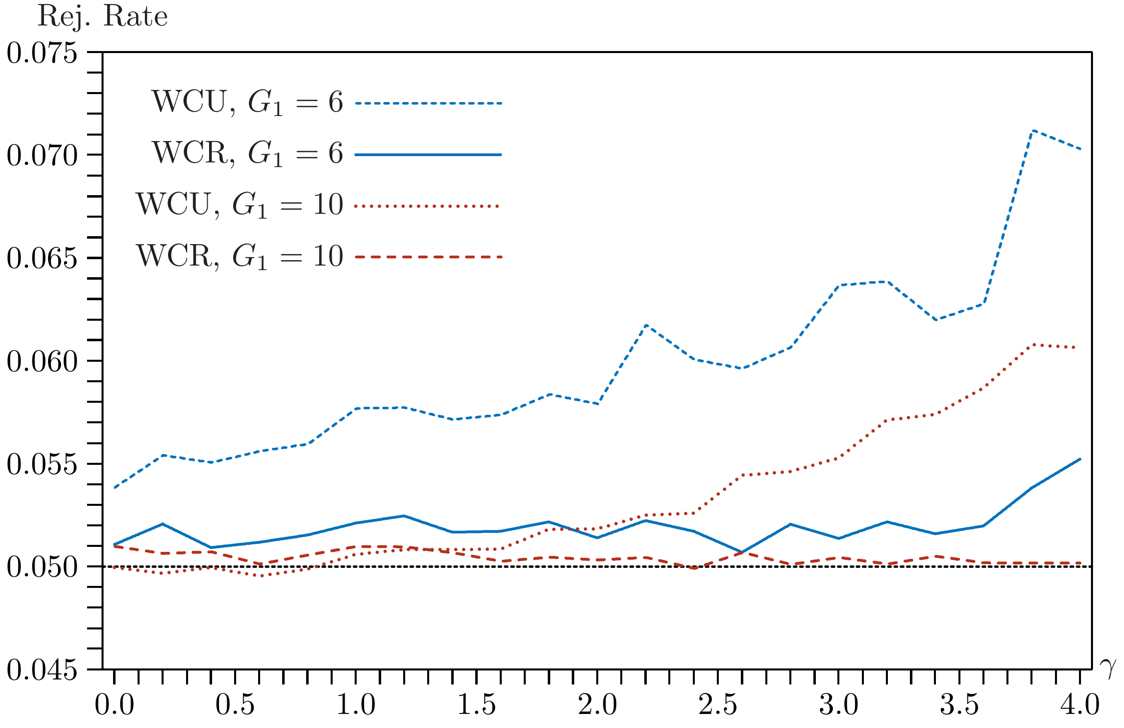
where $[\cdot]$ denotes the integer part of its argument, and $N_G = N - \sum_{j=1}^{G-1} N_g$. When $\gamma = 0$, every N_g is equal to $N/G = 50$. As γ increases, cluster sizes become increasingly unbalanced.

Figure A.7 shows rejection frequencies at the .05 level for WCR and WCU bootstrap tests for three values of G (12, 20, and 32) when $\gamma = 3$ as functions of $\phi = G_1/G$. Cluster sizes are quite unbalanced. The smallest is always 8, and the largest is either 146, 155, or 167 for $G = 12$, 20, or 32, respectively. Clusters are treated from smallest to largest, and the pronounced asymmetry that is evident in the figure reflects this. As in Figure A.4, the vertical axis is subject to a square root transformation.

The WCR bootstrap results in Figure A.7 are quite similar to the ones in Figure 3. There is severe underrejection for extreme values of ϕ. As expected, the range of values of ϕ for which the bootstrap performs acceptably becomes wider as G increases. For $G = 20$ and $G = 32$, this range is $6 \leq G_1 \leq G - 5$. Whether there is any such range for $G = 12$ depends on how one defines “acceptably.” The rejection frequencies fall between 0.053 and 0.055 when $5 \leq G_1 \leq 8$, and this might be considered acceptable.

The WCU bootstrap results in Figure A.7 are also as expected. There is always severe overrejection for extreme values of ϕ. For clarity, the vertical axis has been truncated at 0.20, and rejection frequencies for $G_1 = 1$ and $G_1 = G - 1$ are always much greater than this. For $G = 12$, the WCU bootstrap always overrejects noticeably. For $G = 20$ and $G = 32$, it

Figure A.8: Rejection frequencies for bootstrap tests as functions of γ for $G = 20$



[Corrected 2019-08]

almost always rejects more often than the WCR bootstrap, but it performs acceptably for some intermediate values of ϕ.

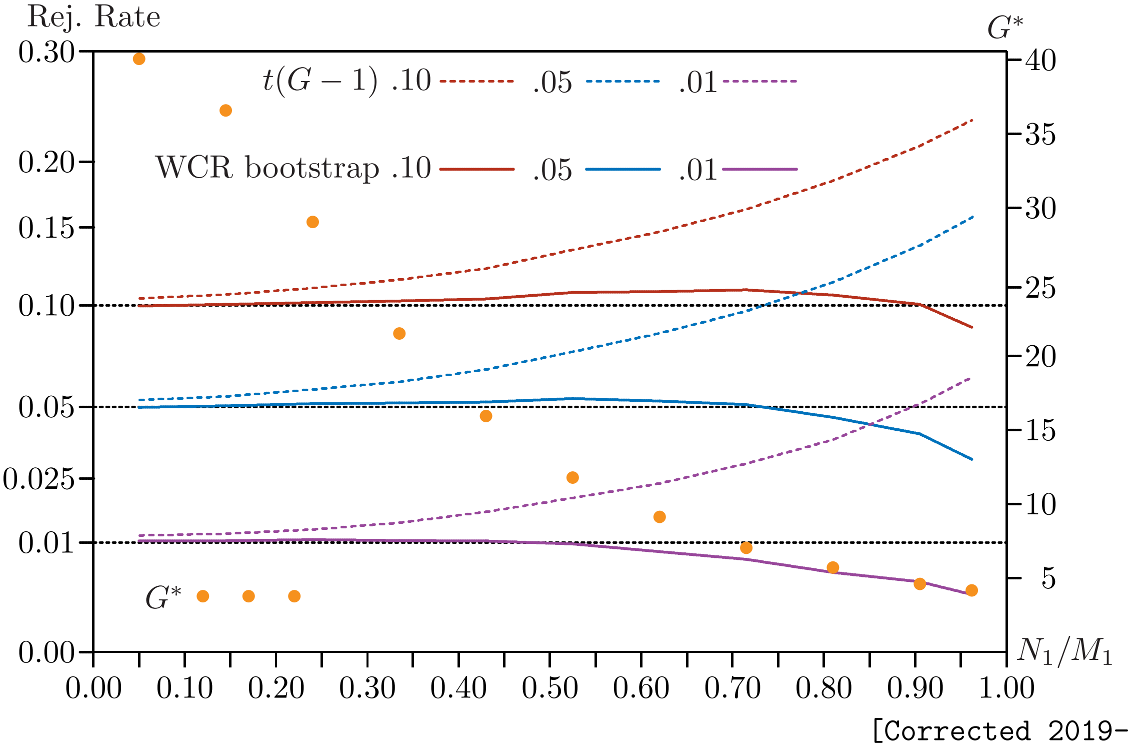
Rejection frequencies for tests based on the $t(G-1)$ distribution have a "U" shape similar to the ones for the WCU bootstrap, but they are always larger (often much larger) than the latter. Results for $G = 12$ are shown in the figure. As the analysis in Section 6 predicts, overrejection is more severe for small values of ϕ than for large values, because clusters are being treated from smallest to largest.

Figure A.8 graphs rejection frequencies against γ for two cases with $G = 20$. In one case, $\phi = 0.3$, so that $G_1 = 6$. In the other, $\phi = 0.5$, so that $G_1 = 10$. The former is just at the edge of the region where the bootstrap seems to work acceptably, while the latter is well inside that region. In both cases, only 20% of the observations in the treated clusters are treated. Two patterns are evident from the figure. The WCR bootstrap almost always outperforms the WCU bootstrap, which overrejects in a large majority of cases. This overrejection becomes more severe as γ increases, probably because the sizes of the treated clusters become more dispersed. In contrast, the WCR bootstrap is much less sensitive to γ, although it clearly performs better for $\phi = 0.5$ than for $\phi = 0.3$.[4]

Equation (A.2) strongly suggests that inference based on CRVE t statistics, and probably also bootstrap inference, will become less reliable as the sizes of the treated clusters

[4] The somewhat ragged shapes of the rejection frequency curves in Figure A.8, especially when $G_1 = 6$, are not primarily due to experimental randomness. They arise because, as γ increases, the numbers of treated observations in clusters of various sizes change in a somewhat irregular fashion.

Figure A.9: Rejection frequencies as functions of N_1/M_1



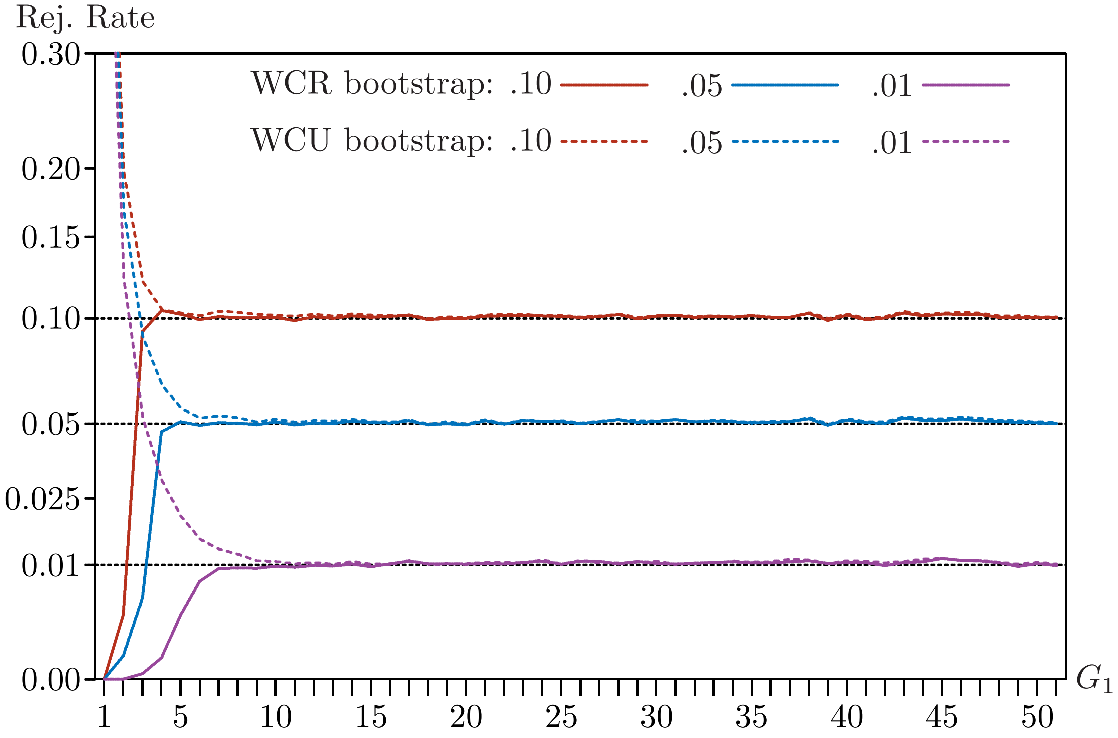
become more variable. In order to investigate this conjecture, we performed one more set of experiments. The model is still the DiD regression of equation (7), with $N = 2000$, $G = 40$, and $G_1 = 20$ in all cases. The 20 untreated clusters each have 50 observations. What varies across the experiments are the sizes of the 20 treated clusters. One cluster, numbered 1, has N_1 observations, with N_1 varying between 50 and 962, and the remaining 19 treated clusters each have $(1000 - N_1)/19$ observations. There are 100,000 replications.

Figure A.9 shows rejection frequencies for CRVE t tests using $t(39)$ critical values and for WCR bootstrap tests at the .01, .05, and .10 levels. When N_1/M_1 is small or moderate in size, both tests perform well, as we would expect for such a large value of G_1. But as N_1/M_1 becomes larger, the performance of both tests deteriorates. The CRVE t tests overreject quite severely, and the WCR bootstrap tests may either overreject or underreject. The latter underrejects quite severely at the .01 level when $N_1/M_1 > 0.75$.

The figure also shows the values of G^*, calculated for the true value of $\rho = 0.05$, for each value of N_1/M_1; the scale is shown on the right-hand vertical axis. These values decline monotonically with N_1/M_1, eventually becoming very much smaller than $G_1 = 20$. These results, and others not reported, suggest that G^* is of considerable diagnostic value, with small values of G^* being associated with poor CRVE test performance.

The experiments of this section yield two principal results beyond the ones in the paper. The first is that wild cluster bootstrap tests perform very well indeed, even when G is very small and clusters are quite unbalanced, provided neither G_1 nor $G - G_1$ is too small and the sizes of the treated clusters are not extremely variable. The second is that the restricted wild cluster bootstrap (WCR) almost always outperforms the unrestricted one (WCU) when

Figure A.10: Rejection frequencies for placebo laws using aggregate data



they both perform reasonably well. The latter usually rejects more often than the former.

A.8 Placebo Laws with Aggregate Data

Figure 10 presents some results for placebo laws experiments using aggregate data. It shows rejection frequencies for CRVE tests based on the $t(G-1)$ distribution and for restricted wild cluster bootstrap tests. However, because of the scale of the vertical axis, the WCR results are difficult to see clearly.

Figure A.10 is comparable to Figure A.4. Like the latter, it shows rejection frequencies for WCR and WCU bootstrap tests, but for aggregate rather than micro data. These are based on 100,000 replications with 399 bootstraps. It is evident that both bootstrap methods work extremely well for G_1 sufficiently large. The minimum value of G_1 that is needed seems to be larger for tests at the .01 level than for tests at higher levels, and perhaps slightly larger for WCU than for WCR.

As the analysis in Section 6 predicts, WCR underrejects severely, and WCU overrejects severely, when G_1 is small. For clarity, rejection frequencies for WCU when $G_1 = 1$ are not shown in the figure. They are 0.7522, 0.7184, and 0.6399 for tests at the .10, .05, and .01 levels, respectively. These are just a little bit lower than the rejection frequencies for CRVE t tests based on $t(G-1)$ critical values that are plotted in Figure 8.

A.9 Aggregation and Test Power

In Section 7, we showed that using aggregate data obtained by averaging over state-year pairs (or, in general, cluster-period pairs) can lead to better finite-sample performance under the null hypothesis. It does so by imposing perfectly balanced cluster sizes. This approach could evidently lead to power loss if the estimates based on aggregate data were less efficient than the ones based on micro data. That seems to be what happens in the empirical example of Section 8. We discuss the issue of power loss in this section and present some simulation evidence.

Consider the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where each observation is associated with a time period t as well as a cluster g. Let the number of time periods be T, and define $M \equiv GT$. Further, let M_m denote the number of observations for cell m, where $m = T(g-1)+t$ if the cells are ordered by cluster and then time period. Define the $M \times N$ averaging matrix $\boldsymbol{A}$ as the matrix with $1/M_m$ in every element of row m that corresponds to an observation belonging to that cell. Then $\boldsymbol{Ay}$ is an $M \times 1$ vector with typical element

$$\bar{y}_m = \frac{1}{M_m}\sum_{i=1}^{M_m} y_{gti},$$

where i indexes the observations associated with cell m. Similarly, $\boldsymbol{AX}$ is an $M \times k$ matrix with typical row $\bar{\boldsymbol{X}}_m$ consisting of averages of the $\boldsymbol{X}_{gti}$ over the observations in cell m.

The OLS estimator for the aggregate data is

$$\acute{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{A}'\boldsymbol{AX})^{-1}\boldsymbol{X}'\boldsymbol{A}'\boldsymbol{Ay},$$

of which the covariance matrix is

$$(\boldsymbol{X}'\boldsymbol{A}'\boldsymbol{AX})^{-1}\boldsymbol{X}'\boldsymbol{A}'\boldsymbol{A\Omega A}'\boldsymbol{AX}(\boldsymbol{X}'\boldsymbol{A}'\boldsymbol{AX})^{-1}. \tag{A.10}$$

This may be compared with the covariance matrix for the OLS estimator $\hat{\boldsymbol{\beta}}$ based on the original micro data:

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega X}(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{A.11}$$

Although it is easy to compare expressions (A.10) and (A.11) numerically, it is impossible to compare them analytically except in very special cases.

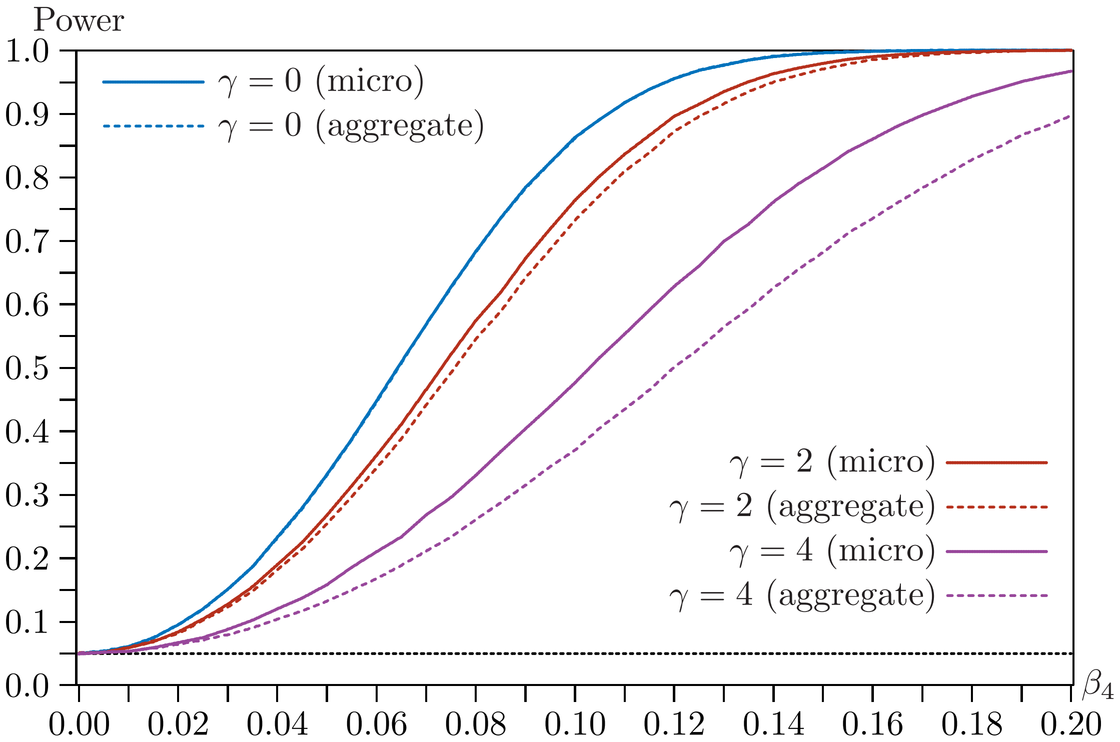
One such case is the dummy variable regression model discussed in Section 6 when every cell has the same number of elements. In this case,

$$\acute{\beta}_2 = \frac{(\boldsymbol{Ad} - \bar{d}\boldsymbol{A\iota})'\boldsymbol{Ay}}{(\boldsymbol{Ad} - \bar{d}\boldsymbol{A\iota})'(\boldsymbol{Ad} - \bar{d}\boldsymbol{A\iota})} = \frac{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{A}'\boldsymbol{Ay}}{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{A}'\boldsymbol{A}(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})}. \tag{A.12}$$

Note that $\bar{d}$ here is the sample mean of the elements of $\boldsymbol{Ad}$. It is the same as the $\bar{d}$ in equation (A.6), but that would not be true if every cell did not have the same number of elements. In this special case,

$$(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{A}'\boldsymbol{Ay} = (M/N)(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{y},$$

Figure A.11: Bootstrap power functions for several values of γ



[Corrected 2019-08]

and

$$(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{A}'\boldsymbol{A}(\boldsymbol{d} - \bar{d}\boldsymbol{\iota}) = (M/N)(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'(\boldsymbol{d} - \bar{d}\boldsymbol{\iota}),$$

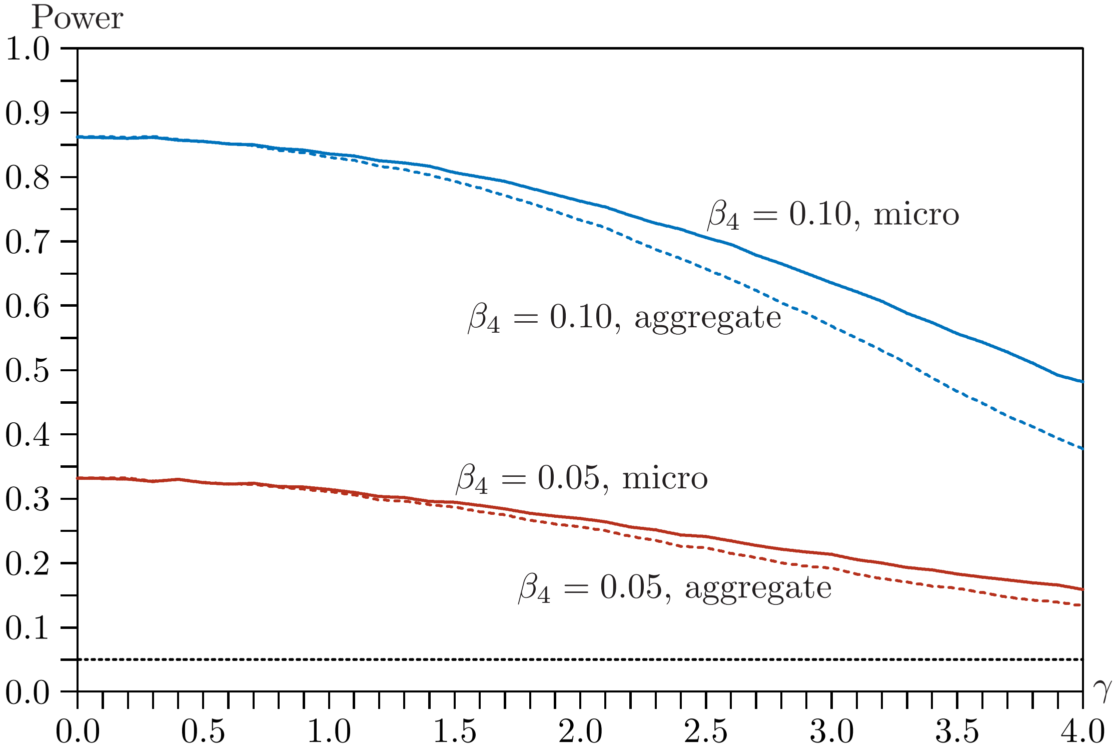
so that the aggregate estimator $\acute{\beta}_2$ given in equation (A.12) is equal to the micro estimator $\hat{\beta}_2$ given in equation (9).

The result that the two estimators are numerically identical is very special. It no longer holds if every cell does not have the same number of elements or if any regressor varies within cells. In fact, when regressors vary within cells, the efficiency loss from aggregation can be severe. For example, when the wage regression for the placebo laws data is estimated with no treatment variable using micro data, the coefficient on age is 0.02462 with a CRVE standard error of 0.00266. When the same regression is estimated using aggregate data, the coefficient is 0.02988 with a standard error of 0.04874. The aggregate standard error is 18.3 times larger than the micro one.

However, since treatment variables typically do not vary within cells, because they apply to every unit in a given cluster for a given time period, it is not clear whether we should expect significant power loss when using aggregate data for DiD regressions. Since neither $\hat{\boldsymbol{\beta}}$ nor $\acute{\boldsymbol{\beta}}$ is an efficient estimator, it is not even clear that the former will be more efficient than the latter.

In order to investigate this issue, we perform a series of experiments based on the DiD regression of equation (7), with 100,000 replications. Cluster sizes are determined by equation (A.9), which depends on a parameter $\gamma \geq 0$. In the experiments, $G = 50$, $N = 15,000$,

Figure A.12: Bootstrap power as a function of γ for two values of β_4



[Corrected 2019-08]

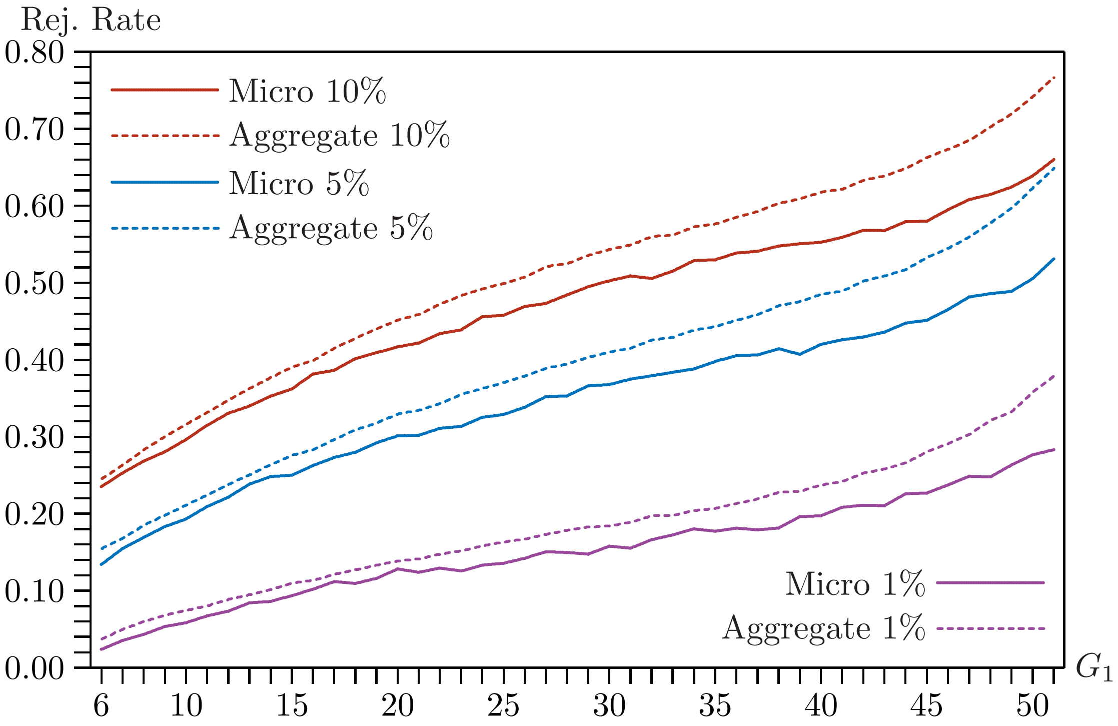
$\rho_\epsilon = 0.05$, and every observation is assigned, with equal frequency, to one of 20 years. Half of the clusters are treated for 8 out of the 20 years. The micro regressions have 15,000 observations, and the aggregate ones have 1000 observations. The value of N is quite large here because we do not want there to be any empty cells. With $\gamma = 4$, there would be empty cells if N were much smaller than $15,000$. When $\gamma = 0$, every cluster has 300 observations, so each cell is an average of 15 observations. As γ increases, cluster sizes become more variable, and so do the numbers of observations per cell. For the largest value of γ that we investigate ($\gamma = 4$), some cells are averages of just one observation, and others are averages of 60.

Figure A.11 shows power functions for bootstrap (WCR) tests for three values of γ. Not surprisingly, there is no power loss when $\gamma = 0$. With equal-sized clusters, homoskedasticity, and no within-cluster variation, the aggregate estimates are very similar (but, in this case, not identical) to the micro estimates. As γ increases, however, power declines, and the power loss from using aggregate data becomes steadily greater.

Both of these results should have been expected. Because of the intra-cluster correlation, the OLS estimates are most efficient when cluster sizes are balanced, so power inevitably declines as γ increases. The aggregate estimates give as much weight to cells based on few observations as to cells based on many observations, and this presumably causes them to be less efficient than the micro estimates.

Another way to see how power loss changes as clusters become more unbalanced is to

Figure A.13: Bootstrap power as a function of G_1 for placebo laws



graph power against γ for a given value of β_4. This is done in Figure A.12, which shows the power of WCR bootstrap tests for two values of β_4 (0.05 and 0.10) as functions of γ for $0 \leq \gamma \leq 4$. Power loss is modest for $\gamma \leq 1$ and quite substantial for $\gamma \geq 2$. This is what we would expect to see in view of the results in Figure A.11.

A very different way to investigate power is to modify the placebo laws experiments of Section 7 by adding a small amount to the regressand whenever an observation or cell is treated. In our experiments, we added 0.02, which is equivalent to increasing wages by 2.02%. We report results only for the WCR bootstrap, because it is generally the most reliable procedure for controlling size. Our experiments used 100,000 replications for the aggregate data and 10,000 for the micro data, with 399 bootstraps in both cases. When we did this for $G_1 \geq 6$ (because the wild cluster bootstrap is not reliable for $G_1 \leq 5$), we obtained the results shown in Figure A.13.

Figure A.13 is surprising. If it is to be believed, aggregation actually increases power, especially when a large number of states is treated. The reason seems to be that, contrary to normal intuition, the micro estimates of β_{treat} are a bit more dispersed than the aggregate estimates. For example, when $G_1 = 20$, the standard errors are 0.01345 and 0.01287, respectively.

Whether the apparent gain in power from aggregation that appears in Figure A.13 should be taken seriously is not clear. Normally, in an experiment designed to measure power, we would hold the DGP constant and draw many realizations of the error terms and, hence, the dependent variable. In the placebo laws experiments, however, we hold most of the data constant and draw many different realizations of the test regressor, arbitrarily adding 0.02 to the regressand whenever the test regressor is equal to 1. Thus the distributions of the estimated coefficients and the distributions of the test statistics are not really the ones that would normally determine test power.

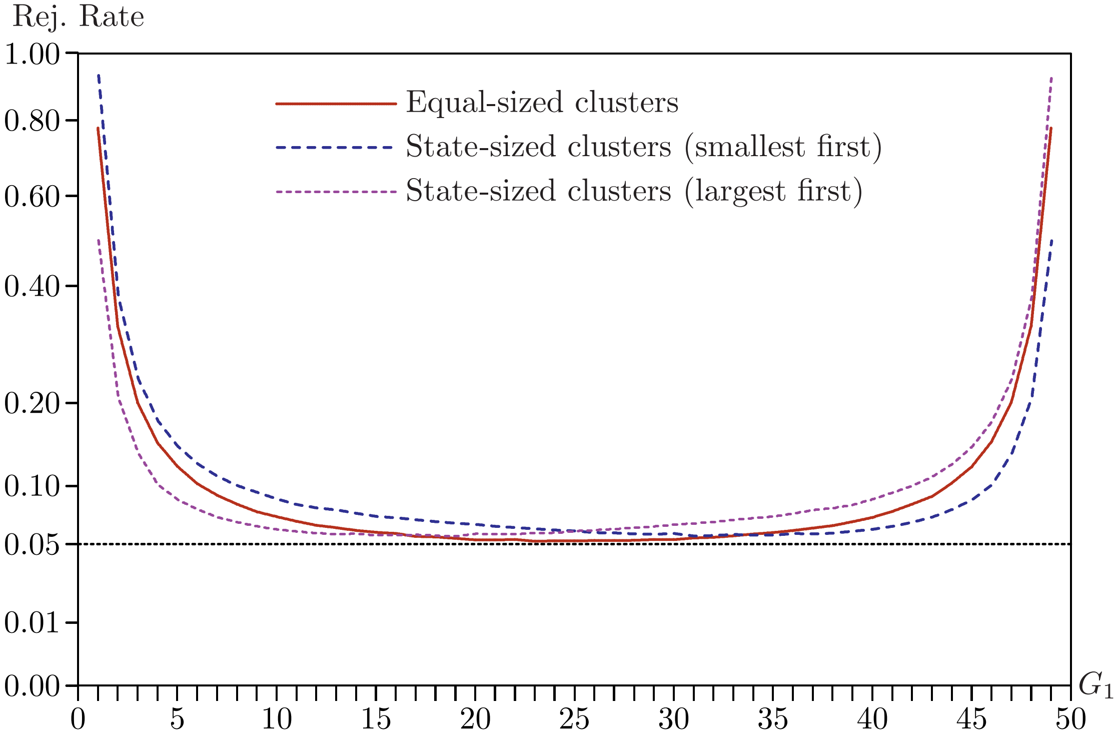
Based on these results, we might be tempted to conclude that, in cases where the test regressor does not vary within cells, aggregation is not likely to lead to very much loss of power. However, the empirical results in Section 8 suggest otherwise. Whether or not power loss due to aggregation is a problem in practice is often very easy to determine. If a test using aggregate data rejects the null hypothesis convincingly, then lack of power is evidently not a problem. Conversely, if a test using micro data that should be reasonably reliable, such as a WCR bootstrap test, rejects the null hypothesis, and a similar test using aggregate data does not, then lack of power is evidently a problem.

Additional Reference

Davidson R, MacKinnon JG. The size distortion of bootstrap tests. *Econometric Theory* **15**: 361–376.

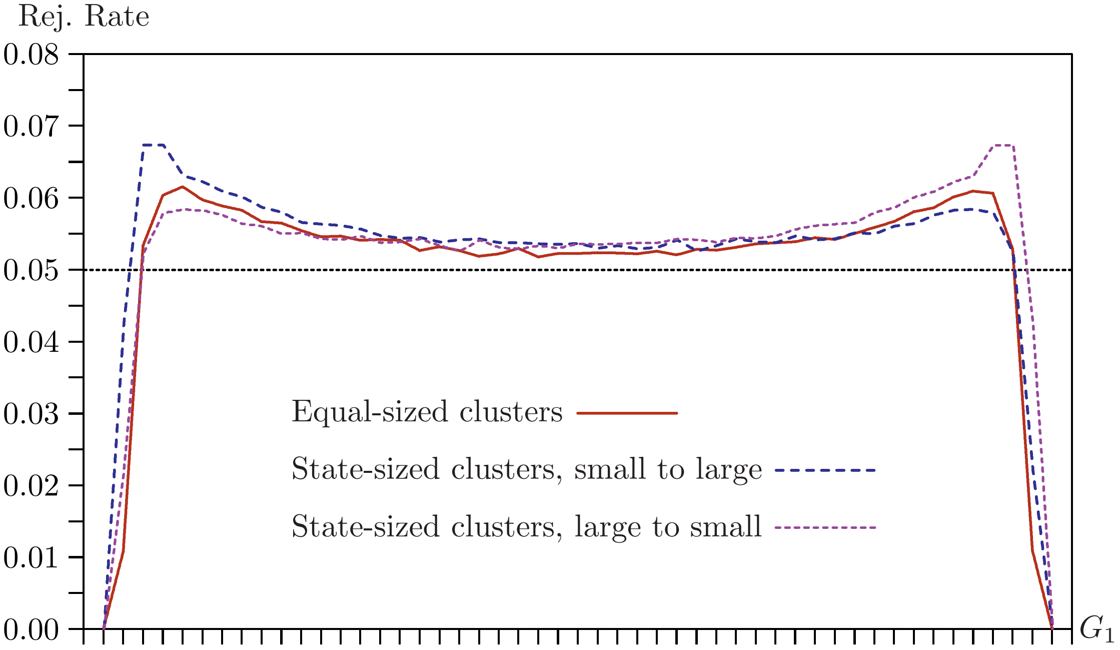
A.10 Corrected Figures from the Paper

Figure 1: Rejection rates and proportion treated, DiD, $t(G-1)$)



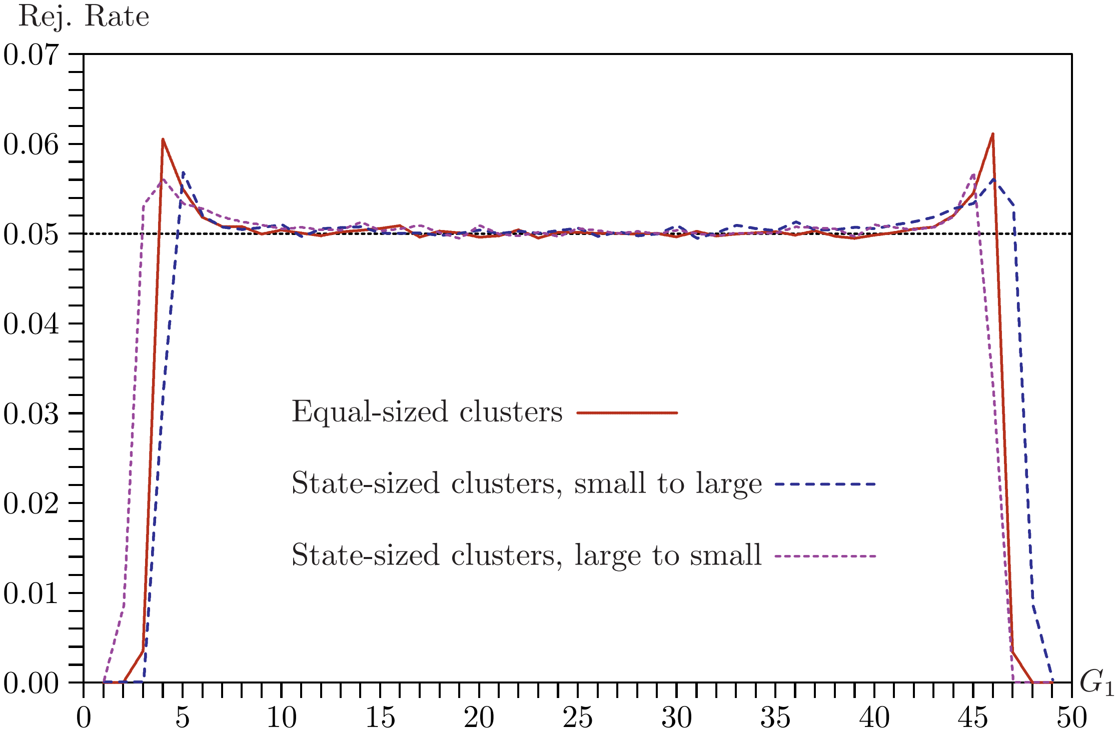
[Corrected 2019-08]

Figure 2: Rejection rates and proportion treated, DiD, $t(G^* - 1)$



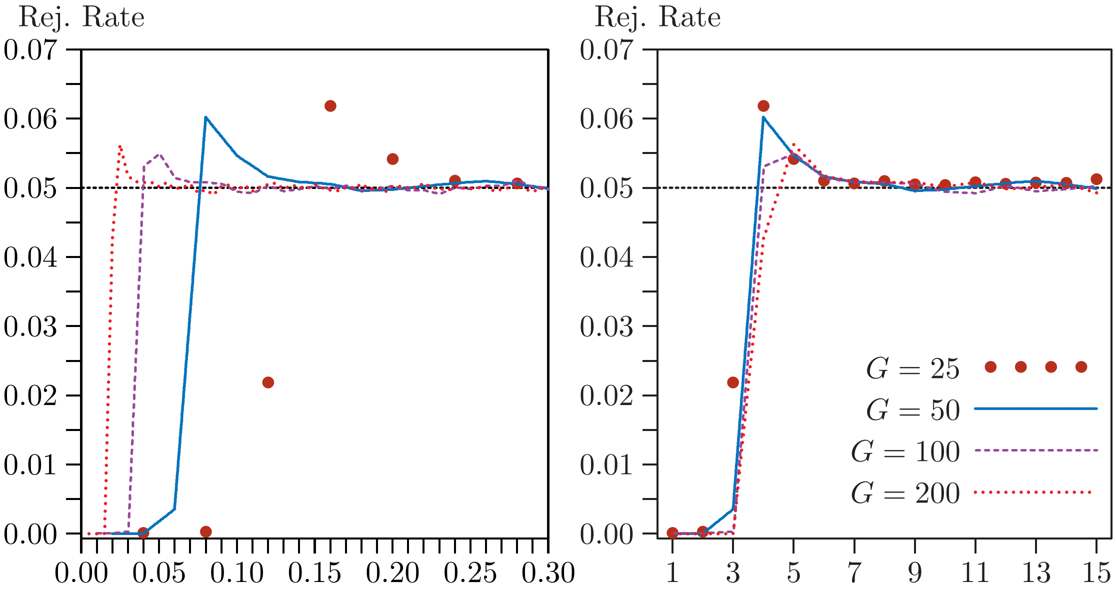
[Corrected 2019-08]

Figure 3: Rejection rates and proportion treated, DiD, wild bootstrap



[Corrected 2019-08]

Figure 6: Restricted wild cluster bootstrap rejection frequencies

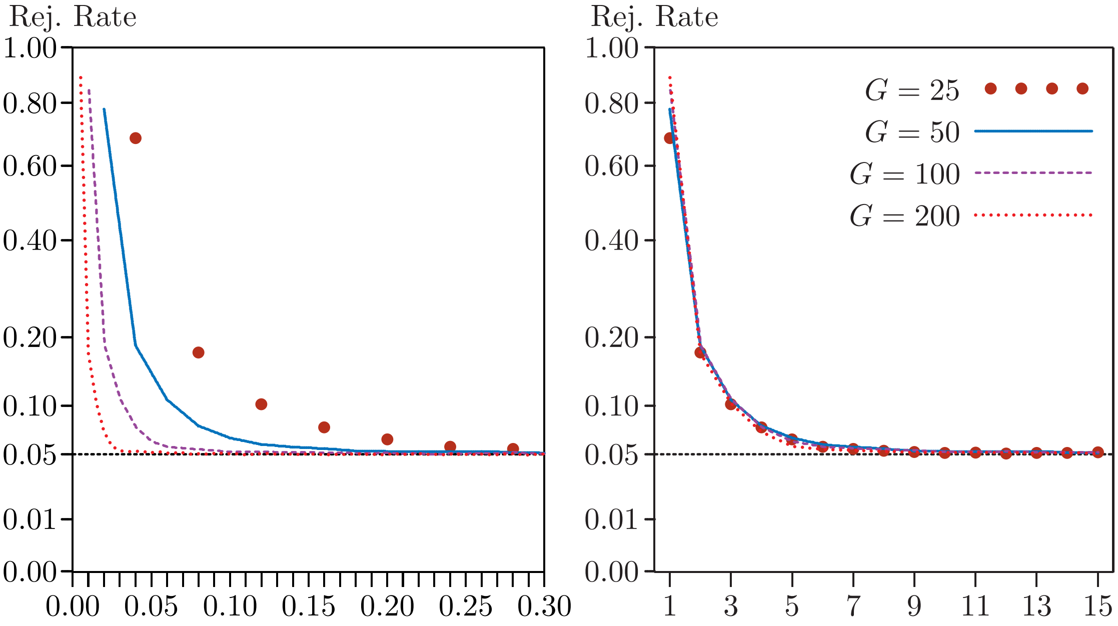


A. Rejection frequencies versus ϕ

B. Rejection frequencies versus G_1

[Corrected 2019-08]

Figure 7: Unrestricted wild cluster bootstrap rejection frequencies



A. Rejection frequencies versus ϕ

B. Rejection frequencies versus G_1

[Corrected 2019-08]