Inference with (very) few Clusters in Applied Research

Camila Steffens
Maria Alexandra Castellanos

Reading Group in Applied Economics

September, 2022

uc3m | Universidad Carlos III de Madrid

Introduction

- How to perform inference when errors are **independent between clusters**, but **correlated within them**
  - e.g.: Difference-in-Differences with county-level observations, but treatment at state level
- **Outline:**
  1. Why clustering?
  2. What to cluster over?
  3. Dealing with few clusters
  4. The case of very few clusters
- Main references: Cameron and Miller (2015) and Roth et al. (2022)

Why clustering? (Cameron and Miller, 2015)

- Canonical inference assumes that errors are **independent across individuals** and homoskedastic
- (Heteroskedastic-) robust standard errors still requires an assumption of random sampling

reg y x, rob

- **In many applications, errors may not be independent (e.g., when the regressor is correlated within a cluster)**

Why clustering? A bit of theory (Cameron and Miller, 2015)

- Simple regression model $y_i = \beta x_i + u_i$, $i = 1, ..., N$, $\mathbb{E}[u_i] = 0$

$$V[\hat{\beta}] = \frac{V\left[\sum_i x_i u_i\right]}{\left(\sum_i x_i^2\right)^2}$$

- Errors are **uncorrelated across** i:

$$V\left[\sum_i x_i u_i\right] = \sum_i x_i^2 V[u_i] = \sum_i x_i^2 E[u_i^2]$$

- **Heteroskedastic-robust variance estimator:**

$$\hat{V}[\hat{\beta}] = \frac{\left(\sum_i x_i^2 \hat{u}_i{}^2\right)}{\left(\sum_i x_i^2\right)^2}$$

Why clustering? A bit of theory (Cameron and Miller, 2015)

- Errors are **correlated across** i:

$$V\left[\sum_i x_i u_i\right] = \sum_i \sum_j Cov\left[x_i u_i, x_j u_j\right] = \sum_i \sum_j x_i x_j E[u_i u_j]$$

- **Cluster-robust variance estimator:**

$$\hat{V}[\hat{\beta}]_{clu} = \frac{\left(\sum_i \sum_j x_i x_j \hat{u}_i \hat{u}_j \cdot \mathbb{I}\left[i, j \text{ are in same cluster}\right]\right)}{\left(\sum_i x_i^2\right)^2}$$

- It assumes that $\mathbb{E}[u_i u_j] = 0$ if i, j are in different clusters (i.e., errors are **independent across clusters**)
- Since we are adding the covariance of $i \neq j$ in same cluster, usually $\hat{V}[\hat{\beta}]_{clu} > \hat{V}[\hat{\beta}]$

Why clustering? (Cameron and Miller, 2015)

⚠ In general, not clustering **underestimates** the standard errors (over-rejection)

- Stronger underestimation as:
  - more correlated are the errors
  - larger is N_g [1]
  - more **positively associated are the regressors** across observations in same cluster
    - ▶ Treatment is fully correlated across counties treated by a policy implemented at state level

[1] Notation: G is the number of clusters, and N_g is the number of observation within cluster $g = 1, ..., G$.

When the number of clusters is large ($G \to \infty$), we can implement **Cluster-Robust Standard Errors** in Stata:

*reg y x, vce(cluster **G**)*

- ✓ It can be used in balanced and unbalanced panels, with fixed or large N_g
- ✓ **In general:** cluster at broader levels to accommodate more correlation
- ⚠ **Trade-off:** larger clusters $\Downarrow$ bias in the estimation of standard errors, but might be too few clusters

Outline

1. Why clustering?
2. What to cluster over?
   - Examples (Cameron and Miller, 2015)
   - Insights from design-based inference (Roth et al., 2022; Rambachan and Roth, 2022)
   - Multiway Cluster (Cameron, Gelbach, and Miller, 2011)
   - Inference for Matching (Abadie and Spiess, 2022)
3. Dealing with few clusters
4. The case of **very few clusters**

Starting with an example:

- Panel data of individuals i within counties c within states s over time t
- Two-Way Fixed Effects:

$$y_{i,c,s,t} = \alpha_i + \alpha_t + \beta D_{i,c,s,t} + \gamma X_{i,c,s,t} + \epsilon_{i,c,s,t}$$

- Assuming that some individuals are treated at $t = t_i$, the treatment variable is:

$$D_{i,c,s,t} = \begin{cases} 0 & i \text{ not-treated; or } i \text{ is treated and } t \leq t_i \\ 1 & i \text{ is treated and } t > t_i \end{cases}$$

⚠ For each individual, the **error term may be correlated over time** (Bertrand, Duflo, and Mullainathan, 2004)

What to cluster over? (Cameron and Miller, 2015)

1) If **regressors are independent across** i, we should cluster at individual level to accommodate serial correlation :

```
xtset i t

xtreg y D x time_dummies, fe vce(rob)
```

```
reghdfe y D x , absorb(i t) vce(cluster i)
```

- "fixed effect generally does not control for all the within-cluster correlation of the error"

  ⚠ ***reghdfe:*** vce(rob) is not consistent under small T in regression with panel data

  ✓ ***xtreg*** adjusts for serial correlation with *fe vce(rob)*

What to cluster over? (Cameron and Miller, 2015)

2) Treatment is **as good as randomly assigned within county**:

$$y_{i,c,s,t} = \alpha_i + \alpha_t + \beta D_{c,s,t} + \gamma X_{i,c,s,t} + \epsilon_{i,c,s,t}$$

```
xtset i t

xtreg y D x time_dummies, fe vce(cluster county)
```

```
reghdfe y D x , absorb(i t) vce(cluster county)
```

✓ Accounts for correlations across i and over time **within counties**

3) Treatment is **correlated at state level** (e.g., $D_{i,c,s,t} = D_{s,t}$):

$$y_{i,c,s,t} = \alpha_i + \alpha_t + \beta D_{s,t} + \gamma X_{i,c,s,t} + \epsilon_{i,c,s,t}$$

```
xtset i t

xtreg y D x time_dummies, fe vce(cluster state)
```

```
reghdfe y D x , absorb(i t) vce(cluster state)
```

✓ Accounts for correlations across i and over time **within states**

What to cluster over?
Insights from design-based inference (Rambachan and Roth, 2022)

- Canonical inference based on **sampling uncertainty** (observations are sampled from a large/infinite population)
- What if we observe the **full population**? (e.g., data aggregated at state/county-level, administrative data)
- **Design-based inference:**
  - ○ view the sample as the fixed population of interest
  - ○ uncertainty arises from random allocation of treatment
  - ✓ **Rule of thumb: cluster at the "level at which the treatment is independently assigned"** (Roth et al., 2022)
- Rambachan and Roth (2022) show that standard inference methods (e.g., robust or cluster-robust SE) are valid from the design-based perspective, but **potentially conservative**

What to cluster over?
Multiway clustering (Cameron, Gelbach, and Miller, 2011)

- Previous examples were based on nested clusters (e.g., individuals within counties within states)
- **In some cases, we have non-nested clusters:**
  - workers' *occupation* and firms' *sector* when regressors are at those levels
  - both *employee* and *employer* levels in matched employer-employee panel data
  - rotating panel survey (e.g., CPS): correlations within a *survey-year* and serial correlation from observing a *household* over multiple years
  - correlations at *cross-sectional* and *temporal* levels in panel data

What to cluster over?
Multiway clustering (Cameron, Gelbach, and Miller, 2011)

Panel data of individuals i across G states and over T years:

- **State level:** correlations across i and over time **within states**
- **Temporal level:** geographical correlations **within years**
- **Multiway Cluster-Robust SE:** requires $G \to \infty$ and $T \to \infty$

```
reghdfe  y D x , absorb(i t) vce(cluster state year)
```

⚠ **Avoid clustering state-per-year** (e.g., *state#year*): it is unlikely that errors in a state at time t are uncorrelated with errors in that state at time $t+1$ (Cameron and Miller, 2015)

✓ Time FE is enough when shocks are constant "across all observations in a given year" (Cameron and Miller, 2015)

What to cluster over? Inference for Matching (Abadie and Spiess, 2022)

- Matching creates a "dependence between the outcomes of treated units and their matches" in two step-estimation

⚠ Ignoring the matching step can *overestimate* or *underestimate* the standard errors from the regression step

✓ **In matching without replacement, the solution is clustering standard errors at matched-sets**

  - with N_1 treated units, each with 2 matches, the sample size is $N = N_1 + 2N_1$
  - there are N_1 clusters, where each cluster is composed by the treated unit and its two respective matched-controls

- Not immediately extended to **matching with replacement**, since untreated units may be part of multiple clusters

Outline

1. Why clustering?
2. What to cluster over?
3. Dealing with few clusters (Cameron and Miller, 2015; Roth et al., 2022)
   - Wild-Cluster Bootstrap (MacKinnon and Webb, 2017)
   - Randomization Inference (MacKinnon and Webb, 2020)
4. The case of **very few** clusters

Dealing with few clusters (Cameron and Miller, 2015; Roth et al., 2022)

⚠ Cluster-Robust SEs require large number of clusters ($G \to \infty$)

- **How large?** (MacKinnon and Webb, 2017)

  ◦ No consensus: the number of clusters increases with clusters' heterogeneity (e.g., 50 might not be enough when cluster sizes are unequal)

  ◦ **(Binary) treatment effects:** also need to consider the proportion of treated clusters compared to the control
  See Figure 1

Rejection Rates with Cluster-Robust SE (MacKinnon and Webb, 2017)

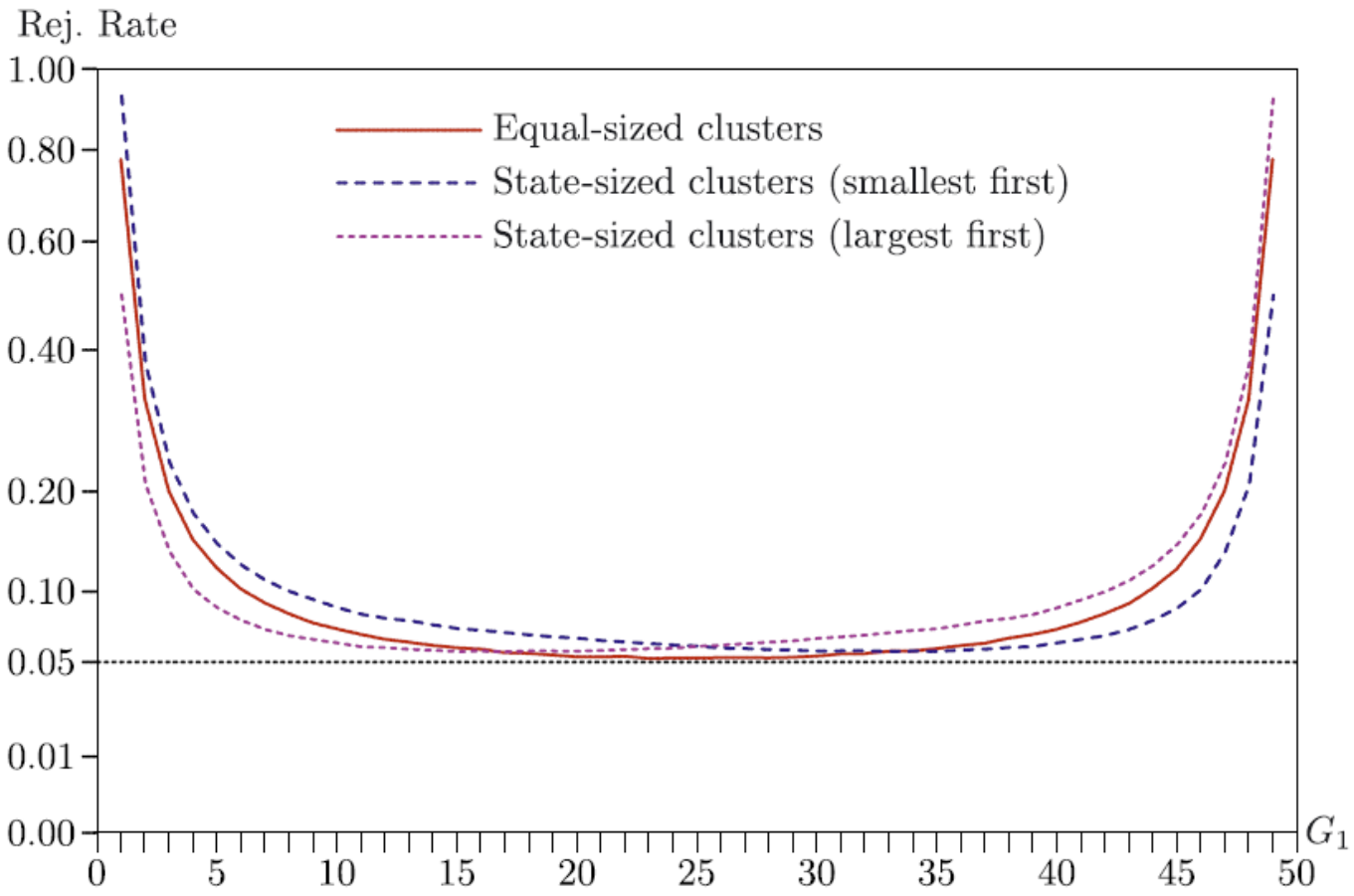

Figure 1: Rejection rates and proportion of treated clusters (G_1) in DiD

The total number of clusters is 50. G_1 is the number of treated cluster, which ranges from 0 (all clusters are untreated) to 50 (all clusters are treated)

Dealing with few clusters (Cameron and Miller, 2015; Roth et al., 2022)

Common solutions:

- ✓ Wild-Cluster Bootstrap
- ✓ Randomization Inference

Wild-Cluster Bootstrap (Cameron and Miller, 2015)

- Differently from a standard bootstrap, instead of resampling clusters, it keeps the regressors fixed and "randomizes" (shocks to) the dependent variable
- First, estimate $\hat{\beta}$, the **cluster-robust standard error** $s_{\hat{\beta}}$ and the *t-statistic* $w = \hat{\beta}/s_{\hat{\beta}}$ using the original sample
- Estimate again the model **imposing the null hypothesis** that you want to test ($\beta = 0$). E.g., for treatment effects of $D_{i,g,t}$, estimate the model only with the covariates and fixed effects:

$$\tilde{y}_{i,g,t} = \tilde{\alpha}_i + \tilde{\alpha}_t + \tilde{\gamma}\mathbf{x}_{i,g,t}$$

and obtain the residual:

$$\tilde{u}_{i,g,t} = y_{i,g,t} - \tilde{y}_{i,g,t}$$

Algorithm of standard Cluster Bootstrap

Algorithm - Do the following B times:

1. Obtain a sample of G clusters $\{(y_1^*, X_1), ..., (y_G^*, X_G)\}$:

1.1) Randomly assign to cluster g a weight $d_g = \{-1 \text{ or } 1\}$ with same probability (note: d_g is equal for all observations in the cluster)

1.2) Generate $u_{i,g,t}^* = d_g \times \tilde{u}_{i,g,t}$ and $y_{i,g,t}^* = \tilde{\alpha}_i + \tilde{\alpha}_t + \tilde{\gamma}\mathbf{x}_{i,g,t} + u_{i,g,t}^*$

2. Estimate the original model with $y_{i,g,t}^*$ as the dependent variable to obtain $\hat{\beta}_b^*$ and the **cluster-robust standard error** $s_{\hat{\beta}_b^*}$

$$\hat{y^*}_{i,g,t} = \hat{\alpha}_i + \hat{\alpha}_t + \hat{\gamma}\mathbf{x}_{i,g,t} + \hat{\beta}_b^* D_{i,g,t}$$

3. Calculate the t-stat $w_b^* = (\hat{\beta}_b^* - \hat{\beta})/s_{\hat{\beta}_b^*}$

- Calculate the **p-value**: $\hat{p} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left(|w_b^*| > |w|\right)$

Single coefficient:

reghdfe y D X time , absorb(i) vce(cluster state)
boottest D, reps(B) boottype(wild) bootcluster(state)

Multiple coefficients:

reghdfe y D_1 D_2 X time , absorb(i) vce(cluster state)

Multiple hypotheses (separate tests):
boottest $\{D_1\}$ $\{D_2\}$, reps(B) boottype(wild) bootcluster(state)

Test for joint null of multiple coefficients:
boottest D_1 D_2 , reps(B) boottype(wild) bootcluster(state)

- ✓ Wild-Cluster Bootstrap can be performed with few clusters (e.g., G = 10)

- ⚠ However, it doesn't perform well with **very few treated clusters (G_1)** compared to control (MacKinnon and Webb, 2017), and in settings with **heterogeneous treatment effect across clusters** (Roth et al., 2022)

Simulations

- ✓ MacKinnon and Webb (2020) show that Randomization Inference performs better in settings with **few treated** and **many untreated** clusters
  - e.g., with $2 \leq G_1 \leq 4$ for a simulation with 40 clusters

Randomization Inference in a nutshell (MacKinnon and Webb, 2020)

- Randomize treatment assignment across clusters and estimate "placebo treatment effect" for R repetitions
- Compute the proportion of times that such "placebo effects" are larger than the "true" estimated effect
- If small (e.g., $p \leq 0.05$), the estimated treatment effect is "unlikely to be observed by chance" (HeSS, 2017)

Implementation in Stata (HeSS, 2017)

$$y_{i,s,t} = \alpha_i + \alpha_t + \beta D_{s,t} + \gamma X_{i,s,t} + \epsilon_{i,s,t}$$

$$D_{s,t} = T_s \times P_t$$

where $P_t = \mathbb{I}\{t = \text{post-treatment}\}$
$T_s = \mathbb{I}\{s \text{ is treated}\}$ is the treatment allocation

ritest T_s _b[T_s#P_t], **cluster(***state***)** reps(R):
reghdfe y T_s#P_t x, absorb(i t) vce(cluster *state*)

- T_s is the variable to be resampled
- _b[T_s#P_t] is the statistic to be compared ($\hat{\beta}$ in this case)
- **cluster(***state***)** keeps the treatment assignment constant within states
- reps(R) defines the number of permutations (R)

Implementation in Stata (HeSS, 2017)

⚠ RI based on distribution of treatment effect ($\hat{\beta}$) doesn't perform well when clusters are heterogeneous (MacKinnon and Webb, 2020)

✓ **Recommendation:** RI on ***t-statistic***:

ritest T_s _b[T_s#P_t]/_se[T_s#P_t], cluster(*state*) reps(R):
reghdfe y T_s#P_t x, absorb(i t) vce(cluster *state*)

Algorithm

Resampling Methods (HeSS, 2017)

⚠ When treatment variable has cross-sectional and temporal dimension, we must be careful with the variable to be randomized

- In general, we are interested in randomizing treatment allocation across observations and not over time
- Resampling on both dimensions ($D_{s,t}$) removes the autocorrelation of the treatment over time

- The resampling of the treatment must be defined at the **same level of treatment assignment** (e.g., not resampling across individuals when treatment is implemented at state level)
- The resample variable is **not restricted to binary treatment**, and *ritest* accommodates alternative resampling specifications from external file or a sampling program

Resampling Methods (HeSS, 2017)

- When the treatment is allocated respecting some **strata**:
  - Example: 2 states are treated in the Midwest, 1 state in the East, and 3 states in the South
  - a variable *region* identifies Midwest, East, South
- ***strata(region)*** fixes the distribution of the treatment across regions: 2 states randomly treated in the Midwest, 1 randomly treated in the East, 3 in the South
- Without specifying **strata**: 6 states randomly treated across all the states of the country

Number of Permutations (MacKinnon and Webb, 2020)

- With G clusters, from which G_1 are treated, the number of possible re-randomizations without replacement is:

$$R = {}_G C_{G_1} - 1 = \frac{G!}{G_1!(G - G_1)!} - 1$$

- If R is very large: set a number # of random permutations in **reps(#)**
- If R is not that large, obtain the "placebo effects" for all the possible permutations
  - we can compute the **precise p-value of the test**
  - need to enumerate permutations using an external file or a program (see example 4.3 in HeSS (2017))
- If $G_1 = 1$, we should compute the "placebo effects" for all the untreated clusters (i.e., $R = G_0 \equiv G - 1$)

Randomization Inference: some caveats (MacKinnon and Webb, 2020)

- RI performs well under the assumption of **random assignment of the treatment**
- When assignment is **conditionally random** and **clusters are heterogeneous**, RI on *t-statistics* is more reliable than RI on estimated coefficients, but at cost of power loss
  - Power loss is potentially larger when the number of clusters decreases, due to inefficiency of Cluster-Robust SE
  - In general, RI on *t-statistics* tends to **under-reject**
- RI tests are valid under the **sharp null** of no treatment effect for all observations, which is stronger than the null hypothesis of no **average treatment effects** (Roth et al., 2022)

Main takeaways so far...

- In general, broad clustering accommodates more correlation
- However, it is often the case in real world data that we end up with **few clusters**
- **Wild-Cluster Bootstrap** works well when the number of treated and untreated clusters is not very different
- **Randomization Inference** performs better with few treated clusters ...
  - but still requires many untreated clusters, random assignment and homogeneity assumptions
- Can we perform inference when the number of treated and untreated clusters is potentially (very) small? Is it possible to accommodate more heterogeneity in such settings?
  - e.g.: PA and NJ comparison by Card and Krueger (1994)

Outline

1. Why clustering?
2. What to cluster over?
3. Dealing with few clusters
4. The case of **very few** clusters
   - Overview of the model-based approach (Roth et al., 2022)
   - Rearrangement with one treated cluster (Hagemann, 2020)
   - Permutation over time (Chernozhukov, Wüthrich, and Zhu, 2021)

Model-based approach (Roth et al., 2022)

- Modelling the dependence within clusters:

$$Y_{igt} = \alpha_g + \gamma_t + D_{gt}\beta + (\nu_{gt} + \epsilon_{igt})$$

  where α_g (or α_i) and γ_t are cluster (or unit) and time fixed effects; D_{gt} indicates whether cluster g is treated in t; ν_{gt} is a common cluster-by-time error; ϵ_{igt} is the unit-level error term.

- Model-based approaches impose restrictions on cluster-specific errors (ν_{gt}), such as:
  - homoskedasticity, mean-zero, normal distribution, and independence across clusters (*iid* normal)
  - homogeneity in both cluster sizes and average treatment effect, and large number of untreated clusters -> untreated error terms could be leveraged for inference

Clustering at the unit-level as an alternative:

- Instead of treating ν_{gt} as random, we could "condition of the values of ν_{gt} and view the remaining uncertainty as coming from the sampling of individuals within clusters"
- Card and Krueger (1994): two states (NJ and PA) considered as fixed, and state-level shocks are violation of parallel trends

✓ **Recommendation:** clustering at individual/unit-level, and provide sensitivity analysis to parallel trends

Outline

1. Why clustering?
2. What to cluster over?
3. Dealing with few clusters
4. The case of **very few** clusters
   - Overview of the model-based approach (Roth et al., 2022)
   - Rearrangement with one treated cluster (Hagemann, 2020)
   - Permutation over time (Chernozhukov, Wüthrich, and Zhu, 2021)

Rearrangement with one treated cluster (Hagemann, 2020)

- Single treated cluster, fixed number of untreated clusters (q), and a large number of observations within clusters
- Treatment effect is identified by between-clusters comparison, and "allows for heterogeneity of unknown form"
- Two-samples problem to test $H_0 : \mu_1 = \mu_0$
  - Treated: $X_1 \sim N(\mu_1, \sigma^2)$
  - Untreated: $X_{0,k} \sim N(\mu_0, \sigma_k^2)$, $k = 1, ..., q$
  - Check how large $|X_1 - \overline{X}_0|$ is compared to any $|X_{0,q} - \overline{X}_0|$, where $\overline{X}_0 = \sum_{k=1}^{q} X_{0,k}/q$
  - If "large enough", there is evidence that $\mu_1 \neq \mu_0$

Rearrangement with one treated cluster (Hagemann, 2020)

- Cluster $g = 1$ is untreated in $t = 0$ and treated in $t = 1$, and $g = 2, ..., (q + 1)$ are untreated in both periods:

$$Y_{i,g,t} = \alpha_i + \gamma_t + D_{g,t}\beta + \delta X_{i,g,t} + \epsilon_{i,g,t}$$

- Taking first differences:

$$\Delta Y_{i,1,t} = \overbrace{\gamma_1 - \gamma_0 + \beta}^{\theta_1} + \delta \Delta X_{i,1,t} + \Delta \epsilon_{i,1,t}$$

$$\Delta Y_{i,g,t} = \overbrace{\gamma_1 - \gamma_0}^{\theta_0} + \delta \Delta X_{i,g,t} + \Delta \epsilon_{i,g,t}, \text{ for } g = 2, ..., q + 1$$

- We want to test the following hypothesis:

$$H_0 : \beta = 0 \Longleftrightarrow \theta_1 = \theta_0$$

Algorithm:

1. Estimate the first differences equations for each cluster:

$$\Delta \hat{Y}_{i,1,t} = \hat{\theta}_1 + \hat{\delta}\Delta X_{i,1,t}$$
$$\Delta \hat{Y}_{i,g,t} = \hat{\theta}_{0,g} + \hat{\delta}\Delta X_{i,g,t}, \text{ for } g = 2, ..., q+1$$

2. Compute $\hat{\theta}_0 = \frac{1}{q}\sum_{g=2}^{q+1}\hat{\theta}_{0,g}$

3. Obtain the following vector with $q+2$ elements:

$$S = \left(\overbrace{(1+\omega)\times(\hat{\theta}_1 - \hat{\theta}_0), (1-\omega)\times(\hat{\theta}_1 - \hat{\theta}_0)}^{\text{2 entries for the treated}}, \overbrace{\hat{\theta}_{0,2} - \hat{\theta}_0, \hat{\theta}_{0,3} - \hat{\theta}_0, ..., \hat{\theta}_{0,q+1} - \hat{\theta}_0}^{\text{one entry for each untreated}} \right)$$

4. Obtain S^{Δ} by sorting the vector S as follows:
   - in a decreasing order if the alternative is $H_1 : \theta_1 > \theta_0$
   - in an increasing order if the alternative is $H_1 : \theta_1 < \theta_0$

Rearrangement with one treated cluster (Hagemann, 2020)

5. Compute the test-statistic:

$$T(S) = \text{mean}\left(S[1, 1:2]\right) - \text{mean}\left(S[1, 3:q+2]\right)$$

$$T(S) = \hat{\theta}_1 - \hat{\theta}_0 - \{\text{average of } \left(\hat{\theta}_{0,g} - \hat{\theta}_0\right) \text{ among } g = 2, ..., q+1\}$$

- We reject H_0 if $T(S) = T(S^{\Delta})$

Rearrangement with one treated cluster (Hagemann, 2020)

- **Intuition:** the test-statistic compares vectors S and S^{Δ}

  If $(1-\omega) \times (\hat{\theta}_1 - \hat{\theta}_0)$ is still larger than all $\left(\hat{\theta}_{0,g} - \hat{\theta}_0\right) \implies$ difference $\hat{\theta}_1 - \hat{\theta}_0$ is "large enough" so that rejects H_0

- The weight $\omega \in (0,1)$ is the key parameter:

$$\uparrow \omega \implies \downarrow (1-\omega) \times (\hat{\theta}_1 - \hat{\theta}_0)$$

Obtaining ω (Hagemann, 2020)

- ω is (numerically) chosen such as the size of the test is α, and depends on the number of untreated clusters q and a measure of heterogeneity ρ
  - smaller significance levels require larger ω
  - ω decreases with q (θ_0 is estimated with more precision)
  - ω increases with ρ
- ω can obtained from the function *stc.weight(q, alpha, rho)* [2]
  - where *alpha* is the level of the test (default is $\alpha = 0.05$)
  - and *rho* is the measure of heterogeneity (default is $\rho = 2$)

Theorem

[2] R-code available in `https://hgmn.github.io/rea/`

Understanding the heterogeneity ρ (Hagemann, 2020)

- ρ "measures how much more variable the estimate from the treated cluster $\hat{\theta}_1$ can be relative to the second-least variable control cluster estimate $\hat{\theta}_{0,k}$"
- With $\rho = 1$, we are assuming that our estimate of θ_1 is at least as precise as almost all the estimates of $\theta_{0,k}$
- $\rho < 1$ still allows for heterogeneity across clusters:
  - $\hat{\theta}_1$ can be "infinitely more variable than the least variable control cluster"
  - but is more precise compared to all other estimates of $\theta_{0,k}$

Some examples of ω (Hagemann, 2020)

		q			
α	ρ	10	20	30	40
0.10	2	.6333	.3294	.2475	.1948
	3		.5543	.4983	.4632
0.05	2		.5020	.4318	.3884
	3		.6703	.6213	.5923
0.01	2		.6986	.6286	.5935
	3		.8058	.7527	.7290

Based on Table 1 (Hagemann, 2020)

Rearrangement with one treated cluster: caveats

- Not clear how to test joint null of multiple coefficients
- Potentially conservative if we observe the "full population"

⚠ "Rules out cluster-specific heterogeneity in trends in untreated potential outcomes" (Roth et al., 2022)

Outline

1. Why clustering?
2. What to cluster over?
3. Dealing with few clusters
4. The case of **very few** clusters
   - Overview of the model-based approach (Roth et al., 2022)
   - Rearrangement with one treated cluster (Hagemann, 2020)
   - Permutation over time (Chernozhukov, Wüthrich, and Zhu, 2021)

Permutation over time (Chernozhukov, Wüthrich, and Zhu, 2021)

- ✓ The method proposed by Chernozhukov, Wüthrich, and Zhu (2021) allows **heterogeneity across clusters** by relying on the **stability of the unobserved shocks over time**
- Inference as a "structural break testing problem" based on the permutation of residuals across time
- **Intuition:** If the error follows the same distribution after the policy, breaks in the outcome are due to policy effects
- Requires a large number of time periods before the treatment

Permutation over time (Chernozhukov, Wüthrich, and Zhu, 2021)

- Cluster $g = 1$ is untreated for T_0 periods, and treated during $T_* = T - T_0$ periods
- Clusters $g = 2, \ldots G$ are untreated for all T
- **Counterfactual model:**

$$Y_{1t}(0) = P_t + u_t$$
$$Y_{1t}(1) = P_t + \theta_t + u_t$$
$$Y_{gt} = Y_{gt}(0) \text{ for } g = 2, \ldots, G$$

Under the sharp null of zero effects:

$$Y_{1t} = Y_{1t}(1) = Y_{1t}(0) = P_t + u_t$$

- **Assumptions:**
  - $E(u_t) = 0$
  - the stochastic shock (error) is stationary and weakly dependent
  - the distribution of the error is invariant to the policy

    $\implies$ P_t is the counterfactual in the absence of policy

Algorithm:

1. Estimate the counterfactual (P_t) for the treated cluster under the null (using data for all t)

$$\textbf{e.g., DiD: } \hat{P}_t = \frac{1}{T}\sum_{s=1}^{T}\left(Y_{1s} - \frac{1}{G-1}\sum_{g=2}^{G} Y_{gs}\right) + \frac{1}{G-1}\sum_{g=2}^{G} Y_{gt}$$

2. Compute the residuals for the treated cluster in each t:

$$\hat{u}_t = Y_{1t} - \hat{P}_t$$

3. Obtain the test-statistic S from the residuals:

$$S(\hat{u}) = \left(\frac{1}{\sqrt{T_*}}\sum_{t=T_0+1}^{T} |\hat{u}_t|\right) \text{ or } S(\hat{u}) = \frac{1}{\sqrt{T_*}}\left|\sum_{t=T_0+1}^{T} \hat{u}_t\right|$$

Algorithm:

4. Obtain the distribution of the S-statistic based on **permutations of the residuals**
   - In each permutation $b = 1, ..., B$, get the vector $\hat{u}_b$ and compute $S(\hat{u}_b)$
   - Block-permutation or random permutation
5. **The p-value is the proportion of times that $S(\hat{u}_b) \geq S(\hat{u})$**

   Formally:

$$\hat{p} = 1 - \hat{F}\left(S(\hat{u})\right), \text{ where}$$
$$\hat{F}\left(S(\hat{u})\right) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\{S(\hat{u}_b) < S(\hat{u})\}$$

Pointwise hypothesis $H_0 : \theta_{t^*} = 0$

For each post-treatment year t^*:

1. Estimate P_t^* using data from $T_0 + 1$ periods ($t = 1, ..., T_0$ and t^*)
2. Obtain a new set of residuals e.g., $\hat{u}_t^* = Y_{1,t} - \hat{P}_t^*$
3. The **pointwise p-value** is the proportion of times that

$$|\hat{u}_t^*| \geq |\hat{u}_{t^*}^*|$$

4. Pointwise confidence intervals can also be computed by inverting the test based on a grid of G candidates values $H0 : \theta_t = \theta_{gt}^0$

Permutation over time: caveats

- T_0 **needs to be large compared to** T_*
  - With $T_0 = 15$, even if $S(\hat{u}_{t^*})$ is the largest statistic-S observed in the data, $p_{t^*} \geq \frac{1}{16} = 0.065$

⚠ Requires parallel trends to hold for many pre-treatment years (Roth et al., 2022)

Many methods have been developed for inference with few clusters:

- **Wild-Cluster Bootstrap** works well when the number of treated and untreated clusters is not very different
- **Randomization Inference** performs better with few treated clusters, *but requires many untreated clusters, random assignment and homogeneity assumptions*
- The **Rearrangement** can be applied for one treated and a fixed number of untreated clusters and allows for heterogeneity "of unknown form", *although it requires homogeneity in trends*
- The **permutation of residuals across the temporal dimension** is more flexible regarding clusters' heterogeneity, *but might be demanding in terms of data requirement*

Which method to apply depends on each context (i.e., reasonable assumptions) and data availability

Thank you!

csteffen@eco.uc3m.es

marcaste@eco.uc3m.es

References I

Abadie, Alberto and Jann Spiess (2022). "Robust post-matching inference". In: *Journal of the American Statistical Association* 117(538), pp. 983–995.

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.

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller (2011). "Robust inference with multiway clustering". In: *Journal of Business & Economic Statistics* 29(2), pp. 238–249.

Cameron, A Colin and Douglas L Miller (2015). "A practitioners guide to cluster-robust inference". In: *Journal of Human Resources* 50(2), pp. 317–372.

References II

Card, David and Alan B Krueger (1994). "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania". In: *The American Economic Review* 84(4), p. 772.

Chernozhukov, Victor, Kaspar Wüthrich, and Yinchu Zhu (2021). "An exact and robust conformal inference method for counterfactual and synthetic controls". In: *Journal of the American Statistical Association* 116(536), pp. 1849–1864.

Hagemann, Andreas (2020). "Inference with a single treated cluster". In: *arXiv preprint arXiv:2010.04076*.

HeSS, Simon (2017). "Randomization inference with Stata: A guide and software". In: *The Stata Journal* 17(3), pp. 630–651.

References III

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.

MacKinnon, James G and Matthew D Webb (2020). "Randomization inference for difference-in-differences with few treated clusters". In: *Journal of Econometrics* 218(2), pp. 435–450.

Rambachan, Ashesh and Jonathan Roth (2022). "Design-Based Uncertainty for Quasi-Experiments". In: *arXiv preprint arXiv:2008.00602v3*.

References IV

Roth, Jonathan, Pedro HC Sant'Anna, Alyssa Bilinski, and John Poe (2022). "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. Section 5: Relaxing sampling assumptions". In: *arXiv preprint arXiv:2201.01194*.

APPENDIX

(Pairs-) Cluster Bootstrap (Cameron and Miller, 2015)

Algorithm:

- Obtain $\hat{\beta}$ and the **cluster-robust standard error** $s_{\hat{\beta}}$ from the original sample
- Do the following B times:
  1. Obtain a sample of G clusters $\{(y_1^*, X_1^*), ..., (y_G^*, X_G^*)\}$ by re-sampling with replacement [3]
  2. Using this sample, compute $\hat{\beta}_b^*$ and the **cluster-robust standard error** $s_{\hat{\beta}_b^*}$
  3. Calculate the t-stat $w_b^* = (\hat{\beta}_b^* - \hat{\beta})/s_{\hat{\beta}_b^*}$
- Obtain the ***p-value*** from the proportion of times that $|w_b^*| > |w|, b = 1, ..., B$, where $w = (\hat{\beta} - \beta)/s_{\hat{\beta}}$ Wild-Cluster Bootstrap

[3] Resampling clusters, keeping all observations fixed within clusters

Caveats of the (Pairs-) Cluster Bootstrap (Cameron and Miller, 2015)

⚠ With few clusters, pairs-cluster bootstrap doesn't eliminate over-rejection issues: we might end up with few (or none) treated or control clusters in some samples

- Abadie and Spiess (2022) show that cluster bootstrap that re-samples on matched sets (i.e., treated units and their untreated matches are drawn together) is valid for post-matching inference when matching without replacement

Wild-Cluster Bootstrap

Rejection Rates: Wild-Cluster Bootstrap (MacKinnon and Webb, 2017)

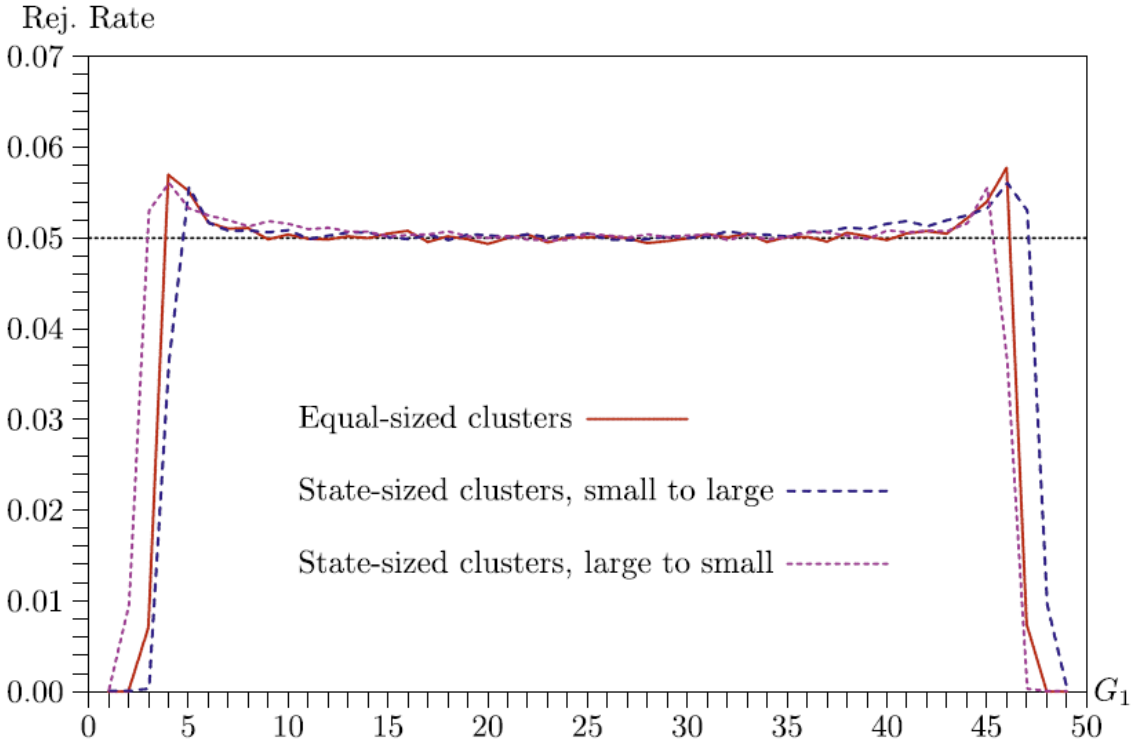

Figure 2: Rejection rates and proportion of treated clusters (G_1) in DiD

The total number of clusters is 50. G_1 is the number of treated cluster, which ranges from 0 (all clusters are untreated) to 50 (all clusters are treated). In general, works well for G_1 between 7 and 43. Otherwise, tends to under-reject. Back

Randomization Inference (MacKinnon and Webb, 2020)

Algorithm:

1. Estimate the model to obtain $\hat{\beta}$ and the **cluster-robust standard error** $s_{\hat{\beta}}$. Calculate the *t-statistic* for $\beta = 0$: $t^* = \hat{\beta}/s_{\hat{\beta}}$

$$\hat{y}_{i,g,t} = \hat{\alpha}_i + \hat{\alpha}_t + \hat{\gamma}\mathbf{x}_{i,g,t} + \hat{\beta}D_{i,g,t}$$

2. For each $r = 1, ..., R$ permutation of the treatment $(D^r_{i,g,t})$, estimate the "placebo treatment" effect and compute

$$t^r = \hat{\beta}^r / s_{\hat{\beta}^r}$$

where $s_{\hat{\beta}^r}$ is the **cluster-robust standard error** of $\hat{\beta}^r$

3. Calculate the ***p-value***: $\hat{p} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}\left(|t^r| > |t^*|\right)$ (two-sided)

Back

Rearrangement with one treated cluster (Hagemann, 2020)

Choice of ω: Theorem 2.1

$$\sup \mathbb{E}[\text{Test}|H_0] \leq \xi_q(\omega, \varrho) \equiv$$
$$\frac{1}{2^{q+1}} + \int_0^\infty \Phi\left((1-\omega)\varrho y\right)^{q-1} \phi(y) dy$$
$$+ \min_{t>0} \left(\Phi\left(\sqrt{q-1}\omega t\right)^{q-1} + 2\Phi(-qt) \right)$$

For each significance level α, number of control clusters q and heterogeneity ϱ, we need to **choose the minimum** ω such that

$$\xi_q(\omega, \varrho) = \alpha$$

- ϱ measures how much more variable is X_1 compared to $X_{0,k}$
- The bound $\xi_q(\omega, \varrho)$ increases with the variation of X_1 (increasing in ϱ) and decreasing with the number of control clusters

Back