When Should You Adjust Standard Errors for Clustering?

Alberto Abadie
MIT

(joint work with S. Athey, G. Imbens and J. Wooldridge)

Mixtape Sessions
April 27-28, 2023

Prologue: Neyman's surprising result

Suppose you are interested in the average treatment effect in a finite population of n units.

$$\tau = \frac{1}{n}\sum_{i=1}^{n} y_i(1) - \frac{1}{n}\sum_{i=1}^{n} y_i(0).$$

$y_i(1)$ and $y_i(0)$ are the potential outcomes with and without treatment for unit i.

You sample N units (out of n) without replacement, and randomize N_1 units to treatment and $N_0 = N - N_1$ units to control.

You observe $(Y_1, W_1), \ldots, (Y_N, W_N)$, where the Y_i's are outcomes and the W_i's are treatment indicators.

The difference in means between treated and non-treated is

$$\widehat{\tau} = \frac{1}{N_1}\sum_{i=1}^{N} W_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - W_i) Y_i.$$

Prologue: Neyman's surprising result

There is no superpopulation. The stochastic variation in $\widehat{\tau}$ comes only from sampling and from randomization.

The variance of $\widehat{\tau}$ is (Neyman's surprising result)

$$\mathrm{var}(\widehat{\tau}) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0} - \frac{\sigma_\tau^2}{n} \tag{1}$$

where

σ_1^2: variance of $y_1(1), \ldots, y_n(1)$

σ_0^2: variance of $y_1(0), \ldots, y_n(0)$

σ_τ^2: variance of $(y_1(1) - y_1(0)), \ldots, (y_n(1) - y_n(0))$

The usual estimator of the variance of $\mathrm{var}(\widehat{\tau})$ takes care only of the first two terms in (1). Three reasons:

- The third term is small when n is large
- σ_τ^2 is not identified anyway
- $\sigma_\tau^2 = 0$ under constant treatment effects

Prologue: Neyman's surprising result

However, it is easy to come up with examples such that

$$\frac{\sigma_\tau^2}{n} = \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}$$

so $\text{var}(\widehat{\tau}) = 0$. Imagine $n = N = 4$, $N_1 = N_0 = 2$ and

i	$y_i(1)$	$y_i(0)$
1	1	-1
2	2	-2
3	3	-3
4	4	-4

No matter which two units get treatment, $\widehat{\tau} = 5$, so $\text{var}(\widehat{\tau}) = 0$. The usual variance estimator is (very!) conservative.

There is not much we can do in this case, because we cannot estimate σ_τ^2. However, when treatment depends on cluster, the across-clusters variance of the treatment effect may be identified.

College effects in a U.S. Census sample

Dependent variable: Log labor earnings		
Panel A		
Treatment: State indicator for share of some college greater than 0.55		
	OLS	
coefficient	0.1022	
standard error:		
robust	(0.0012)	
cluster	(0.0312)	
Panel B		
Treatment: Individual indicator for some college		
	OLS	FE
coefficient	0.4656	0.4570
standard error:		
robust	(0.0012)	(0.0012)
cluster	(0.0269)	(0.0276)

FAQ's@OH's

At office hours:

- "Should I cluster my standard errors?"
- "If I cluster, at what level should I cluster?"

Basic questions, but surprisingly hard to answer.

Two common frameworks for clustering in econometrics/statistics:

- **Model-based inference:** E.g., DGP is

  $$Y_i = \mu + \eta_{c(i)} + \varepsilon_i.$$

  Error component $\eta_{c(i)}$ is common at the cluster level. Inference for μ takes into account variation induced by cluster-level error components.

- **Sampling-based inference:** Two-stage sampling:
  1. Sample clusters.
  2. Sample observations within the sampled clusters.

Conventional frameworks for clustered inference

Model-based inference:

- Researcher is forced to take a stand on the error components for outcome process in the super-population/DGP.
- Why cluster at the state level, but not at the gender level?

Sampling-based inference:

- Conventional variance estimators assume that we observe a small fraction of the clusters in the population.

None of these approaches incorporates a model for treatment assignment, which will be important to:

- Obtain identification of treatment effects.
- Derive variance formulas that take into account the uncertainty induced by the treatment assignment.

A framework for clustered inference

We propose a framework for clustered inference.

- Shifts the focus of interest from features of infinite super-populations/data-generating processes to average treatment effects defined for the finite population at hand.
- Incorporates a design component that accounts for the variability induced on the estimator by the treatment assignment mechanism.
- Accommodates settings where we observe all or a large fraction of the clusters in the population.

When/How to cluster

In the new framework, the decision on when/how to cluster depends on information potentially available to the researcher.

- The sampling process and the treatment assignment mechanism solely determine the correct level of clustering.
- Clustered sampling requires clustered inference.
- Clustered assignment amounts to clustered sampling of potential outcomes.
- Because the parameter of interest is causal (difference of averages of potential outcomes), clustered assignment also requires clustered inference.
- The presence of cluster-level unobserved components of the outcome variable becomes irrelevant for the choice of clustering level.

Independent sampling

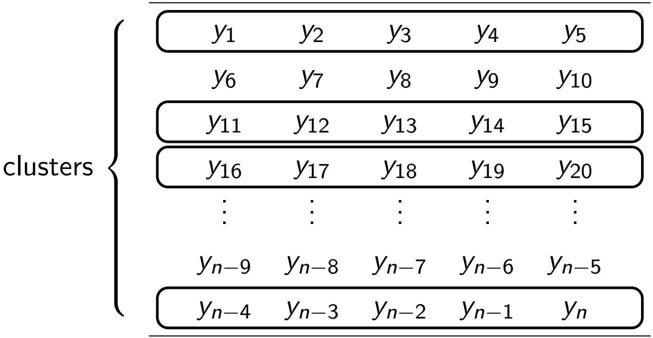
$$\text{clusters}\left\{\begin{array}{ccccc}
\hline
\boxed{y_1} & y_2 & y_3 & \boxed{y_4} & \boxed{y_5} \\
y_6 & \boxed{y_7} & y_8 & \boxed{y_9} & y_{10} \\
\boxed{y_{11}} & \boxed{y_{12}} & \boxed{y_{13}} & y_{14} & y_{15} \\
y_{16} & y_{17} & y_{18} & \boxed{y_{19}} & y_{20} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
y_{n-9} & \boxed{y_{n-8}} & \boxed{y_{n-7}} & y_{n-6} & y_{n-5} \\
y_{n-4} & y_{n-3} & y_{n-2} & y_{n-1} & \boxed{y_n} \\
\hline
\end{array}\right.$$

We aim to estimate the mean

$$\theta = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

with a random sample. No need for cluster estimator of the variance.

Clustered sampling



clusters

y_1	y_2	y_3	y_4	y_5
y_6	y_7	y_8	y_9	y_{10}
y_{11}	y_{12}	y_{13}	y_{14}	y_{15}
y_{16}	y_{17}	y_{18}	y_{19}	y_{20}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{n-9}	y_{n-8}	y_{n-7}	y_{n-6}	y_{n-5}
y_{n-4}	y_{n-3}	y_{n-2}	y_{n-1}	y_n

We aim to estimate the mean

$$\theta = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

with a clustered sample. Need cluster estimator of the variance.

Clustered assignment

$$\text{clusters}\left\{\begin{array}{ccccc}
y_1(0) & y_2(0) & y_3(0) & y_4(0) & y_5(0) \\
y_6(1) & y_7(1) & y_8(1) & y_9(1) & y_{10}(1) \\
y_{11}(1) & y_{12}(1) & y_{13}(1) & y_{14}(1) & y_{15}(1) \\
y_{16}(0) & y_{17}(0) & y_{18}(0) & y_{19}(0) & y_{20}(0) \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
y_{n-9}(0) & y_{n-8}(0) & y_{n-7}(0) & y_{n-6}(0) & y_{n-5}(0) \\
y_{n-4}(1) & y_{n-3}(1) & y_{n-2}(1) & y_{n-1}(1) & y_n(1)
\end{array}\right.$$

We aim to estimate

$$\tau = \frac{1}{n}\sum_{i=1}^{n} y_i(1) - \frac{1}{n}\sum_{i=1}^{n} y_i(0),$$

using a difference in means between treated and nontreated. For each of the two averages in τ we have a clustered sample. Need cluster estimator of the variance.

Applying the framework to derive variance formulas

We demonstrate how this framework produces precise answers on the properties of conventional variance estimators and can be used to derive improved variance estimators.

Aside from a definition of estimand and estimator, applying the framework to derive variance formulas involves careful characterizations of three key elements:

- A sequence of populations
- A sampling mechanism
- An assignment mechanism

We will go through this exercise for a particular (yet widely applicable) setting.

A sequence of populations

Elements of the model:

- We have a sequence of populations indexed by k. The k-th population has n_k units, indexed by $i = 1, \ldots, n_k$.
- The population is partitioned into m_k clusters. $m_{k,i} \in \{1, \ldots, m_k\}$ is the cluster that unit i of population k belongs to.
- The number of units in cluster m of population k is $n_{k,m} \geq 1$.
- For each unit, i, there are bounded potential outcomes, $y_{k,i}(1)$ and $y_{k,i}(0)$, corresponding to treatment and no treatment.

Remarks:

- We condition on the population. Potential outcomes are fixed.
- The population is finite, which allows us to sample a non-zero fraction of it.
- To obtain limit results we embed the population in a sequence, with population size increasing in k.

A sequence of populations

The object of interest is the population average treatment effect

$$\tau_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \big(y_{k,i}(1) - y_{k,i}(0) \big).$$

The population average treatment effect by cluster is

$$\tau_{k,m} = \frac{1}{n_{k,m}} \sum_{i=1}^{n_k} 1\{m_{k,i} = m\}(y_{k,i}(1) - y_{k,i}(0)).$$

Therefore,

$$\tau_k = \sum_{m=1}^{m_k} \frac{n_{k,m}}{n_k} \tau_{k,m}.$$

A sequence of populations

Stochastic objects in the population and the sample:

- For each unit in the population, we define the stochastic treatment indicator, $W_{k,i} \in \{0, 1\}$.
- The realized outcome for unit i in population k is $Y_{k,i} = y_{k,i}(W_{k,i})$.
- For a sample of the population, we observe the triple $(Y_{k,i}, W_{k,i}, m_{k,i})$.
- Inclusion in the sample is represented by the random variable $R_{k,i}$, which takes value one if unit i belongs to the sample, and value zero if not.

We next describe the two components of the stochastic nature of the sample:

- The sampling process that determines the values of $R_{k,i}$.
- The assignment process that determines the values of $W_{k,i}$.

The sampling process

Elements of the sampling process:

- Clusters are sampled with probability $q_k \in (0, 1]$.
- Units in sampled clusters are sampled with probability $p_k \in (0, 1]$.
- If $q_k = 1$, we sample all clusters. If $p_k = 1$, we sample all units from the sampled clusters. If $q_k = p_k = 1$, all units in the population are sampled.

Remarks:

- The conventional framework for analyzing cluster sampling focuses on the special case $q_k \to 0$.
- In many applications sampled clusters comprise a large fraction of the overall set of clusters.

We refer to the case of $q_k = 1$ as *random sampling* and to the case of $q_k < 1$ as *clustered sampling*.

The assignment mechanism

Elements of the assignment mechanism:

- In a first stage, for cluster m in population k, $A_{k,m} \in [0,1]$ is drawn randomly from a distribution with mean μ_k and variance σ_k^2, independently for each cluster.
- In a second stage, each unit in cluster m is assigned to treatment independently, with probability $A_{k,m}$.

Because $A_{k,m}^2 \leq A_{k,m}$, it follows that $0 \leq \sigma_k^2 \leq \mu_k(1-\mu_k)$.

- *Random assignment:* $\sigma_k^2 = 0$. All units have the same probability of treatment.
- *Clustered assignment:* $\sigma_k^2 = \mu_k(1-\mu_k)$, all units within clusters have the same treatment values.
- *Partially clustered assignment:* $0 < \sigma_k^2 < \mu_k(1-\mu_k)$. Assignment probability depends on cluster, but not all units in the same cluster necessarily have the same value of $W_{k,i}$.

Least squares

Let

$$N_{k,1} = \sum_{i=1}^{n_k} R_{k,i} W_{k,i} \quad \text{and} \quad N_{k,0} = \sum_{i=1}^{n_k} R_{k,i}(1 - W_{k,i})$$

and $N_k = N_{k,1} + N_{k,0}$.

We first analyze the OLS estimator of a regression of the outcome $Y_{k,i}$ on an intercept and the treatment indicator $W_{k,i}$.

The OLS estimator (modified so it is well-defined even when $N_{k,1} = 0$ or $N_{k,0} = 0$) is equal to the difference in means:

$$\widehat{\tau}_k = \frac{1}{N_{k,1} \vee 1} \sum_{i=1}^{n_k} R_{k,i} W_{k,i} Y_{k,i} - \frac{1}{N_{k,0} \vee 1} \sum_{i=1}^{n_k} R_{k,i}(1 - W_{k,i}) Y_{k,i}.$$

Conventional variance estimators

Let $\widehat{U}_{k,i}$ be the residuals from the regression of $Y_{k,i}$ or a constant and $W_{k,i}$. Let $\overline{W}_k$ be the sample average of $W_{k,i}$.

The conventional heteroskedasticity-robust estimator of the asymptotic variance ("robust") is

$$\widehat{V}_k^{\text{robust}} = \frac{1}{\overline{W}_k^2(1-\overline{W}_k)^2}\left\{\frac{1}{N_k}\sum_{i=1}^{n_k} R_{k,i}\widehat{U}_{k,i}^2(W_{k,i}-\overline{W}_k)^2\right\}.$$

The conventional clustered estimator of the asymptotic variance ("cluster") is

$$\widehat{V}_k^{\text{cluster}} = \frac{1}{\overline{W}_k^2(1-\overline{W}_k)^2}\left\{\frac{1}{N_k}\sum_{m=1}^{m_k}\left(\sum_{i=1}^{n_k} 1\{m_{k,i}=m\} R_{k,i}\widehat{U}_{k,i}(W_{k,i}-\overline{W}_k)\right)^2\right\}.$$

Variance of least squares

We obtain

$$\sqrt{N_k}(\widehat{\tau}_k - \tau_k)/v_k^{1/2} \xrightarrow{d} N(0,1),$$

where

$$\widehat{V}_k^{\text{cluster}} - v_k \approx p_k q_k \left\{ \sum_{m=1}^{m_k} \frac{n_{k,m}^2}{n_k} (\tau_{k,m} - \tau_k)^2 \right\},$$

and typically

$$\widehat{V}_k^{\text{robust}} \ll \widehat{V}_k^{\text{cluster}}.$$

Comparison of $\widehat{V}_k^{\text{cluster}}$ and v_k

$$\widehat{V}_k^{\text{cluster}} - v_k \approx p_k q_k \left\{ \sum_{m=1}^{m_k} \frac{n_{k,m}^2}{n_k} (\tau_{k,m} - \tau_k)^2 \right\}.$$

When q_k is small, or when the average treatment effect is nearly constant between clusters, then $v_k^{\text{cluster}} \approx v_k$.

Aside from these special cases, the formula indicates that cluster standard errors can be extremely conservative in general.

We propose two new variance estimators

- CCV: causal cluster variance
- TSCB: two-stage cluster bootstrap

to address over-estimation of the variance. They require within-cluster variation in treatment assignment.

Similar calculations for the fixed-effects estimator (FE).

College effects in a U.S. Census sample

Dependent variable: Log labor earnings

Panel A

Treatment: State indicator for share of some college greater than 0.55

	OLS
coefficient	0.1022
standard error:	
robust	(0.0012)
cluster	(0.0312)

Panel B

Treatment: Individual indicator for some college

	OLS	FE
coefficient	0.4656	0.4570
standard error:		
robust	(0.0012)	(0.0012)
cluster	(0.0269)	(0.0276)
cluster (CCV)	(0.0035)	(0.0014)
cluster (TSCB)	(0.0036)	(0.0014)

Simulation design

- Simulations based on the Census data on log earnings ($Y_{k,i}$), and indicators for college attendance ($W_{k,i}$) and state of residence ($m_{k,i}$) for 2,632,838 individuals.
- 50 states plus Washington DC and Puerto Rico for a total of 52 clusters.
- We calibrate the distribution of $A_{k,m}$ using the distribution of $\overline{W}_{k,m}$.
- We assign potential outcomes as $y_{k,i}(0) = Y_{k,i} - \widehat{\tau}_{k,m} W_{k,i}$ and $y_{k,i}(1) = Y_{k,i} + \widehat{\tau}_{k,m}(1 - W_{k,i})$.
- We create samples for given values of p_k and q_k.

Simulation results

Average standard errors across simulations

					normalized standard error			
		$N_k^{1/2}$s.d.	$v_k^{1/2}$	$\widetilde{v}_k^{1/2}$	robust	cluster	CCV	TSCB
Baseline design:								
$p_k = 1,\ q_k = 1,$	OLS	5.91	5.90		1.90	44.86	6.32	5.80
$\sigma_{\tau_k} = .120,\ \sigma_k = .057$	FE	2.34		2.32	1.90	44.63	2.31	2.29
Second design:								
$p_k = .1,\ q_k = 1,$	OLS	2.61	2.59		1.90	14.28	3.78	2.60
$\sigma_{\tau_k} = .120,\ \sigma_k = .057$	FE	1.95		1.95	1.90	14.21	1.95	1.94
Third design:								
$p_k = .1,\ q_k = 1,$	OLS	14.50	14.17		1.98	56.46	13.70	14.33
$\sigma_{\tau_k} = .480,\ \sigma_k = .206$	FE	12.14		11.89	2.13	56.79	11.61	12.07
Fourth design:								
$p_k = .1,\ q_k = 1,$	OLS	9.39	9.39		1.90	8.20	9.19	9.37
$\sigma_{\tau_k} = 0,\ \sigma_k = .206$	FE	2.04		2.04	2.04	1.97	2.04	2.09
Fifth design:								
$p_k = .1,\ q_k = 1,$	OLS	1.95	1.97		1.97	56.42	4.53	2.04
$\sigma_{\tau_k} = .480,\ \sigma_k = 0$	FE	1.91		1.94	1.94	56.42	1.96	1.90

Conclusions

- We propose a inferential framework aimed to address a question of central relevance for empirical practice: when and how we should cluster standard errors.
- We shift the attention from estimation of features of a data-generating process (i.e., infinite superpopulation) to estimation of average treatment effects of the finite population at hand.
- In this framework, the decision on when and how to cluster standard errors depends on the nature of the sampling and the assignment processes only, and not on the presence of within-cluster error components in the outcome variable.

Conclusions

- We derive expressions of the large sample variances of the OLS and FE estimators of the average treatment effect for a setting with clustered sampling and where assignment is random within clusters with assignment probabilities that may vary across clusters.
- For this setting, we demonstrate that robust standard errors can be too small and conventional cluster standard errors can be unnecessarily large.
- We propose two novel procedures, CCV and TSCB, that can be used to calculate more precise standard errors in settings with large clusters and where there is enough variation in treatment assignment within cluster.
- While the variance calculations are for a particular setting, the general principles of the framework are valid for other settings and estimators.