Clustering as a Design Problem

Alberto Abadie, Susan Athey,

Guido Imbens, & Jeffrey Wooldridge

Harvard-MIT Econometrics Seminar

Cambridge, February 4, 2016

• Adjusting standard errors for clustering is common in empirical work.

• Motivation not always clear.

• Implementation is not always clear.

• We present a coherent framework for thinking about clustering that clarifies when and how to adjust for clustering.

• Mostly exact calculations in simple cases.

• Clarifies role of large number of clusters.

NOT about small sample issues, either small number of clusters or small number of units, **NOT** about serial correlation issues. (Important, but not key to issues discussed here)

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

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

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 is on properties of $\widehat{\tau}$:

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

Standard approach:

View $\mathbf{D}$ and $\mathbf{G}$ as fixed, assume

$$\varepsilon \sim \mathcal{N}(0, \Omega)$$

Ω block diagonal, correspondig to clusters

$$\Omega = \begin{pmatrix} \Omega_1 & 0 & \dots & 0 \\ 0 & \Omega_2 & \dots & 0 \\ \vdots & & \ddots & \\ 0 & \dots & & \Omega_G \end{pmatrix}.$$

Variance estimators differ by assumptions on Ω_g: diagonal (robust, Eicker-Huber-White), unrestricted (cluster, Liang-Zeger/Stata), constant off-diagonal (Moulton/Kloek)

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

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

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

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

Moulton/Kloek (constant covariance within-clusters)

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

where ρ_ε, ρ_D are the within-cluster correlations of $\widehat{\varepsilon}$ 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)

Views from the Literature

• "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, p. 671)

• "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, p. 333)

• Clustering does not matter when the regressors are not correlated within clusters.

• Use $\widehat{\mathbb{V}}_{\text{cluster}}$ when in doubt.

Questions

1. Is there any harm in using $\widehat{\mathbb{V}}_{\text{cluster}}$ when $\widehat{\mathbb{V}}_{\text{robust}}$ is valid?

2. Can we infer from the data whether $\widehat{\mathbb{V}}_{\text{cluster}}$ or $\widehat{\mathbb{V}}_{\text{robust}}$ is appropriate?

3. When are $\widehat{\mathbb{V}}_{\text{cluster}}$, $\widehat{\mathbb{V}}_{\text{robust}}$, or $\widehat{\mathbb{V}}_{\text{moulton}}$ appropriate?

4. Is $\widehat{\mathbb{V}}_{\text{cluster}}$ superior to $\widehat{\mathbb{V}}_{\text{robust}}$ in large samples?

5. What is the role of within-cluster correlation of regressors?

We develop a framework within which these questions can be answered.

Key features:

- Specify population and estimand
- Specify data generating process

Answers

1. Is there any harm in using $\widehat{\mathbb{V}}_{\text{cluster}}$ when $\widehat{\mathbb{V}}_{\text{robust}}$ is valid? **YES**

2. Can we infer from the data whether $\widehat{\mathbb{V}}_{\text{cluster}}$ or $\widehat{\mathbb{V}}_{\text{robust}}$ is appropriate? **NO**

3. When are $\widehat{\mathbb{V}}_{\text{cluster}}$ or $\widehat{\mathbb{V}}_{\text{robust}}$ appropriate? **DEPENDS ON DESIGN**

4. Is $\widehat{\mathbb{V}}_{\text{cluster}}$ superior to $\widehat{\mathbb{V}}_{\text{robust}}$ in large samples? **DEPENDS ON DESIGN**

5. What is the role of within-cluster correlation of regressors? **DEPENDS ON DESIGN**

First, Define the Population and Estimand

Population of size M.

Population is partioned into G groups/clusters.

The population size in cluster g is M_g, here $M_g = M/G$ for all clusters for convenience.

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

M may be large/infinite, G may be large/infinite, M_g may be large/infinite.

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

1. **Descriptive Setting**:

Outcome Y_i

Estimand is population average

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

Estimator is sample average

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

2. Causal Setting:

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

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

Descriptive Setting: population definitions

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

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

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

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

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

Estimator is

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

• (random sampling) Suppose sampling is completely random,

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

Exact variance, normalized by sample size:

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

What do the variance estimators give us here?

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

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

- **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. The set of clusters that we see in the sample is just a small subset of that large population of clusters.

In that case we dont have a random sample from the population of interest.

• Be explicit about the population of interest. Do we see all clusters in the population or not.

• This issue is distinct from the use of distributional approximations based on increasing the number of clusters.

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

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

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

In a random sample we will eventually see units from all clusters, and we do not need to resample the η_g.

The random sampling variance 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:

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

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

Now the exact variance is

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

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

Four Causal Settings

- Random sample, random assignment of units.
- Random sample, random assignment of clusters.
- Clustered sample, random assignment of units.
- Random sample, assignment prob varying across clusters.

Questions

1. Is $\widehat{\mathbb{V}}_{\text{robust}}$ valid?
2. Is $\widehat{\mathbb{V}}_{\text{cluster}}$ valid?

Answers

- Random sample, random assignment of units.
  $\widehat{\mathbb{V}}_{\text{robust}}$ **valid** $\widehat{\mathbb{V}}_{\text{cluster}}$ **not generally valid**

- Random sample, random assignment of clusters.
  $\widehat{\mathbb{V}}_{\text{robust}}$ **not generally valid** $\widehat{\mathbb{V}}_{\text{cluster}}$ **valid**

- Clustered sample, random assignment of units.
  depends on estimand: average effect in population versus average effect in sample

- Random sample, assignment prob varying across clusters.
  neither generally valid

Causal Setting: Random Sampling, Random Assignment

Points:

1. Should not cluster.

2. $\widehat{\mathbb{V}}_{\text{robust}}$ is valid

3. $\widehat{\mathbb{V}}_{\text{cluster}}$ can be different from $\widehat{\mathbb{V}}_{\text{robust}}$ in large samples, with many clusters, even with $\rho_\varepsilon = \rho_D = 0$.

4. $\widehat{\mathbb{V}}_{\text{moulton}}$ and $\widehat{\mathbb{V}}_{\text{cluster}}$ are conceptually quite different.

Example. Data generating process:

$R_i = 1$ (all units are sampled)

$W_i \sim \mathcal{B}(1, 1/2) \quad D_i = 2 \cdot (W_i - 1) \in \{-1, 1\}$

$\tau_i = 1 + \xi_{G_i}, \quad \xi_g \sim \mathcal{B}(1, 1/2)$

$Y_i = \tau_i \cdot D_i + \nu_i \quad \nu_i \sim \mathcal{N}(0, 1)$

Estimated regression

$Y_i = \alpha + \tau \cdot D_i + \varepsilon \quad \mathbf{NOTE}: \ \rho_D = 0, \quad \rho_\varepsilon = 0$

Random Sampling, Random or Clustered Assignment

	random assignm	clustered assignm
standard deviation	0.05	0.13
$\sqrt{\widehat{\mathbb{V}}_{\text{robust}}}$	0.05	0.04
coverage rate $\widehat{\mathbb{V}}_{\text{robust}}$	0.96	0.48
$\sqrt{\widehat{\mathbb{V}}_{\text{cluster}}}$	0.12	0.16
coverage rate $\widehat{\mathbb{V}}_{\text{cluster}}$	1.00	0.97

Causal Setting: Fuzzy Clustering

Suppose:

$$\mathbb{E}[D_i] = 0,\ \mathbb{E}[D_i \cdot D_j | G_i = G_j] = \gamma,\quad \mathbb{E}[D_i \cdot D_j | G_i \neq G_j] = 0.$$

Assignment is correlated within clusters, but not perfectly correlated.

- $\widehat{\mathbb{V}}_{\text{cluster}}$ cannot be right, because it is wrong if $\gamma = 0$
- $\widehat{\mathbb{V}}_{\text{robust}}$ cannot be right, because it is wrong if $\gamma = 1$
- So, what do we do?

- Will look at simple case where exact calculations are possible.

- Will propose new variance estimator that can deal with
  - random assignment (where it reduces to robust variance),
  - clustered assignment (where it reduces to clustered variance),
  - intermediate correlated assignment cases (for which there is no variance estimator).

Example Data Generating Process

Population size M, G clusters, all equal size, all units sampled, $R_i = 1$.

In $G/2$ randomly selected clusters the fraction of treated units is $1/2 - \delta$, in the remaining clusters the fraction of treated units is $1/2 + \delta$. Hence $\sum_{i=1}^{M} D_i = 0$, $\sum_{i=1}^{M} D_i^2 = M$.

For each unit there are two values $Y_i(-1)$ and $Y_i(1)$. The estimand is $\tau = \sum_{i=1}^{M}(Y_i(1) - Y_i(-1))/(2M)$. The estimator is the ols estimator in a regression

$$Y_i = \alpha + \tau \cdot D_i + \varepsilon$$

leading to

$$\widehat{\tau} = \frac{1}{M} \sum_{i=1} D_i \cdot Y_i$$

Define

$$\varepsilon_i(-1) = Y_i(-1) - \frac{1}{M}\sum_{i=1}^{M} Y_i(-1), \qquad \varepsilon_i(1) = Y_i(1) - \frac{1}{M}\sum_{i=1}^{M} Y_i(1)$$

$$\underline{\varepsilon}_i = \frac{\varepsilon_i(-1) + \varepsilon_i(1)}{2}$$

$$\varepsilon_i = \varepsilon_i(D_i)$$

$$\widehat{\tau} = \frac{\mathbf{Y}'\mathbf{D}}{N} = \tau + \frac{\underline{\varepsilon}'\mathbf{D}}{N}$$

So, true (infeasible) variance is

$$\mathbb{V}(\widehat{\tau}) = \underline{\varepsilon}'\mathbb{V}(\mathbf{D})\underline{\varepsilon}/N^2$$

We need to figure out the exact variance of $\mathbf{D}$ and find an estimator for $\underline{\varepsilon}$. Note that $\mathbb{E}[\mathbf{D}] = 0$, so just need second moments.

Elements of $\mathbb{V}(\mathbf{D}) = \mathbb{E}[\mathbf{D}\mathbf{D}']$:

$$\mathbb{E}[D_i^2] = 1$$

$$\mathbb{E}[D_i \cdot D_j | G_i = G_j, i \neq j] = \frac{4M\delta^2 - G}{M - G} \approx 4\delta^2$$

$$\mathbb{E}[D_i \cdot D_j | G_i \neq G_j] = -\frac{4\delta^2}{G - 1} \approx 0$$

Approximate variance of $\mathbf{D}$ is $\widehat{\mathbb{V}}(\mathbf{D})$:

$$\widehat{\mathbb{V}}(\mathbf{D})_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 4 \cdot \delta^2 & \text{if } i \neq j, G_i = G_j, \\ 0 & \text{otherwise.} \end{cases}$$

We do not observe $\underline{\varepsilon}_i$, but can estimate it:

$$\mathbb{E}[\varepsilon_i] = \underline{\varepsilon}_i$$

So proposed feasible variance estimator is

$$\widehat{\mathbb{V}}_{\mathsf{fc}} = \widehat{\varepsilon}'\widehat{\mathbb{V}}(\mathbf{D})\widehat{\varepsilon}/N^2 \quad \textbf{NEW VARIANCE ESTIMATOR}$$

- If $\delta = 0$ (random assignment), then $\widehat{\mathbb{V}}_{\mathsf{fc}} = \widehat{\mathbb{V}}_{\mathsf{robust}}$
- If $\delta = 1/2$ (clustered assignment), then $\widehat{\mathbb{V}}_{\mathsf{fc}} = \widehat{\mathbb{V}}_{\mathsf{cluster}}$
- Can deal with intermediate cases.

Simulation random sampling, correlated assignment within clusters

$$\frac{Y_i(-1) + Y_i(1)}{2} = \nu_{G_i} + \eta_i, \qquad \nu_g \sim \mathcal{N}(0,1),\ \eta_i \sim \mathcal{N}(0,1)$$

$$\tau_i = \frac{Y_i(1) - Y_i(-1)}{2} = \xi_{G_i} + \omega_i \sim, \qquad \xi_g \sim \mathcal{N}(0,1),\ \omega_i \sim \mathcal{N}(0,1)$$

Three values for δ:

1. $\delta = 0$, stratified assignment

2. $\delta = 1/4$ correlated assignment / fuzzy clustering

3. $\delta = 1/2$ clustered assignment

- std is standard deviation of estimator
- strat is variance est for stratified randomized experiment,
- robust is eicker-huber-white robust variance estimator,
- cluster is liang-zeger (stata) variance estimator,
- fc is proposed feasible variance est for fuzzy clustering
- ifc is infeasible true variance.

Random or Clustered Assignment

		strat		robust		cluster		fc		ifc	
δ	std	se	size	se	size	se	size	se	size	se	size
random	.01	.01	.03	.02	.00	.05	.00	.03	.02	.01	.05
correlated	.05	.01	.62	.02	.51	.07	.01	.05	.04	.05	.05
clust	.10	.01	.81	.02	.74	.11	.03	.11	.03	.10	.05

Summary

What to do depends on the sampling scheme and the assignment of the regressors.

		Sampling		
		random	correlated	clustered
Assignment	random	**robust/fc**	??	**cluster**
	correlated	**fc**	??	**cluster**
	clustered	**cluster/fc**	??	**cluster**