Chapter 5: Clustering

Peter Hull

Applied Econometrics II
Brown University
Spring 2024

Outline

1. Cluster-Robust SEs

2. When to Cluster?

General OLS Asymptotics

Recall: the OLS estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ can be rearranged as:

$$\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{\mathbf{X}'\mathbf{X}}{N}\right)^{-1} \cdot \left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{\sqrt{N}}\right)$$

where $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$ stacks observations of the population regression

- Under rather mild conditions (a LLN), $\frac{\mathbf{X}'\mathbf{X}}{N} \xrightarrow{p} E\left[\frac{1}{N}\sum_i X_i X_i'\right]$
- W/slightly stronger conditions (a CLT), $\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{\sqrt{N}} \Rightarrow \mathrm{N}(0, Var(\frac{1}{\sqrt{N}}\sum_i X_i\varepsilon_i))$

This leads to our general asymptotic approximation for OLS: $\hat{\beta} \approx \beta^*$ where

$$\beta^* \sim \mathrm{N}(\beta, V/N), \quad V = E\left[\frac{1}{N}\sum_i X_i X_i'\right]^{-1} Var\left(\frac{1}{\sqrt{N}}\sum_i X_i\varepsilon_i\right) E\left[\frac{1}{N}\sum_i X_i X_i'\right]^{-}$$

Inferences on β from $\hat{V} = \left(\frac{1}{N}\sum_i X_i X_i'\right)^{-1} \widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i X_i\varepsilon_i\right)\left(\frac{1}{N}\sum_i X_i X_i'\right)^{-1}$

Getting to the Meat of the "Sandwich Estimator," $\hat{V}$

Key question: how do we form the variance estimate $\widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i X_i\varepsilon_i\right)$?

In *iid* data, we know $Var\left(\frac{1}{\sqrt{N}}\sum_i X_i\varepsilon_i\right) = \frac{1}{N}\sum_i Var(X_i\varepsilon_i) = E[X_iX_i'\varepsilon_i^2]$

- This suggests $\widehat{Var}\left(\frac{1}{\sqrt{N}}\sum_i X_i\varepsilon_i^2\right) = \frac{1}{N}\sum_i X_iX_i'\hat{\varepsilon}_i$, which leads to our usual heteroskedasticity-robust estimator

The motivation for alternative estimators comes from the possibility that $X_i\varepsilon_i$ and $X_j\varepsilon_j$ may be correlated for $i \neq j$

- Generally, $Var\left(\frac{1}{\sqrt{N}}\sum_i X_i\varepsilon_i\right) = \frac{1}{N}\sum_i Var(X_i\varepsilon_i) + 2\sum_{i,j\neq i} Cov(X_i\varepsilon_i, X_j\varepsilon_j)$
- But we can't allow for arbitrary cross-sectional correlations, since then we couldn't guarantee $\frac{1}{\sqrt{N}}\sum_i X_i\varepsilon_i$ converges ...
- We need to zero out some covariances to make progress

Cluster-Robust Estimators

Suppose we can partition observations into clusters, $c(i) \in 1, \ldots, C$

- To ease notation, suppose equal sizes: $|i : c(i) = c| = N/C \equiv T$
- With $N = CT$, OLS can be rewritten: $\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{\mathbf{X}'\mathbf{X}}{N}\right)^{-1} \cdot \left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{\sqrt{CT}}\right)$

Define $Q_c = \frac{1}{\sqrt{T}} \sum_{i:c(i)=c} X_i \varepsilon_i$ and note that $\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{\sqrt{CT}} = \frac{1}{\sqrt{C}} \sum_c Q_c$

- If the Q_c clusters are *iid*, a CLT applies: $\frac{1}{\sqrt{C}} \sum_c Q_c \Rightarrow \mathrm{N}(0, Var(Q_c))$
- E.g. in a balanced panel, could have *iid* series $(X_{c1}\varepsilon_{c1} \ldots, X_{cT}\varepsilon_{cT})$

This gives us a new "clustered" variance estimate to plug into $\hat{V}$:

$$\widehat{Var}\left(\frac{1}{\sqrt{N}} \sum_i X_i \varepsilon_i\right) = \frac{1}{C} \sum_c \hat{Q}_c^2, \text{ for } \hat{Q} = \frac{1}{\sqrt{T}} \sum_{i:c(i)=c} X_i \hat{\varepsilon}_i$$

This is (basically) what's going on under the hood when you ", *cluster(c)*"

Canonical Example: State-Level Diff-in-Diff

Bertrand et al. (2004) famously show how clustering can matter in DiD

- Good model of an "applied 'metrics" paper: critical but constructive!

The problem: in the late 1990s / early 2000s lots of people were running state-level diff-in-diffs with plain-vanilla ", *r*"

- But of course state-year outcomes Y_{it} are likely serially correlated
- We can model autocorrelation, e.g. $Corr(Y_{it}, Y_{i,t-1}) = \rho$, or we can be completely agnostic and allow for arbitrarily "clustering" over time

BDM illustrate the problem in a very creative / influential way: *placebos*

- Randomly generate fake law changes in state-level wage data and test for effects at the 5% level
- If SEs are well-calibrated, should only reject the null 5% of the time...

Uh-Oh

The Quarterly Journal of Economics, February 2004

HOW MUCH SHOULD WE TRUST DIFFERENCES-IN-DIFFERENCES ESTIMATES?*

MARIANNE BERTRAND
ESTHER DUFLO
SENDHIL MULLAINATHAN

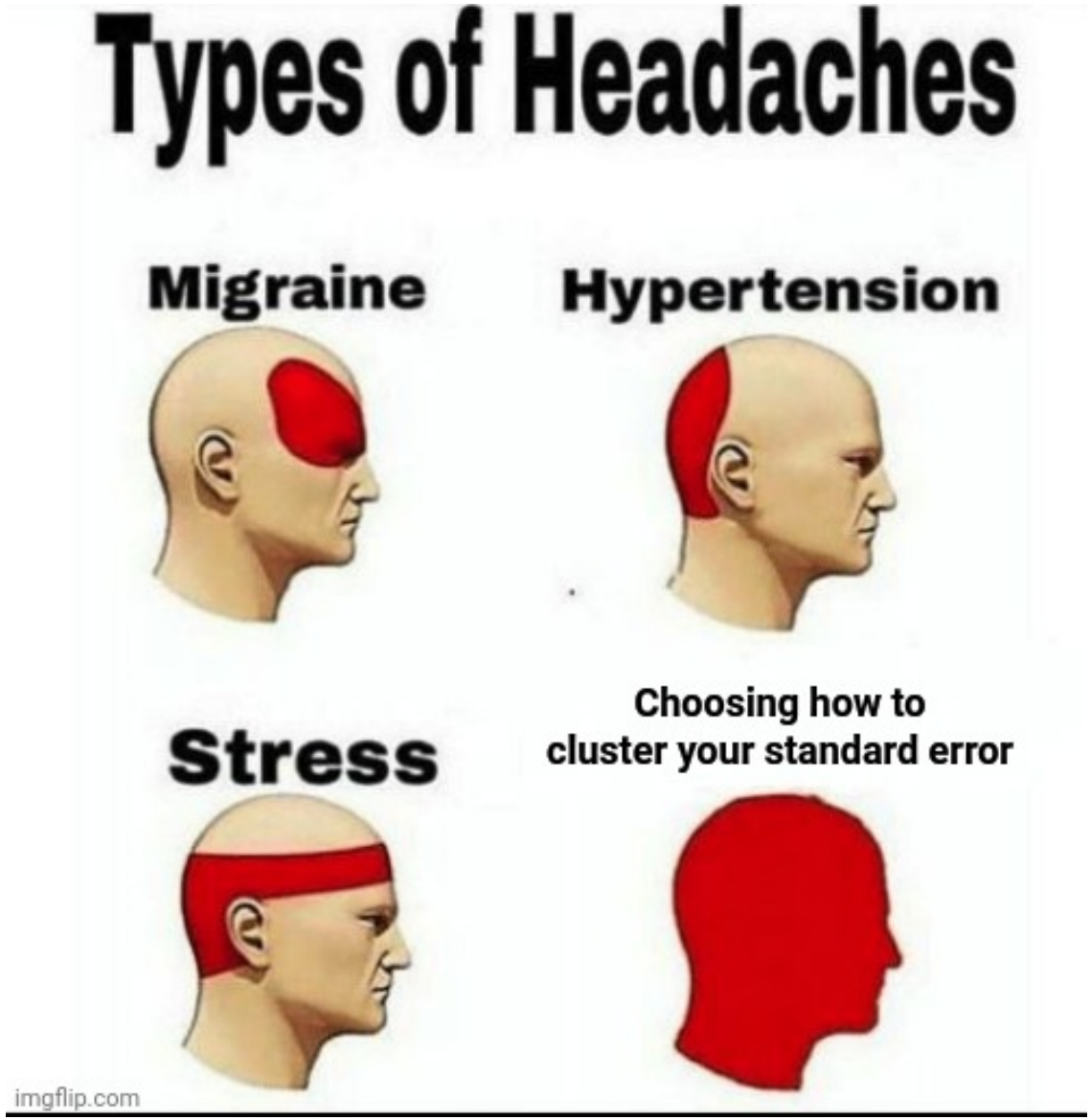
Most papers that employ Differences-in-Differences estimation (DD) use many years of data and focus on serially correlated outcomes but ignore that the resulting standard errors are inconsistent. To illustrate the severity of this issue, we randomly generate placebo laws in state-level data on female wages from the Current Population Survey. For each law, we use OLS to compute the DD estimate of its "effect" as well as the standard error of this estimate. These conventional DD standard errors severely understate the standard deviation of the estimators: we find an "effect" significant at the 5 percent level for up to 45 percent of the placebo interventions. We use Monte Carlo simulations to investi-

Outline

1. Cluster-Robust SEs✓

2. When to Cluster?

, *cluster(head)*



Source: Khoa Vu (of course)

Cluster At The Level of Treatment?

At a (rather unhelpfully) high level, we know when to cluster together i and j: when you expect $Cov(X_i\varepsilon_i, X_j\varepsilon_j) \neq 0$!

In a simple RCT, this is actually not too hard to figure out

- Suppose $(X_1, \ldots, X_N)$ is mean-zero and independent of $(\varepsilon_1, \ldots, \varepsilon_N)$, with $X_i \perp X_j$ whenever $c(i) \neq c(j)$ (e.g. village-level RCT)
- Then whenever $c(i) \neq c(j)$:

$$Cov(X_i\varepsilon_i, X_j\varepsilon_j) = E[X_iX_j'\varepsilon_i\varepsilon_j] = E[E[X_iX_j' \mid \varepsilon_i, \varepsilon_j]\varepsilon_i\varepsilon_j] = 0$$

- So we only need to cluster by $c(i)$: treatment randomization saves us!

This leads to the popular (and sometimes misused) heuristic: cluster at the level of treatment / identifying variation

- See Abadie et al. (2023) for a more complete version of this argument

Where Intuition Can Fall Short: Paired Randomization

Suppose (as is often done) we pair individuals up by some baseline characteristics, then within each pair c we randomly treat one individual

- Treatment is at the individual level... so should we just ", *r*" ?

de Chaisemartin and Ramirez-Cuellar (2022) show the answer is no: non-clustered SEs will generally be downward-biased (maybe very badly)

- Under constant effects, $E[\hat{V}] = V/2$; severe over-rejection!

Paired randomization makes X_i and X_j *negatively* correlated within pairs

- Clustering by pair solves this; treatment assignment is *iid* across pairs
- Alternatively, you could ", *r*" with pair fixed effects (and the standard Stata d.f. correction). Why? Because FE = FD when $T = 2$

Where Clustering is Surely Called For: Stacking

We previously saw how one might "stack" partially-overlapping subsamples to deal with negative weights in TWFE

- We might also stack in order to test across regressions (see `https://twitter.com/instrumenthull/status/1492915860763250691`)

Whenever we repeat observations in a regression, we should always remember to cluster by the original i (or c)

- Consider a silly example: you duplicate your dataset and re-run OLS. What happens to the coefficients? Nothing
- But what happens to the SEs? Fall by a factor of $\sqrt{2}$
- Clustering allows mechanical $Cov(X_i\varepsilon_i, X_j\varepsilon_j) = 1$ among duplicates

Too Few Clusters

Unfortunately, clustering aggressively can come at a cost: with few clusters, the cluster-robust $\hat{V}$ is generally downward-biased

- Version of classic Behrens-Fisher problem (Imbens and Kolesar, 2016)
- Incidentally, also the case for "$, r$" with few observations

There are some parametric solutions that sometimes work well: e.g. Bell and McCaffrey (2002); see again Imbens and Kolesar '16

- See also the wild bootstrap (MacKinnon 2002, 2012)

We usually expect clustering to increase SEs (though not a theorem); if SEs go down a lot with aggressive clustering, you should worry!

Practical Advice

Getting your standard errors "right" is difficult. My advice:

- Start w/a clustering scheme that mimics the unit of *randomization*
- Check other "reasonable" schemes to make sure your results aren't too dependent on one (unless you can strongly justify it)

Sometimes clustered regressions can be aggregated to the level of "randomization:" e.g. reg village-avg outcomes on village-level treatment

- This is a good idea for many reasons: for one thing ", *r*" is enough
- We'll see a slightly more complicated version of this soon

Turns Out, Statistical Inference is Hard!

Journal of Econometrics Session: What is a Standard Error?

Panel Session

Saturday, Jan. 7, 2023 · 2:30 PM - 4:30 PM (CST)

Hilton Riverside, Grand Salon A Sec 3

Hosted By: ECONOMETRIC SOCIETY

Moderators: Serena Ng, Columbia University

Elie Tamer, Harvard University

Panelist(s)

Andrew Gelman, Columbia University

Patrick Kline, University of California-Berkeley

James Powell, University of California-Berkeley

Jeffrey M. Wooldridge, Michigan State University

Bin Yu, University of California-Berkeley