

Wild Bootstrap Inference for Wildly Different Cluster Sizes

Matthew D. Webb

October 9, 2013

- This paper is joint with:

James G. MacKinnon
Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

- Contributions:
 - Presents Monte Carlo evidence that overturns the 'rule of 42'
 - Shows that DiD and CRVE estimation works poorly when the proportion of clusters treated is very small or very large

Standard Method for Difference-in-Differences

- Suppose you want to estimate the following linear difference-in-differences equation

$$Y_{igt} = \beta_0 + \beta_1 * treat_g + \beta_2 * year_t + \beta_3 * treat_g * year_t + X_{igt}\gamma + \epsilon_{igt} \quad (1)$$

- Y_{igt} observation for person i in group g and time t
- $treat_g$ dummy for if the person is in the treatment group
- $year_t$ dummy if in time period after treatment
- X_{igt} other independent variables
- $treat_g * year_t$ is the DiD term

Estimating β

- We are interested in inference for the OLS estimate of $\hat{\beta}_3$
- With the assumptions that data are independent over g , but errors are correlated within cluster
- $E[u_g] = 0$
- $E[u_g u_g'] = \Sigma_g$, violates the i.i.d. assumption
- $E[u_g u_h'] = 0$ for cluster $h \neq g$
- When then have
- $\sqrt{N}(\hat{\beta} - \beta) \sim \mathcal{N}[0, NV[\hat{\beta}]]$

Cluster Robust Variance Estimator (CRVE)

$$\hat{V}_{CR}[\hat{\beta}] = (X'X)^{-1} \left(\sum_{g=1}^G X_g \tilde{u}_g \tilde{u}_g' X_g' \right) (X'X)^{-1}$$

- in the simplest case the OLS residuals are used $\tilde{u}_g = \hat{u}_g = y_g - X_g \beta$
- in other cases $\sum_{g=1}^G X_g \hat{u}_g \hat{u}_g' X_g'$, is replaced by $\sum_{g=1}^G \tilde{U}_g \tilde{U}_g'$
- Stata uses:

$$\tilde{U}_g = \sum_{i=1}^{N_g} \hat{u}_{ig} \begin{pmatrix} 1 \\ X_g \end{pmatrix},$$

Asymptotics Underlying CRVE

General results on covariance matrix estimation in White (1984) imply of the CRVE is consistent under three key assumptions:

- A1. The number of clusters goes to infinity.
- A2. The within-cluster correlation is constant across clusters.
- A3. The individual clusters contain an equal number of observations.

Carter, Schnepel and Steigerwald (2012) relax A1 and A2.

This talk concerns A3.

- Clustered Errors
 - Kloek (1981)
 - Moulton (1990)
- Inference in Difference-in-Difference
 - Conley and Taber (2011)
 - Donald and Lang (2007)
- Bootstrap Inference in Difference-in-Differences
 - Bertrand, Duflo and Mullainathan (2004)
 - Cameron, Gelbach and Miller (2008)
 - Webb (2013)

Rejection Frequencies by Number of Clusters

	Number of Groups (G)					
	5	10	15	20	25	30
OLS $\sim N(0, 1)$	0.468	0.486	0.493	0.494	0.489	0.499
CRVE $\sim N(0, 1)$	0.211	0.133	0.108	0.094	0.084	0.080
CRVE $\sim T(G - 1)$	0.100	0.090	0.081	0.075	0.070	0.069

Notes: Replication of simulations performed by CGM. Rejection frequencies estimated with 50,000 replications.

The “Rule of 42”

Claim: “In a DD scenario where you’d like to cluster on state or some other cross-sectional dimension, the relevant dimension for counting clusters is the number of states or cross-sectional groups. Therefore, following Douglas Adam’s dictum that the answer to life, the universe, and everything is 42, we believe the question is: How many clusters are enough for reliable inference using the standard cluster adjustment?”

Angrist and Pischke, *Mostly Harmless Econometrics*, page 319.

Response:

True if clusters are of equal size, false otherwise.

The “Rule of 42”

	6	10	20	50
OLS	0.383	0.443	0.390	0.490
CRVE	0.153	0.105	0.080	0.055

Notes: Bertrand, Duflo and Mullainathan (2004) Monte Carlo Simulations using CPS aggregate data.

Simulations such as these, and those by Cameron, Gelbach and Miller (2008) have led to a shorthand ‘rule of 42’, when $A1$ is approximately satisfied. “Current consensus appears to be that $G = 50$ is enough for state-year panel data.” Cameron and Miller (2013)

Procedure for Wild Cluster Bootstrap-t

1. estimate equation 1 and obtain estimates of $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\epsilon}_{igt}$
- 1a. estimate a restricted version of equation 1 which imposes the null hypothesis, obtain $\tilde{\epsilon}_{igt}$ and equivalent
2. we are interested in the significance of $\hat{\beta}_3$ so calculate the t-statistic, \hat{t} , using cluster robust standard errors
3. choose a number of bootstraps, B , and for each iteration generate a new bootstrap sample from the bootstrap DGP:

$$y_{igt}^* = \tilde{\beta}_0 + \tilde{\beta}_1 * treat_g + \tilde{\beta}_2 * year_t + \tilde{\beta}_3 * treat_g * year_t + X_{igt} \tilde{\gamma} + f(\tilde{u}_{igt}) v_g^*, \quad (2)$$

where $f(\tilde{u}_{igt})$ transforms the i^{th} residual in time t from group g , \tilde{u}_{igt} , and v_g is a bootstrap weight. Impose the null by setting $\tilde{\beta}_3 = 0$

Procedure for Wild Cluster Bootstrap-t

4. estimate equation 1 again using the bootstrap sample
5. calculate the t-statistic, t_j^* on $\hat{\beta}_3$ by using the cluster robust standard errors
6. after the B^{th} iteration calculate the bootstrap p-value by

$$\hat{p}^*(\hat{t}) = 2\min \left(\frac{1}{B} \sum_{j=1}^B I(t_j^* \leq \hat{t}), \frac{1}{B} \sum_{j=1}^B I(t_j^* > \hat{t}) \right). \quad (3)$$

Bootstrap Weight Distribution

- Consider the $f(\tilde{u}_{igt})v_g^*$ term in equation 2
- With the bootstrap techniques considered here $f(\tilde{u}_{igt}) = \tilde{u}_{igt}$
- However, v_g changes according to the bootstrap weight distribution
- One common distribution is the Mammen distribution

$$v_g = -\frac{\sqrt{5}-1}{2} \text{ w.p. } p = \frac{\sqrt{5}+1}{2\sqrt{5}}$$
$$\text{and } v_g = \frac{\sqrt{5}+1}{2} \text{ w.p. } 1-p$$

- The other common distribution, with preferable characteristics, is the Rademacher distribution

$$v_g = \pm 1 \text{ w.p. } 0.5$$

- **However**, both of these result in only 2^G possible bootstrap samples

Monte Carlo Simulation Design

The model is:

$$y_{ig} = \beta_1 + \beta_2 X_{ig} + \epsilon_{ig}, \quad i = 1, \dots, N_g, \quad g = 1, \dots, G. \quad (4)$$

Each simulation proceeds as follows:

- 1 Specify $\rho_x \in \{0, 0.2, \dots, 0.8, 1\}$ and $\rho_\epsilon \in \{0, 0.1, \dots, 0.8, 0.9\}$.
- 2 For each simulated sample, generate X_{ig} and ϵ_{ig} and use equation (4) to compute y_{ig} , with $\beta_1 = 0$ and $\beta_2 = 0$.
- 3 Estimate equation (4) by OLS.
- 4 Test the hypothesis that $\beta_2 = 0$, using either a t test based on the CRVE or a wild bootstrap test, as discussed above.
- 5 Repeat steps 2, 3, and 4 100,000 times, and estimate the rejection frequencies of each test at the .01, .05, and .10 levels.

Rejection Frequencies with 50 Equal-Sized Clusters

ρ_ϵ		ρ_x					
		0	0.2	0.4	0.6	0.8	1
0	t(G-1)	0.0512	0.0502	0.0512	0.0509	0.0572	0.0663
	wild	0.0510	0.0495	0.0505	0.0483	0.0505	0.0503
0.3	t(G-1)	0.0501	0.0518	0.0536	0.0568	0.0616	0.0667
	wild	0.0496	0.0508	0.0504	0.0504	0.0505	0.0503
0.5	t(G-1)	0.0506	0.0502	0.0543	0.0581	0.0634	0.0662
	wild	0.0497	0.0495	0.0506	0.0500	0.0501	0.0501
0.7	t(G-1)	0.0507	0.0521	0.0543	0.0590	0.0637	0.0676
	wild	0.0498	0.0502	0.0500	0.0500	0.0507	0.0515
0.9	t(G-1)	0.0503	0.0517	0.0545	0.0578	0.0641	0.0657
	wild	0.0498	0.0509	0.0498	0.0494	0.0509	0.0495

Notes: Rejection frequencies at the 5% level are based on 100,000 replications. There are 50 equal-sized clusters with 2000 observations. Wild bootstrap P values are based on 399 bootstraps using the Rademacher distribution.

Rejection Frequencies with 50 State-Sized Clusters

		ρ_x					
		0	0.2	0.4	0.6	0.8	1
0	t(G-1)	0.0583	0.0596	0.0600	0.0612	0.0684	0.0818
	wild	0.0489	0.0503	0.0506	0.0498	0.0515	0.0518
0.3	t(G-1)	0.0581	0.0639	0.0706	0.0815	0.0970	0.1051
	wild	0.0498	0.0512	0.0513	0.0519	0.0533	0.0518
0.5	t(G-1)	0.0586	0.0652	0.0746	0.0865	0.0975	0.1064
	wild	0.0506	0.0503	0.0516	0.0538	0.0518	0.0509
0.7	t(G-1)	0.0575	0.0666	0.0771	0.0871	0.0995	0.1086
	wild	0.0494	0.0502	0.0530	0.0522	0.0520	0.0520
0.9	t(G-1)	0.0570	0.0674	0.0769	0.0868	0.0983	0.1077
	wild	0.0519	0.0520	0.0527	0.0519	0.0515	0.0521

Notes: Rejection frequencies at the 5% level are based on 100,000 replications.

There are 50 clusters proportional to US state populations with 2000 observations. Wild bootstrap P values are based on 399 bootstraps using the Rademacher distribution.

Set up for Percentage Treated Monte Carlo

- Many applications to clustered data involve treatment effects at the cluster level.
- We conduct another set of experiments in which the test regressor is a dummy variable that equals one for some proportion P of the clusters.
- The limitations of the CRVE when P is low were presented in Conley and Taber (2011)
- We report results for 50 clusters with 1000 observations, $\rho_\epsilon = 0.50$, and P that varies between 0.02 and 0.98 at intervals of 0.02.
- In “cluster indicator” experiments all observations in a cluster are “treated”.
- In “DiD” experiments one half of observations in a cluster are “treated”.
- The CRVE rejection frequencies are presented in figures 1, 3.
- The Wild bootstrap rejection frequencies are presented in figures 2, 4.

Figure : 1 - CRVE rejection frequencies and proportion treated (cluster indicators)

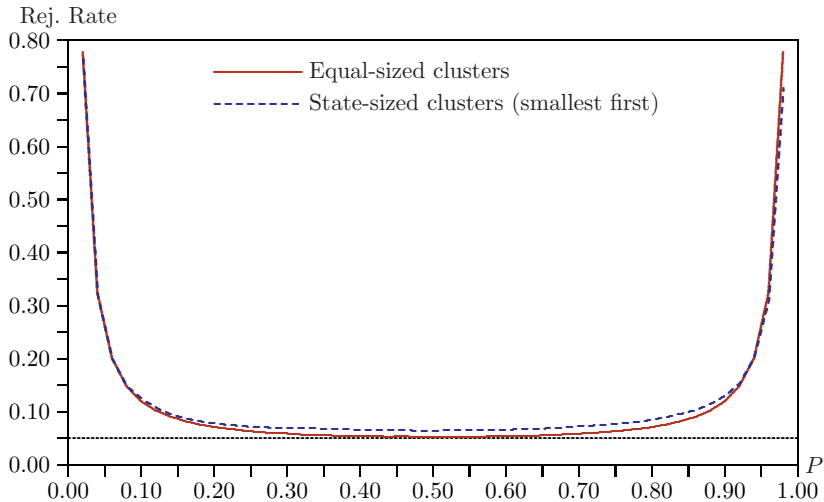


Figure : 2 - Wild bootstrap rejection frequencies and proportion treated (cluster indicators)

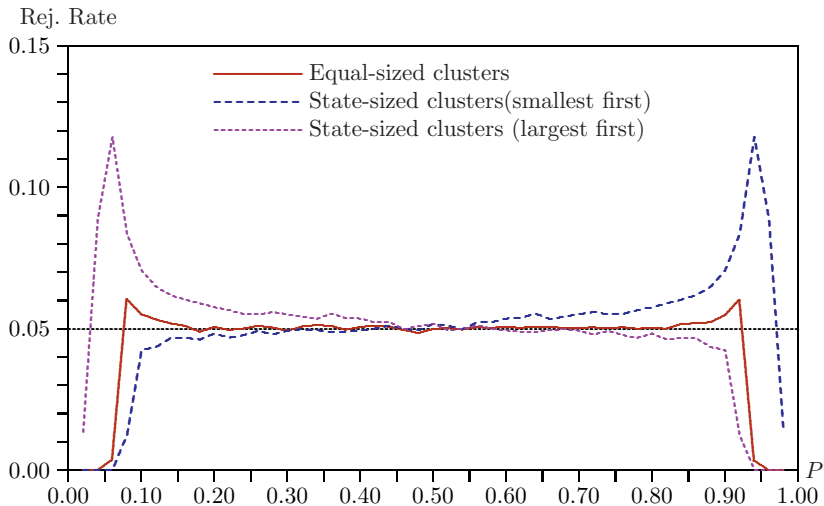


Figure : 3 - Rejection frequencies and proportion treated - DiD

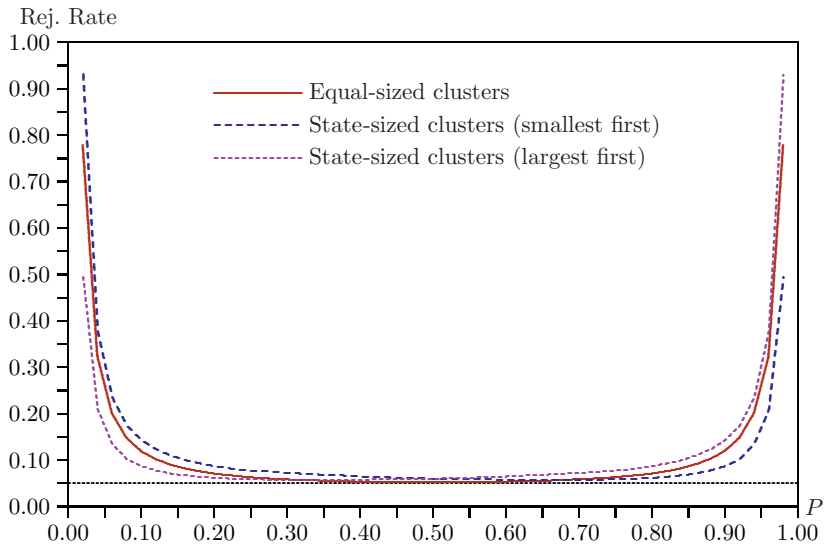


Figure : 4 - Wild bootstrap rejection frequencies DID

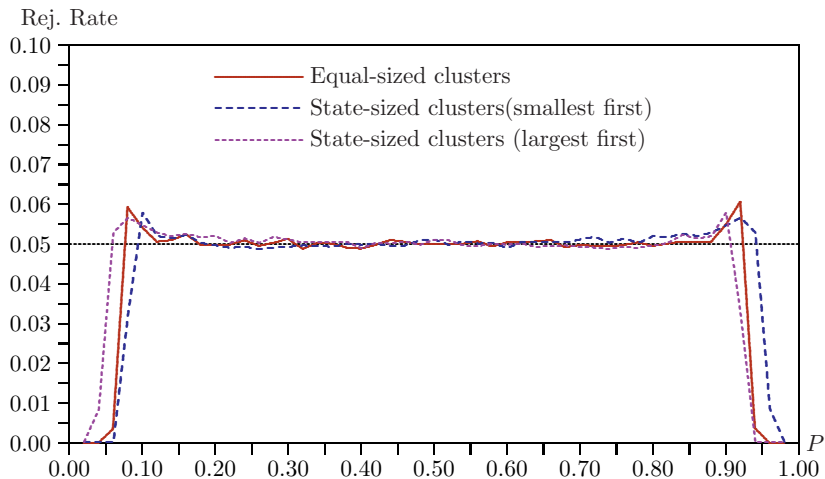
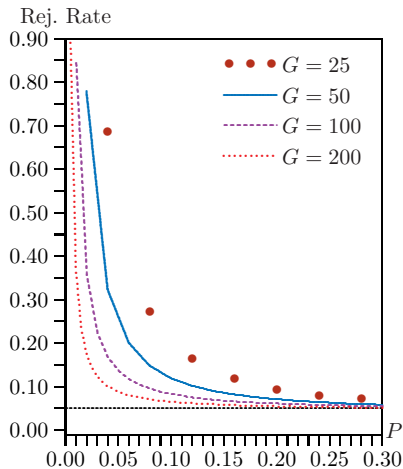
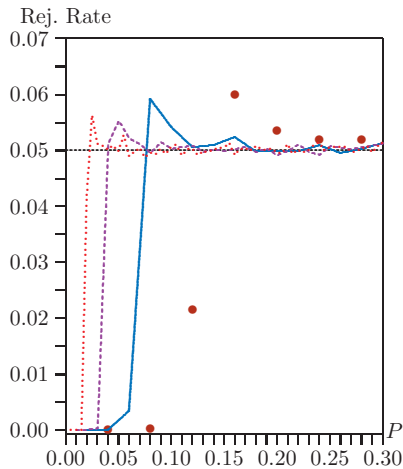


Figure : 5 - Rejection frequencies and proportion treated - equal sized clusters DiD



A. CRVE rejection frequencies



B. Wild bootstrap rejection frequencies

- Replication of Bertrand, Duflo and Mullainathan (2004)
- Estimate DiD coefficients on fake laws for women's wages
- Data from US Current Population Survey, for women aged 25-50 from 1979 to 1999
- For each replication generate a fake treatment which starts in a random year between 1985 - 1995

The regression for the log of women's wages is

$$\ln(\text{wage}) = \beta_1 + \beta_{\text{treat}} \text{TREAT} + \text{YEARS } \beta_{\text{years}} + \text{STATES } \beta_{\text{states}} + \text{controls} + \epsilon, \quad (5)$$

where YEARS and STATES are full sets of fixed effects, and the controls are a quadratic in age and a set of education dummy variables.

Rejection Frequencies of Placebo Law Monte Carlo Simulations Using Current Population Survey Data

		HCCME	t(G-1)	Wild
level 0.10	Random 25	0.706	0.182	0.143
level 0.10	Random 10	0.754	0.222	0.106
level 0.10	Random 1	0.712	0.804	0.000
		HCCME	t(G-1)	Wild
level 0.05	Random 25	0.652	0.118	0.059
level 0.05	Random 10	0.713	0.134	0.049
level 0.05	Random 1	0.640	0.762	0.000
		HCCME	t(G-1)	Wild
level 0.01	Random 25	0.560	0.023	0.011
level 0.01	Random 10	0.618	0.052	0.012
level 0.01	Random 1	0.498	0.709	0.000

Notes: Rejection frequencies based on 1000 replications.

Conclusions

- Even with many clusters, CRVE inference can be unreliable, especially when:
 - Clusters are of wildly different sizes
 - The proportion of clusters treated is either very large or very small
- The wild cluster bootstrap allows for reliable inference with variable cluster sizes
- The wild cluster bootstrap will underreject when the proportion treated is very large or very small

Bibliography I

- Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004) 'How much should we trust differences-in-differences estimates?' *The Quarterly Journal of Economics* 119(1), pp. 249–275
- Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2008) 'Bootstrap-based improvements for inference with clustered errors.' *The Review of Economics and Statistics* 90(3), 414–427
- Cameron, A.C., and D.L. Miller (2013) 'A practitioner's guide to cluster robust inference.' *Journal of Human Resources* p. forthcoming
- Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald (2012) 'Cluster robust inference for heterogeneous cluster samples.' Technical Report, University of California, Santa Barbara
- Conley, Timothy G., and Christopher R. Taber (2011) 'Inference with "Difference in Differences"; with a small number of policy changes.' *The Review of Economics and Statistics* 93(1), 113–125

Bibliography II

- Donald, Stephen G, and Kevin Lang (2007) 'Inference with difference-in-differences and other panel data.' *The Review of Economics and Statistics* 89(2), 221–233
- Kloek, T. (1981) 'OLS estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated.' *Econometrica* 49(1), pp. 205–207
- Moulton, Brent R. (1990) 'An illustration of a pitfall in estimating the effects of aggregate variables on micro units.' *Review of Economics & Statistics* 72(2), 334
- Webb, Matthew D. (2013) 'Reworking wild bootstrap based inference for clustered errors.' Working Papers 1315, Queen's University, Department of Economics, August
- White, Halbert (1984) *Asymptotic theory for econometricians* (Orlando: Academic Press)