Clustering as a Design Problem

Alberto Abadie, Susan Athey,
Guido Imbens, & Jeffrey Wooldridge

UIUC, February 10, 2017
correlation of $X_i \times \varepsilon_i$
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

\[ \hat{\tau} = 0.058 \text{ (s.e. 0.011)} \]

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.
• Would your answer change if the RA suggested clustering by gender?

• How would you explain the answers (and possibly their difference)?
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, 2007, 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, 2015, p. 333)
The RA goes ahead anyway, and, using the Liang-Zeger (STATA) clustered standard errors, comes back with:

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

(where the standard error was 0.011 before).

- Are you confident that the program has a non-zero effect?
- Which standard errors should you report?
• Adjusting standard errors for clustering is common in empirical work.
  
  – Motivation not always clear.

  – Choice of level of clustering is not always clear.

• We present a coherent framework for thinking about clustering that clarifies when and how to adjust for clustering.
  
  – Currently 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), **NOT** about serial correlation issues (Bertrand et al, 2003). (Important, but not key to issues discussed here)
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

\[ Y_i = \alpha + \tau \cdot D_i + \varepsilon_i = X_i'\beta + \varepsilon, \quad X_i' = (1, D_i), \quad \beta' = (\alpha, \tau) \]
Least squares estimator (not generalized least squares)

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

Residuals

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

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

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

- How do we estimate the variance of $\hat{\tau}$?
Standard Textbook Approach:

View D and G as fixed, assume

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

\[ \Omega = \begin{pmatrix} \Omega_1 & 0 & \ldots & 0 \\ 0 & \Omega_2 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & \Omega_G \end{pmatrix} \]

Variance estimators differ by assumptions on \( \Omega_g \):

- diagonal (robust, Eicker-Huber-White),
- constant off-diagonal (Moulton/Kloek)
- unrestricted (cluster, Liang-Zeger/Stata)
Common Variance estimators (normalized by sample size)

Eicker-Huber-White, standard robust var (zero error covar):

\[
\hat{\sigma}^2_{\text{robust}} = N \left( \sum_{i=1}^{N} X_i X_i' \right)^{-1} \left( \sum_{i=1}^{N} X_i X_i' \widehat{\epsilon}_i^2 \right) \left( \sum_{i=1}^{N} X_i X_i' \right)^{-1}
\]

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

\[
\hat{\sigma}^2_{\text{cluster}} = N \left( \sum_{i=1}^{N} X_i X_i' \right)^{-1} \sum_{g=1}^{G} \left( \sum_{i:G_i=g} X_i \widehat{\epsilon}_i \right) \left( \sum_{i:G_i=g} X_i \widehat{\epsilon}_i \right)' \left( \sum_{i=1}^{N} X_i X_i' \right)
\]

Moulton/Kloek (constant covariance within-clusters)

\[
\hat{\sigma}^2_{\text{moulton}} = \hat{\sigma}^2_{\text{robust}} \cdot \left( 1 + \rho_{\epsilon} \cdot \rho_{D} \cdot \frac{N}{G} \right)
\]

where \( \rho_{\epsilon}, \rho_{D} \) are the within-cluster correlations of \( \widehat{\epsilon} \) and \( D \).
Related Literature


• Sample Design: Kish (1965)


Questions

1. Is there any harm in using $\hat{V}_{\text{cluster}}$ when $\hat{V}_{\text{robust}}$ is valid? Is $\hat{V}_{\text{cluster}}$ superior to $\hat{V}_{\text{robust}}$ in large samples?

2. Can we infer from the data whether $\hat{V}_{\text{cluster}}$ or $\hat{V}_{\text{robust}}$ is appropriate? What is the relevance of within-cluster correlation of regressors and residuals?
We develop a framework within which these questions can be answered.

Key features:

- Specify population and estimand
- Specify design
Answers

1. Is there any harm in using $\hat{\Sigma}_{\text{cluster}}$ when $\hat{\Sigma}_{\text{robust}}$ is valid? Is $\hat{\Sigma}_{\text{cluster}}$ superior to $\hat{\Sigma}_{\text{robust}}$ in large samples?

   **Yes, there is harm:** $\hat{\Sigma}_{\text{cluster}}$ can be unnecessarily conservative relative to $\hat{\Sigma}_{\text{robust}}$, even in large samples.

2. Can we infer from the data whether $\hat{\Sigma}_{\text{cluster}}$ or $\hat{\Sigma}_{\text{robust}}$ is appropriate? What is the relevance of within-cluster correlation of regressors and residuals?

   **The data cannot tell us which variance estimator is right, it depends on the population/estimand/design. The values of within-cluster correlation of regressors and residuals are immaterial.**
Digression Based on Abadie, Athey, Imbens, Wooldridge, 2015.

- Specify the population and what you are interested in.

- Specify what is missing, what is observed, and why these things are missing or observed.
### Sampling-based Uncertainty (✓ is observed, ? is missing)

<table>
<thead>
<tr>
<th>Unit</th>
<th>Actual Sample</th>
<th>Alt Sample I</th>
<th>Alt Sample II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Y_i$ $Z_i$</td>
<td>$Y_i$ $Z_i$</td>
<td>$Y_i$ $Z_i$</td>
</tr>
<tr>
<td>1</td>
<td>✓ ✓</td>
<td>? ?</td>
<td>? ?</td>
</tr>
<tr>
<td>3</td>
<td>✓ ✓</td>
<td>? ?</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>4</td>
<td>? ?</td>
<td>✓ ✓</td>
<td>? ?</td>
</tr>
<tr>
<td>M</td>
<td>✓ ✓</td>
<td>? ?</td>
<td>? ?</td>
</tr>
</tbody>
</table>
Design-based Uncertainty (✓ is observed, ? is missing)

<table>
<thead>
<tr>
<th>Unit</th>
<th>Actual Sample</th>
<th>Alt Sample I</th>
<th>Alt Sample II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Y_i(-1)$</td>
<td>$Y_i(1)$</td>
<td>$Y_i(-1)$</td>
</tr>
<tr>
<td>1</td>
<td>✓</td>
<td>?</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>?</td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>?</td>
<td>✓</td>
</tr>
<tr>
<td>3</td>
<td>?</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>?</td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>$M$</td>
<td>✓</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
EHW and bootstrap standard errors capture sampling uncertainty, assuming population size $M$ is large relative to sample size $N$. They ignore design-based uncertainty.

If the population size is large relative to the sample size, this does not matter.
Back to Clustering. First, Define the Population and Estimand

Population of size $M$.

Population is partitioned into $G$ groups/clusters.

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

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

$M$ may be large/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 Question:

Outcome $Y_i$

Estimand is population average

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

Estimator is sample average

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

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

Estimator is

\[
\hat{\theta} = \frac{\sum_{i=1}^{M} R_i \cdot Y_i \cdot (D_i - \bar{D})}{\sum_{i=1}^{M} R_i \cdot (D_i - \bar{D})^2} \quad \text{where} \quad \bar{D} = \frac{\sum_{i=1}^{M} R_i \cdot D_i}{\sum_{i=1}^{M} R_i}
\]
Descriptive Question: population definitions

\[
\sigma_g^2 = \frac{1}{M_g - 1} \sum_{i:G_i=g} (Y_i - \overline{Y}_{M,g})^2 \quad \overline{Y}_{M,g} = \frac{G}{M} \sum_{i:G_i=g} Y_i
\]

\[
\sigma_{\text{cluster}}^2 = \frac{1}{G - 1} \sum_{g=1}^{G} (\overline{Y}_{M,g} - \overline{Y}_M)^2
\]

\[
\sigma_{\text{cond}}^2 = \frac{1}{G} \sum_{g=1}^{G} \sigma_g^2
\]

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

\[
\sigma^2 = \frac{1}{M - 1} \sum_{i=1}^{M} (Y_i - \overline{Y}_M)^2 \approx \sigma_{\text{cluster}}^2 + \sigma_{\text{cond}}^2
\]
Estimator is

\[ \hat{\theta} = \frac{1}{N} \sum_{i=1}^{M} R_i \cdot Y_i \]

• (random sampling) Suppose sampling is completely random,

\[ \text{pr}(R = r) = \left( \frac{M}{N} \right)^{-1}, \quad \forall r \text{ s.t. } \sum_{i=1}^{M} r_i = N. \]

Exact variance, normalized by sample size:

\[ N \cdot \mathbb{V}(\hat{\theta}|\text{RS}) = \sigma^2 \cdot \left(1 - \frac{N}{M}\right) \approx \sigma^2 \]
What do the variance estimators give us here?

\[
\mathbb{E}\left[ \hat{V}_{\text{robust}} \mid RS \right] \approx \sigma^2
\]

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

- **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 viewed as just a small subset of that large population of clusters.

In that case we don’t have a random sample from the population of interest.

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

- This issue is distinct from the use of distributional approximations based on increasing the number of clusters.
Consider a model-based approach:

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

The standard ols variance expression

\[
\nabla(\hat{\beta}) = (X'X)^{-1}(X'\Omega X)(X'X)^{-1}
\]

is based on resampling units, or resampling both \(\varepsilon\) and \(\eta\).

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

The random sampling variance keeps the \(\eta_g\) 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:

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

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

Now the exact variance is

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

**Adjusting standard errors for clustering here can make a difference and is correct here. Failure to do so leads to invalid confidence intervals.**
Four Settings with Design-based Uncertainty

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

Questions

1. Is $\hat{V}_{\text{robust}}$ valid?
2. Is $\hat{V}_{\text{cluster}}$ valid?
Answers

- Random sample, random assignment of units.
  \( \hat{V}_{\text{robust}} \text{ valid} \quad \hat{V}_{\text{cluster}} \text{ not generally valid} \)

- Random sample, random assignment of clusters.
  \( \hat{V}_{\text{robust}} \text{ not generally valid} \quad \hat{V}_{\text{cluster}} \text{ valid} \)

- 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. \( \hat{V}_{\text{robust}} \) is valid

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

4. \( \hat{V}_{\text{moulton}} \) and \( \hat{V}_{\text{cluster}} \) are conceptually quite different.
Example. Data generating process:

\[ R_i = 1 \quad \text{(all units are sampled)} \]

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

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

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

Estimated regression

\[ Y_i = \alpha + \tau \cdot D_i + \varepsilon \quad \text{NOTE} : \quad \rho_D = 0, \quad \rho_\varepsilon = 0 \]
Random Sampling, Random or Clustered Assignment

<table>
<thead>
<tr>
<th></th>
<th>random assignm</th>
<th>clustered assignm</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard deviation</td>
<td>0.05</td>
<td>0.13</td>
</tr>
<tr>
<td>(\sqrt{V_{\text{robust}}})</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>coverage rate (\hat{V}_{\text{robust}})</td>
<td>0.96</td>
<td>0.48</td>
</tr>
<tr>
<td>(\sqrt{V_{\text{cluster}}})</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>coverage rate (\hat{V}_{\text{cluster}})</td>
<td>1.00</td>
<td>0.97</td>
</tr>
</tbody>
</table>
Causal Setting: Fuzzy Clustering

Suppose:

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

Assignment is correlated within clusters, but not perfectly correlated.

- \( \hat{V}_{\text{cluster}} \) cannot be right, because it is wrong if \( \gamma = 0 \)
- \( \hat{V}_{\text{robust}} \) 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

\[
\hat{\tau} = \frac{1}{M} \sum_{i=1}^{M} D_i \cdot Y_i
\]
Define

\[ \varepsilon_i(-1) = Y_i(-1) - \frac{1}{M} \sum_{i=1}^{M} Y_i(-1), \quad \varepsilon_i(1) = Y_i(1) - \frac{1}{M} \sum_{i=1}^{M} Y_i(1) \]

\[ \varepsilon_i = \frac{\varepsilon_i(-1) + \varepsilon_i(1)}{2} \]

\[ \varepsilon_i = \varepsilon_i(D_i) \]

\[ \hat{\tau} = \frac{Y'D}{N} = \tau + \frac{\varepsilon'D}{N} \]

So, true (infeasible) variance is

\[ \nabla(\hat{\tau}) = \varepsilon'\nabla(D)\varepsilon/N^2 \]

We need to figure out the exact variance of \( D \) and find an estimator for \( \varepsilon \). Note that \( \mathbb{E}[D] = 0 \), so just need second moments.
Elements of $\mathbb{V}(D) = \mathbb{E}[DD']$:

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

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

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

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

$$\hat{\mathbb{V}}(D)_{ij} = \begin{cases} 
1 & \text{if } i = j, \\
4 \cdot \delta^2 & \text{if } i \neq j, G_i = G_j, \\
0 & \text{otherwise.}
\end{cases}$$
We do not observe $\varepsilon_i$, but can estimate it:

$$\mathbb{E}[\varepsilon_i] = \varepsilon_i$$

So proposed feasible variance estimator is

$$\hat{\mathbb{V}}_{fc} = \hat{\varepsilon}' \hat{\mathbb{V}}(D) \hat{\varepsilon} / N^2$$  NEW VARIANCE ESTIMATOR

- If $\delta = 0$ (random assignment), then $\hat{\mathbb{V}}_{fc} = \hat{\mathbb{V}}_{\text{robust}}$

- If $\delta = 1/2$ (clustered assignment), then $\hat{\mathbb{V}}_{fc} = \hat{\mathbb{V}}_{\text{cluster}}$

- Can deal with intermediate cases.
**Simulation** random sampling, correlated assignment within clusters

\[
\frac{Y_i(-1) + Y_i(1)}{2} = \nu_{Gi} + \eta_i, \quad \nu_g \sim \mathcal{N}(0, 1), \quad \eta_i \sim \mathcal{N}(0, 1)
\]

\[
\tau_i = \frac{Y_i(1) - Y_i(-1)}{2} = \xi_{Gi} + \omega_i \sim, \quad \xi_g \sim \mathcal{N}(0, 1), \quad \omega_i \sim \mathcal{N}(0, 1)
\]

Three values for $\delta$:

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

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>std</th>
<th>se</th>
<th>size</th>
<th>se</th>
<th>size</th>
<th>se</th>
<th>size</th>
<th>se</th>
<th>size</th>
<th>se</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>.01</td>
<td>.01</td>
<td>.03</td>
<td>.02</td>
<td>.00</td>
<td>.05</td>
<td>.00</td>
<td>.03</td>
<td>.02</td>
<td>.01</td>
<td>.05</td>
</tr>
<tr>
<td>correlated</td>
<td>.05</td>
<td>.01</td>
<td>.62</td>
<td>.02</td>
<td>.51</td>
<td>.07</td>
<td>.01</td>
<td>.05</td>
<td>.04</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>clust</td>
<td>.10</td>
<td>.01</td>
<td>.81</td>
<td>.02</td>
<td>.74</td>
<td>.11</td>
<td>.03</td>
<td>.11</td>
<td>.03</td>
<td>.10</td>
<td>.05</td>
</tr>
</tbody>
</table>
# Summary

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

<table>
<thead>
<tr>
<th>Sampling</th>
<th>random</th>
<th>correlated</th>
<th>clustered</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>robust(fc)</td>
<td>??</td>
<td>cluster</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assignment</th>
<th>correlated</th>
<th>fc</th>
<th>??</th>
<th>cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>clustered</td>
<td>cluster(fc)</td>
<td>??</td>
<td></td>
<td>cluster</td>
</tr>
</tbody>
</table>
Conclusion

In most studies clustering is not about sampling, but about design of causal variables: observational studies equivalent of clustered randomized experiment.

Need to be clear about distribution of causal variables.