When Should You Adjust Standard Errors for Clustering?

by Alberto Abadie, Susan Athey, Guido W. Imbens and Jeffrey Wooldridge

Presented by Maren Vairo
Applied Economics Reading Group
UC3M

February, 2018

Motivation

In empirical work, it is common to report standard errors that account for clustering of units.

What is the motivation for this adjustment?

- Typically, the stated motivation is that unobserved components of outcomes are correlated for units within clusters.
- Hansen (2007): "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"
- This motivation makes it difficult to justify clustering over some dimension, and not others.

Motivation

What is the appropriate level of clustering?

- Typically, go for the most aggregate level feasible.
- Cameron and Miller (2015): "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."
- But there is harm in clustering at too aggregate level.

Clustering is a design problem

The confusion rises from the dominant model-based perspective on clustering.
Clustering is in essence a **design problem**:

- *Sampling design*: the sampling follows a two stage process, where first a subset of clusters are randomly sampled and second units are sampled randomly from the sampled clusters
- *Experimental design*: clusters of units, rather than units, are assigned to treatment

Clustering is a design problem

This design perspective, applied to randomization inference, clarifies the role of clustering adjustments and aids in the decision whether to, and at what level to, cluster.

- Contrary to common wisdom, correlation between residuals or/and regressors are neither necessary nor sufficient conditions for cluster adjustments to matter
- The data are informative about whether clustering matters, but they are only partially informative about whether one should cluster
- The question of whether to, and at what level to, cluster cannot be informed solely by data
- Researchers need to address two issues: how units in the sample were selected and how units were assigned to the various treatments

The model-based approach to clustering

The model is:

$$Y_i = \alpha + \tau W_i + \epsilon_i = \beta' X_i + \epsilon_i$$

- A scalar outcome variable Y_i
- A binary covariate W_i
- Units belong to clusters $C_i \in \{1, ..., C\}$
- $\beta' = (\alpha, \tau)$ and $X_i' = (1, W_i)$

Under $E(\epsilon|\mathbf{X}, \mathbf{C}) = 0$ and $E(\epsilon\epsilon'|\mathbf{X}, \mathbf{C}) = \mathbf{\Omega}$:

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

The model-based approach to clustering

Different assumptions on Ω give a different variance structure:

- Without clustering and under homoskedasticity:

$$\mathbb{V}_{OLS}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- Without clustering and allowing for heteroskedasticity:

$$\mathbb{V}_{EHW}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}(\sum_{i=1}^{N} \Omega_{ii} X_i X_i')(\mathbf{X}'\mathbf{X})^{-1}$$

The model-based approach to clustering

- Assuming that clusters are equal size and under:

$$\Omega_{ij} = \begin{cases} 0 \text{ if } C_i \neq C_j \\ \rho\sigma^2 \text{ if } C_i = C_j, i \neq j \\ \sigma^2 \text{ if } i = j \end{cases}$$

$$\mathbb{V}_{KLOEK}(\hat{\tau}) = \mathbb{V}_{OLS}(\hat{\tau})(1 + \rho_\epsilon \rho_W \frac{N}{C})$$

- While letting Ω_{ij} with $C_i = C_j$ unrestricted:

$$\mathbb{V}_{LZ}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}(\sum_{c=1}^{C} \mathbf{X}_\mathbf{c}'\mathbf{\Omega}_\mathbf{c}\mathbf{X}_\mathbf{c})(\mathbf{X}'\mathbf{X})^{-1}$$

Misconception 1: Clustering matters only if the residuals and regressors are both correlated within clusters

There appears to be a view, captured by the expression for V_{KLOEK} that clustering does not matter if either ρ_ϵ or ρ_W are zero. If this was true, clustering would not matter when:

- Treatment is completely randomly assigned
- Cluster fixed effects are included in the regression

To illustrate the fallacy of this view, simulate a data set with $N = 100,323$, $C = 100$, where the number of units in each cluster ranges from 950 to 1063. For each unit, observe Y_i, W_i and C_i, and estimate the regression function above.

Misconception 1: Clustering matters only if the residuals and regressors are both correlated within clusters

Results:

$$\hat{\rho}_{\hat{\epsilon}} = 0.001 \quad \hat{\rho}_W = 0.001$$

Both within cluster correlations are close to zero, and because there is only modest variation on cluster sizes, the standard Moulton-Kloek adjustment would essentially be zero.
However:

$$\hat{\tau}^{ls} = -0.120 \quad (se_{EHW} = 0.004) \quad [se_{LZ} = 0.100]$$

Clustering matters substantially!

So inspecting the within-cluster correlations of the residuals and regressors is not necessarily informative about whether clustering standard errors using the Liang-Zenger estimator matters.

Misconception 2: If clustering matters, one should cluster

- There is also a common view that there is no harm, at least in large samples, to adjusting the standard errors for clustering.
- If clustering matters it should be done, and if it does not matter it does no harm.
- Based on this perception, many discussions recommend to calculate diagnosis on the sample to inform the decision whether or not one should cluster.
- However, the decision should be based on substantive information, not solely on whether it makes a difference.

Misconception 2: If clustering matters, one should cluster

To see this, consider:

- A population with 10,000,000 units, equally partitioned in 100 clusters
- A sample of 100,000 is 10,000 times randomly drawn
- $W_i \in \{0, 1\}$ is randomly assigned with probability $1/2$, independent of everything else
- Y_i is generated as: $Y_i = \tau_{C_i} W_i + \nu_i$, with $\tau_c = -1$ and $\tau_c = 1$ for half the clusters, ν_i are iid.
- The ATE is $\tau = 0$ in this context

Misconception 2: If clustering matters, one should cluster

Table 1: Standard Errors and Coverage Rates Random Sampling, Random Assignment (10,000 replications)

No Fixed Effects				Fixed Effects			
EHW variance		LZ variance		EHW variance		LZ variance	
$\sqrt{\mathbb{V}_{\text{EHW}}}$	cov rate	$\sqrt{\mathbb{V}_{\text{LZ}}}$	cov rate	$\sqrt{\mathbb{V}_{\text{EHW}}}$	cov rate	$\sqrt{\mathbb{V}_{\text{LZ}}}$	cov rate
0.007	0.950	0.051	1.000	0.007	0.950	0.131	0.986

In this example, EHW standard errors are the appropriate ones, although the LZ standard errors are substantially larger!

Misconception 2: If clustering matters, one should cluster

The LZ standard errors are based on the presumption that there are clusters in the population beyond the 100 clusters that are seen in the sample.It is this presumption that is critical, and often implicit, in the model-based motivation for clustering standard errors.

Obviously, one can not tell from the sample itself if such clusters exist in the population. So one needs to choose between the two standard errors on the basis of substantive knowledge of the study design.

A Formal Result: The Sequence of Populations

- A sequence of populations, each with M_n units and C_n clusters, where M_n is strictly increasing and C_n is weakly increasing in n
- For each unit, there are two potential outcomes: $Y_{in}(0)$ and $Y_{in}(1)$, and two treatment-specific residuals
  $\epsilon_{in}(w) = Y_{in}(w) - \frac{1}{M_n} \sum_{j=1}^{M} {}_n Y_{jn}(w)$ for $w = 0, 1$

- We are interested in the population average treatment effect:
  $\tau_n = \frac{1}{M_n} \sum_{i=1}^{M} {}_n (Y_{in}(1) - Y_{in}(0)) = \overline{Y_n}(1) - \overline{Y_n}(0)$

A Formal Result: The Sampling Process and the Assignment Mechanism

- We observe (Y_{in}, W_{in}, C_{in}) for a subset of the population, where $Y_{in} = Y_{in}(W_{in})$. R_{in} indicates if unit i was sampled, $N_n = \sum_{i=1}^{M_n} R_{in}$ is the number of units sampled
- The sampling process is independent of the potential outcomes and the assignment. It consists of two stages:
  - Clusters are sampled with probability P_{C_n}
  - Units are sampled from the population of sampled clusters with probability P_{U_n}
- The assignment process that determines W_{in} also follows two stages:
  - For cluster c in population n, an assignment probability q_{cn} is randomly drawn from a distribution $f(.)$ with mean μ_n and variance ${\sigma^2}_n$. For simplicity, assume $\mu_n = 1/2$
  - Each unit in c is assigned treatment with cluster-specific probability q_{cn}

A Formal Result: The Estimator

The LS estimator of τ_n is:

$$\hat{\tau_n} = \frac{\sum_{i=1}^{n} R_{in}(W_{in} - \overline{W_n})Y_{in}}{\sum_{i=1}^{n} R_{in}(W_{in} - \overline{W_n})^2} = \overline{Y}_{n1} - \overline{Y}_{n0}$$

In the current setting and under some regularity conditions (Ass.1: the number of clusters increases without limit as n increases, the relative cluster sizes are bounded, and the potential outcomes are bounded):

$$\sqrt{N_n}(\hat{\tau_n} - \tau_n) - \eta_n = o_p(1)$$

Where:

$$\eta_n = \frac{2}{\sqrt{M_n P_{C_n} P_{U_n}}} \sum_{i=1}^{M_n} R_{in}(2W_{in} - 1)\epsilon_{in}$$

A Formal Result: The Estimator

In this setting, the main results are:
i) The variance of interest is:

$$\mathbb{V}(\eta_n) = \frac{1}{M_n}\sum_{i=1}^{M_n}\{2(\epsilon_{in}(1)^2 + \epsilon_{in}(0)^2) - P_{U_n}(\epsilon_{in}(1) - \epsilon_{in}(0))^2 + 4P_{U_n}{\sigma_n}^2(\epsilon_{in}(1) - \epsilon_{in}(0))^2\}$$
$$+\frac{P_{U_n}}{M_n}\sum_{c=1}^{C_n}{M_{cn}}^2\{(1 - P_{C_n})(\overline{\epsilon}_{cn}(1) - \overline{\epsilon}_{cn}(0))^2 + 4{\sigma_n}^2(\overline{\epsilon}_{cn}(1) + \overline{\epsilon}_{cn}(0))^2\}$$

Where:

- The first sum is approximately the EHW variance
- The second sum captures the effects of clustered sampling and clustered assignment

A Formal Result: The Estimator

ii) How does the LZ variance compare to the correct variance?

$$\mathbb{V}_{LZ} - \mathbb{V}(\eta_n) = \frac{P_{C_n} P_{U_n}}{M_n} \sum_{c=1}^{C_n} {M_{cn}}^2 (\bar{\epsilon}_{cn}(1) - \bar{\epsilon}_{cn}(0))^2 \geq 0$$

- LZ captures correctly the component of clustering due to clustered assignment.
- It does not capture correctly the component due to clustered sampling unless P_{C_n} is close to zero

A Formal Result: The Estimator

iii) When does the cluster adjustment matter?

$$\mathbb{V}_{LZ} - \mathbb{V}_{EHW} = -\frac{2P_{U_n}}{M_n}\sum_{i=1}^{M_n}\{(\epsilon_{in}(1) - \epsilon_{in}(0))^2 + 4{\sigma_n}^2(\epsilon_{in}(1) + \epsilon_{in}(0))^2\}$$
$$+\frac{P_{U_n}}{M_n}\sum_{c=1}^{C_n}{M_{cn}}^2\{(\bar{\epsilon}_{cn}(1) - \bar{\epsilon}_{cn}(0))^2 + 4{\sigma_n}^2(\bar{\epsilon}_{cn}(1) + \bar{\epsilon}_{cn}(0))^2\}$$

- The first sum is small relative to the second if there is a substantial number of units per cluster relative to to the number of clusters.
- In that case, clustering matters if there is heterogeneity of treatment effects or there is clustering in the assignment
- The difference does not depend on whether the sampling is clustered

A Formal Result: The Estimator

Corollary 1

Standard errors need to account for clustering unless one of the following pairs of conditions holds: (i) there is no clustering in the sample ($P_{C_n} = 1$) and there is no clustering in the assignment (${\sigma_n}^2 = 0$); or (ii) there is no heterogeneity in the treatment effects ($Y_{in}(1) - Y_{in}(0) = \tau_n$) and there is no clustering in the assignment (${\sigma_n}^2 = 0$).

Corollary 2

The LZ variance is approximately correct if one of three conditions holds: (i) there is no heterogeneity in treatment effects ($Y_{in}(1) - Y_{in}(0) = \tau_n$); (ii) P_{C_n} is close to zero, so that we observe only few clusters in the population of clusters; (iii) P_{U_n} is close to zero, so that there is at most one sampled unit per cluster.

A Formal Result: The Estimator

When P_{C_n} is high, the LZ variance is in general too conservative:

- If the assignment is perfectly correlated within clusters, there is no general improvement over LZ available
- If there is within-cluster variation in the treatment, an improvement is available. Consider the estimated cluster specific treatment effect: $\hat{\tau_c} = \overline{Y}_{c1} - \overline{Y}_{c0}$. Then, the proposed cluster-adjusted variance estimator is:

$$\hat{\mathbb{V}}_{CA}(\hat{\tau}) = \hat{\mathbb{V}}_{LZ}(\hat{\tau}) - \frac{1}{N^2} \sum_{c=1}^{C} {N_c}^2 (\hat{\tau_c} - \hat{\tau})^2$$

Fixed effects

What happens if we include cluster fixed effects in the regression?

With fixed effects, one should cluster if either: (i) both $P_{C_n} < 1$ *and there is heterogeneity in the treatment effects, or (ii)* $\sigma^2 > 0$ *and there is heterogeneity in the treatment effects.*

Heterogeneity in the treatment effects is now a requirement for clustering adjustments to be necessary.

Conclusion

The practical implications from the results are:

- The researcher should assess whether the sampling and assignment mechanisms are clustered or not
- If the answer to both is no, one should not adjust the standard errors for clustering, irrespective of whether such an adjustment would change the standard errors.
- The Liang-Zeger adjustment is in general conservative and can be improved upon if there is within-cluster variation in treatment assignment and the fraction of clusters that is observed is known.
- This analysis extends to the fixed effects case with the provision that, if there is no heterogeneity in treatment effects, one need not adjust standard errors for clustering once fixed effects are included.