When Should You Adjust Standard Errors for Clustering?

Alberto Abadie, Susan Athey,

Guido Imbens, & Jeffrey Wooldridge

Session in Honor of Gary Chamberlain

MAWM Econometric Society, January 3rd, 2021


EXIT
HARVARD

Motivating Example: You have a random sample of individuals from the US of size $N = 10,000$. You flip a fair coin to assign each of them to a job training program or not. Regressing the outcome on the treatment, and using robust standard errors, you find

$$\widehat{\tau} = 0.058 \quad \text{(s.e. 0.011)}$$

Your RA realizes you know which of the 50 states these individuals live in, and suggests clustering by state.

Question: What do you tell the RA, and why?

1. Yes, definitely cluster.

2. No, definitely do not cluster.

3. Does not matter.

Second Question: Would your answer change if the RA suggested clustering by gender?

How would you explain the answers (and possibly the difference between the answer for clustering on state and the clustering on gender)?

What is the principle that governs the choice to cluster or not?

Some Views from the Econometric Literature

Key is random component that is common to units within group:

• "The clustering problem is caused by the presence of a common unobserved random shock at the group level that will lead to correlation between all observations within each group" (Hansen, 2007, p. 671)

When in doubt, cluster:

• "The consensus is to be conservative and avoid bias and to use bigger and more aggregate clusters when possible, up to and including the point at which there is concern about having too few clusters." (Cameron and Miller, 2015, p. 333)

The RA goes ahead anyway, and, using the Liang-Zeger (STATA) clustered standard errors, comes back with:

$$\widehat{\tau} = 0.058 \quad \text{(s.e. } 0.067\text{)}$$

(where the standard error was 0.011 before).

- Are you confident that the program has a non-zero effect?

- Which standard errors would you report?

- Adjusting standard errors for clustering is common in empirical work.
  - Formal motivation not always clear.
  - Choice of level of clustering is not always clear.
- We present a framework for thinking about clustering that clarifies when/how to adjust for clustering.
  - Mostly exact calculations in simple cases.
  - Clarifies role of large number of clusters asymptotics.

NOT about small sample issues, either small number of clusters or small number of units (Donald and Lang, 2007),
NOT about serial correlation issues (Bertrand et al, 2003).
(both important, but not key to issues discussed here)

Context

Part of set of papers focusing on design/randomization-based inference for causal effects.

• Uncertainty comes at least partly, and sometimes entirely, from assignment process, rather than from sampling process.

• The econometrics literature traditionally (mistakenly) focuses exclusively on sampling based uncertainty, which leads to confusion and incorrect standard errors

• See Abadie-Athey-Imbens-Wooldridge (2020) paper on standard errors for regression estimators when you observe the entire population.

• Other cases: staggered adoption (Athey-Imbens, 2015), difference-in-differences (Bottmer, Imbens, Spiess, Warnick, 2021)

Sampling-based Uncertainty (✓ is observed, ? is missing)

Unit	Actual Sample			Alternative Sample I			Alternative Sample II			...
	Y_i	Z_i	R_i	Y_i	Z_i	R_i	Y_i	Z_i	R_i	...
1	✓	✓	1	?	?	0	?	?	0	...
2	?	?	0	?	?	0	?	?	0	...
3	?	?	0	✓	✓	1	✓	✓	1	...
4	?	?	0	✓	✓	1	?	?	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
M	✓	✓	1	?	?	0	?	?	0	...

Design-based Uncertainty (✓ is observed, ? is missing)

Unit	Actual Sample				Alternative Sample I				Alterna Sampl	
	$Y_i(1)$	$Y_i(0)$	X_i	R_i	$Y_i(1)$	$Y_i(0)$	X_i	R_i	$Y_i(1)$	$Y_i(0)$
1	✓	?	1	1	✓	?	1	1	?	✓
2	?	✓	0	1	?	✓	0	1	?	✓
3	?	✓	0	1	✓	?	1	1	✓	?
4	?	✓	0	1	?	✓	0	1	✓	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
M	✓	?	1	1	?	✓	0	1	?	✓

Clustering Setup

Data on (Y_i, D_i, G_i), $i = 1, \ldots, N$

Y_i is outcome

D_i is regressor, mainly focus on special case where $D_i \in \{-1, 1\}$ (to allow for exact results).

$G_i \in \{1, \ldots, G\}$ is group/cluster indicator.

Estimate regression function

$$Y_i = \alpha + \tau \cdot D_i + \varepsilon_i = X_i'\beta + \varepsilon, \qquad X_i' = (1, D_i), \qquad \beta' = (\alpha, \tau)$$

Least squares estimator (not generalized least squares)

$$(\widehat{\alpha}, \widehat{\tau}) = \arg\min \sum_{i=1}^{N} (Y_i - \alpha - \tau \cdot D_i)^2 \qquad \widehat{\beta} = (\widehat{\alpha}, \widehat{\tau})'$$

Residuals

$$\widehat{\varepsilon}_i = Y_i - \widehat{\alpha} - \widehat{\tau} \cdot D_i$$

Focus of the paper is on properties of $\widehat{\tau}$:

- What is variance of $\widehat{\tau}$?

- How do we estimate the variance of $\widehat{\tau}$?

Standard Textbook Approach:

View D and G as fixed, assume

$$\varepsilon \sim \mathcal{N}(0, \Omega) \qquad \Omega = \begin{pmatrix} \Omega_1 & 0 & \dots & 0 \\ 0 & \Omega_2 & \dots & 0 \\ \vdots & & \ddots & \\ 0 & \dots & & \Omega_G \end{pmatrix}.$$

Variance estimators differ by assumptions on Ω_g:

- diagonal (robust, Eicker-Huber-White),
- constant off-diagonal within clusters (Moulton/Kloek)
- unrestricted (cluster, Liang-Zeger/Stata)

Common Variance estimators (normalized by sample size)
Eicker-Huber-White, standard robust var (zero error covar):

$$\hat{\mathbb{V}}_{\text{robust}} = N \left(\sum_{i=1}^{N} X_i X_i' \right)^{-1} \left(\sum_{i=1}^{N} X_i X_i' \hat{\varepsilon}_i^2 \right) \left(\sum_{i=1}^{N} X_i X_i' \right)^{-1}$$

Liang-Zeger, STATA, standard clustering adjustment, (unrestricted within-cluster covariance matrix):

$$\hat{\mathbb{V}}_{\text{cluster}} = N \left(\sum_{i=1}^{N} X_i X_i' \right)^{-1} \sum_{g=1}^{G} \left(\sum_{i:G_i=g} X_i \hat{\varepsilon}_i \right) \left(\sum_{i:G_i=g} X_i \hat{\varepsilon}_i \right)' \left(\sum_{i=1}^{N} X_i X_i' \right)^{-1}$$

Moulton/Kloek (constant covariance within-clusters)

$$\hat{\mathbb{V}}_{\text{moulton}} = \hat{\mathbb{V}}_{\text{ols}} \cdot \left(1 + \rho_\varepsilon \cdot \rho_D \cdot \frac{N}{G} \right)$$

where ρ_ε, ρ_D are the within-cluster correlations of $\hat{\varepsilon}$ and D.

Related Literature

• Clustering: Moulton (1986, 1987, 1990), Kloek (1981) Hansen (2007), Cameron & Miller (2015), Angrist & Pischke (2008), Liang and Zeger (1986), Wooldridge (2010), Donald and Lang (2007), Bertrand, Duflo, and Mullainathan (2004)

• Sample Design: Kish (1965)

• Causal Literature: Neyman (1935, 1990), Rubin (1976, 2006), Rosenbaum (2000), Imbens and Rubin (2015)

• Exper. Design: Murray (1998), Donner and Klar (2000)

• Finite Population Issues: Abadie, Athey, Imbens, and Wooldridge (2014)

Define the Population and Estimand

Population of size M.

Population is partitioned into G groups/clusters.

The population size in cluster g is M_g, sometimes $M_g = M/G$ for all clusters (for convenience, not essential).

$G_i \in \{1, \ldots, G\}$ is group/cluster indicator.

M may be large, G may be large, M_g may be large, but all finite.

$R_i \in \{0, 1\}$ is sampling indicator, $\sum_{i=1}^{M} R_i = N$ is sample size.

1. Descriptive Question:

Outcome Y_i

Estimand is population average

$$\theta^* = \frac{1}{M} \sum_{i=1}^{M} Y_i$$

Estimator is sample average

$$\widehat{\theta} = \frac{1}{N} \sum_{i=1}^{M} R_i \cdot Y_i$$

definitions:

$$\sigma_g^2 = \frac{1}{M_g - 1} \sum_{i:G_i=g} \left(Y_i - \overline{Y}_{M,g}\right)^2 \qquad \overline{Y}_g = \frac{G}{M} \sum_{i:G_i=g} Y_i$$

$$\sigma^2_{\text{cluster}} = \frac{1}{G-1} \sum_{g=1}^{G} \left(\overline{Y}_g - \overline{\overline{Y}}\right)^2$$

$$\sigma^2_{\text{cond}} = \frac{1}{G} \sum_{g=1}^{G} \sigma_g^2$$

$$\rho = \frac{G}{M(M-G)} \sum_{i \neq j, G_i = G_j} \frac{(Y_i - \overline{\overline{Y}})(Y_j - \overline{\overline{Y}})}{\sigma^2} \approx \frac{\sigma^2_{\text{cluster}}}{\sigma^2_{\text{cluster}} + \sigma^2_{\text{cond}}}$$

$$\sigma^2 = \frac{1}{M-1} \sum_{i=1}^{M} (Y_i - \overline{\overline{Y}})^2 \approx \sigma^2_{\text{cluster}} + \sigma^2_{\text{cond}}$$

Estimator is

$$\widehat{\theta} = \frac{1}{N} \sum_{i=1}^{M} R_i \cdot Y_i$$

• (random sampling) Suppose sampling is completely random,

$$\text{pr}(R = r) = \left(\begin{array}{c} M \\ N \end{array} \right)^{-1}, \quad \forall r \text{ s.t. } \sum_{i=1}^{M} r_i = N.$$

Exact variance, normalized by sample size:

$$N \cdot \mathbb{V}(\widehat{\theta}|\text{RS}) = \sigma^2 \cdot \left(1 - \frac{N}{M}\right) \approx \sigma^2$$

(if $N << M$)

What do the variance estimators give us here?

$$\mathbb{E}\left[\widehat{\mathbb{V}}_{\text{robust}} \middle| \text{RS}\right] \approx \sigma^2 \quad \text{correct}$$

$$\mathbb{E}\left[\widehat{\mathbb{V}}_{\text{cluster}} \middle| \text{RS}\right] \approx \sigma^2 \cdot \left\{1 + \rho \cdot \left(\frac{N}{G} - 1\right)\right\} \quad \text{wrong}$$

- **Adjusting the standard errors for clustering can make a difference here**

- **Adjusting standard errors for clustering is wrong here**

Why is the cluster variance wrong here?

Implicitly the cluster variance takes as the estimand the average outcome in a super-population with a large number of clusters. (In that case we dont have a random sample from the population of interest.)

Two takeaways:

1. Be explicit about the estimand / population of interest. Do we have a random sample or a clustered random sample where we only see some of the clusters?

2. You can not tell from the data which variance is correct, because it depends on the question.

Consider a model-based approach:

$$Y_i = X_i'\beta + \varepsilon_i + \eta_{G_i} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2), \quad \eta_g \sim \mathcal{N}(0, \sigma_\eta^2)$$

The standard ols variance expression

$$\mathbb{V}(\widehat{\beta}) = (X'X)^{-1}(X'\Omega X)(X'X)^{-1}$$

is based on resampling units, or resampling both ε and η.

- Cluster variance resamples η_g
- Design-based approach keeps the η_g fixed.

• (clustered sampling) Suppose we randomly select H clusters out of G, and then select N/H units randomly from each of the sampled clusters:

$$\text{pr}(R = r) = \left(\begin{array}{c} G \\ H \end{array} \right)^{-1} \cdot \left(\begin{array}{c} M/G \\ N/H \end{array} \right)^{-H},$$

$$\text{for all } r \text{ s.t. } \forall g \sum_{i:G_i=g} r_i = N/G \vee \sum_{i:G_i=g} r_i = 0.$$

Now the exact variance is

$$N \cdot \mathbb{V}(\widehat{\theta}|\text{CS}) = \sigma^2_{\text{cluster}} \cdot \frac{N}{H} \cdot \left(1 - \frac{H}{G}\right) + \sigma^2_{\text{cond}} \cdot \left(1 - \frac{N}{M}\right)$$

Adjusting standard errors for clustering here can make a difference and is correct here. Failure to do so leads to invalid confidence intervals.

2. Causal Question:

potential outcomes $Y_i(-1), Y_i(1)$, treatment $D_i \in \{-1, 1\}$, realized outcome $Y_i = Y_i(D_i)$,

Estimand is 0.5 times average treatment effect (to make estimand equal to limit of regression coefficient, simplifies calculations later, but not of essence)

$$\theta^* = \frac{1}{M} \sum_{i=1}^{M} (Y_i(1) - Y_i(-1))/2$$

Estimator is

$$\widehat{\theta} = \frac{\sum_{i=1}^{M} R_i \cdot Y_i \cdot (D_i - \overline{D})}{\sum_{i=1}^{M} R_i \cdot (D_i - \overline{D})^2} \quad \text{where} \quad \overline{D} = \frac{\sum_{i=1}^{M} R_i \cdot D_i}{\sum_{i=1}^{M} R_i}$$

$$\varepsilon_i(d) = Y_i(1) - \frac{1}{N} Y_i(d) \quad \overline{\varepsilon}_g(d) = \frac{1}{M_g} \sum_{i:G_i=g} \varepsilon(d)$$

Two special cases where we know what to do:

1. Random Sampling, Random Assignment

• Should not cluster. Clustering variance can be unnecessarily conservative (see example at second slide)

2. Random Sampling, Clustered Assignment (fixed within clusters)

• Should cluster.

Question: what to do if assignment is correlated within clusters (but not perfectly correlated)?

neither cluster nor standard robust variance is correct

- Standard robust variance is correct if no correlation (but not if correlation positive)

- Clustering variance is correct if correlation is perfect (but over-estimates variance if correlation is less than perfect)

no variance estimator available for the general case.

In between case: Random Sampling of Units, Imperfectly Correlated Assignment

Assignment probabilities for clusters are sampled from distribution with mean $1/2$ and standard deviation σ.

- random assignment if $\sigma = 0$
- cluster assignment if $\sigma = 1/2$

$$\text{normalized var}\,(\hat{\tau}) = \text{robust var} + \text{cluster adj}$$

$$\text{robust var} \approx \frac{1}{M}\sum_{i=1}^{M}\left\{2\left(\varepsilon_i(1)^2 + \varepsilon_i(-1)^2\right)\right\}$$

$$\text{correct cluster adj} = \frac{4\sigma^2 N}{C}\left\{\frac{1}{C}\sum_{c=1}^{C} p_c^2(\overline{\varepsilon}_c(1) + \overline{\varepsilon}_c(-1))^2\right\}$$

Liang-Zeger clustering adjustment fixes σ^2 at maximum value of 1/4:

$$\frac{N}{C}\left\{\frac{1}{C}\sum_{c=1}^{C} p_c^2(\overline{\varepsilon}_c(1) + \overline{\varepsilon}_c(-1))^2\right\} \geq \text{correct cluster adj}$$

Conclusion

• Econometric textbook discussions of need and methods for clustering are misguided (more than empirical practice) by focusing on sampling based justification for clustering.

• In empirical work clustering comes from assignment mechanism, not from sampling mechanism.

• Standard Liang-Zeger / STATA clustering adjustment is unneccesarily conservative if assignment is not perfectly correlated.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." Econometrica 88, no. 1 (2020): 265-296.

Angrist, Joshua D., and Jörn-Steffen Pischke. Mostly harmless econometrics: An empiricist's companion. Princeton university press, 2008.

Athey, Susan, and Guido W. Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. No. w24963. National Bureau of Economic Research, 2018.

Bottmer, Lea, Guido Imbens, Jann Spiess, and Merrill Warnick, (2021), "A Design-based Perspective on Synthetic Control Methods."

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How much should we trust differences-in-differences estimates?." The Quarterly journal of economics 119, no. 1 (2004): 249-275.

Cameron, A. Colin, and Douglas L. Miller. "A practitioner's guide to cluster-robust inference." Journal of human resources 50, no. 2 (2015): 317-372.

Donald, Stephen G., and Kevin Lang. "Inference with difference-in-differences and other panel data." *The review of Economics and Statistics* 89, no. 2 (2007): 221-233.

Hansen, Christian B. "Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects." Journal of econometrics 140, no. 2 (2007): 670-694.

Imbens, Guido W., and Donald B. Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, 2015.

Kloek, Teunis. "OLS estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated." Econometrica: (1981): 205-207.

Leslie Kish. Survey sampling. Vol. 60. Wiley-Interscience, 1995.

Liang, Kung-Yee, and Scott L. Zeger. "Longitudinal data analysis using generalized linear models." Biometrika 73, no. 1 (1986): 13-22.

Moulton, Brent R. "Random group effects and the precision of regression estimates." Journal of econometrics 32, no. 3 (1986): 385-397.

Moulton, Brent R. "Diagnostics for group effects in regression analysis." Journal of Business & Economic Statistics 5, no. 2 (1987): 275-282.

Moulton, Brent R. "An illustration of a pitfall in estimating the effects of aggregate variables on micro units." The review of Economics and Statistics (1990): 334-338.

Murray, David M. Design and analysis of group-randomized trials. Vol. 29. Oxford University Press, USA, 1998.