When should we adjust standard errors for clustering ? A discussion of Abadie et al. 2017

PSE Doctoral program:

Labor & public economics

Arthur Heim

October, 2nd 2019

Outline

Outline

Introduction

The Clusterjerk[1] of every seminar

- "Did you cluster your standard error ?"
- Yet, most of the time, it is not clear whether one should cluster or not and on which level of grouping.
- There is also a big confusion on the role of fixed effects to account for clustering.

Econometricians *Haiku* from Angrist and Pischke 2008, end of chapter one:

T-stats looks too good
Try cluster standard errors
significance gone.

1. From a debate on Chris Blattman's blog

Introduction

A simple example

- **Imagine** you wrote a *not-desk-rejected* paper estimating a Mincerian equation using Labor force survey (e.g. Enquête emploi in France):

$$Y_i = \alpha + \delta S_i + \gamma_1 e_i + \gamma_2 e_i^2 + \boldsymbol{X_i'}\boldsymbol{\beta} + \varepsilon_i$$

- You are considering whether you should cluster your SE.
- Referees strongly encourage you to do so:
  1. Referee 1 tells you "the wage residual is likely to be correlated within local labor markets, so you should cluster your standard errors by state or village."
  2. Referee 2 argues "The wage residual is likely to be correlated for people working in the same industry, so you should cluster your standard errors by industry"
  3. Referee 3 argues that "the wage residual is likely to be correlated by age cohort, so you should cluster your standard errors by cohort".
- **What should you do?**

Introduction

A slightly more sophisticated one

- You conduct a field experiment where first, a sample of 120 middle schools are randomly selected to participate in a teacher training program.
- Second, you randomly select the teachers (whichever school they belong to) who are to participate in the first year. The others represent a control group for the first year.
- Outcomes are test scores retrieved from national student assessments and concerns all students from the classrooms taughts by these teachers (let's assume that the students to teacher assignment is also fairly random)
- Should you cluster SE:
  1. Yes/no ?
  2. at the teacher level ?
  3. at the school level ?

Introduction

Answer from Abadie et al. 2017:

- Whether one should cluster (or not) **should not** be decided based on whether or not it changes something to the results.
- **Clustering will almost always matter**, even when there is no correlation between residuals within cluster and no correlation between regressors within cluster.
- Inspecting data is not sufficient to determine whether clustering adjustment is needed.

Introduction

Answer from Abadie et al. 2017:

- There are three rationals for clustering:
  1. **Sampling Design**: The sampling process consists in selecting **a small share** of clusters from a larger population of many more clusters.
     - French Labor force survey samples "grapes" of households
  2. **Experimental Design**: There exist a correlation between belonging to a certain cluster and the values of your variable of interest.
     - Clustered randomized control trials
  3. **Heterogenity** in treatment effect w.r.t clusters ;
     - Different cluster-specific-ATE

- Abadie et al. 2017 explain the situations when one should/shouldn't adjust w.r.t these rationals.

Outline

The textbook case

What is usually meant when one talks about clusters

- Most econometrics texbooks[2] approches the clustering issue as something close to omitted variable bias where, the initial model:

$$Y_{ic} = \alpha + \boldsymbol{X}'\beta + \mu_{ic}$$

actually hides the fact that the error term μ_c has a group structure s.t.:

$$\mu_{ic} = \upsilon_c + \varepsilon_{ic}$$

- And thus, estimating the model without accounting for that yields biased standard errors because $\mathbb{E}[\mu_{ic}\mu_{jc}] = \rho\sigma^2_\mu > 0$

♭ This presentation, although pedagogical, reinforce the confusion between fixed effect and clustering.

$$(Y_{ic} - \bar{Y}_c) = (\boldsymbol{X_{ic}} - \boldsymbol{\bar{X}_c})'\boldsymbol{\beta} + \mu_{ic} - \bar{\mu}_c$$

2. For instance Cameron and Trivedi 2005; Angrist and Pischke 2008; Wooldridge 2010; Wooldridge 2012

The textbook case

What is usually meant when one talks about clusters

- The second approach is usually through panel data and especially Dif in Dif issues.
- The very influential paper by Bertrand, Duflo, and Mullainathan 2004 (QJE) emphasizes the issue of serial correlation in DiD models such as the classic group-time fixed effect estimand:

$$Y_{ict} = \gamma_c + \lambda_t + \boldsymbol{X'}\boldsymbol{\beta} + \varepsilon_{ict}$$

- The problem is that individuals in a given group are likely to suffer from common shocks at some time t such that there is another component hiden in the error above:

$$\varepsilon = \upsilon_{ct} + \eta_{ict}$$

- If these group-time shocks are (assumed) independents, then the situation is closed to the one before and one could cluster by group-time.
- Yet, this is often not true (e.g. if groups are states or region, a bad situation one period is likely to be bad too the next period)

The textbook case

The group structure problem

- Heteroskedasticity robust standard errors assume that the $(N \times N)$ matrix $\mathbb{E}\left[\varepsilon\varepsilon'|\boldsymbol{X}\right]$ is diagonal, meaning there is no correlation between errors accross observations. Memo
- This assumption is false in many settings among which:
  - Non-stationary time series or panel data
  - Identical values of one or more regressors for groups of individuals = clusters
  - ...
- From a setting where potentially all errors are correlated together, we cannot use the estimated residuals as in the robust SE (White 1980) (because $\sum \hat{X}_i \hat{\epsilon}_i = 0$ by construction)
- Hence, one has to allow correlation up to a certain point: in time (Newey and West 1987), or among members of a group (Kloek 1981; Moulton 1986)

The textbook case

The group structure problem

- Assuming homoskedasticity:

$$\mathbb{E}\left[\varepsilon\varepsilon|\boldsymbol{X}\right] \equiv \Omega_{ij} = \begin{cases} 0 & \text{if } C_i \neq C_j \\ \rho\sigma^2 & \text{if } C_i = C_j, \ i \neq j \\ \sigma^2 & \text{if } i = j \end{cases}$$

- Suppose just 2 groups, this matrix looks something like:

$$\Omega_{ij} = \begin{pmatrix} \sigma^2_{(1,1)1} & \cdots & \rho\sigma^2_{(1,n_1)1} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2_{(n_1,1)1} & \cdots & \sigma^2_{(n_1,n_1)1} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \sigma^2_{(n_1+1,n_1+1)2} & \cdots & \rho\sigma^2_{(n_1+1,N)2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \rho\sigma^2_{(N,1)2} & \cdots & \sigma^2_{(N,N)2} \end{pmatrix}$$

The textbook case

Assuming homoskedasticity & group size

- Assuming homoskedasticity & same group size:

$$\mathbb{V}_{kloek}(\hat{\beta}|\boldsymbol{X}) = \mathbb{V}_{OLS} \times \left(1 + \rho_{\varepsilon}\rho_{X}\frac{N}{C}\right) \quad (1)$$

- Where ρ_{ε} is the within cluster correlation of the errors
- Where ρ_{X} is the within cluster correlation of the regressors

Relaxing homoskedasticity

- The cluster adjustment by Liang and Zeger 1986 used in most statistical packages:

$$\mathbb{V}_{LZ}(\hat{\beta}|\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{c=1}^{C}\boldsymbol{X}_c'\Omega_c\boldsymbol{X}_c\right)(\boldsymbol{X}'\boldsymbol{X})^{-1} \quad (2)$$

The textbook case

Estimated versions

- The estimated version of the so called robust (EHW) variance is:

$$\hat{\mathbb{V}}_{EWH}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X'X})^{-1}\left(\sum_{i=1}^{N}(\boldsymbol{Y_i} - \hat{\beta}'\boldsymbol{X_i})^2 \boldsymbol{X_i X_i'}\right)(\boldsymbol{X'X})^{-1} \tag{3}$$

- The estimated version of the cluster robust (LZ) variance is:

$$\hat{\mathbb{V}}_{LZ}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X'X})^{-1}\left(\sum_{c=1}^{C}\Big(\sum_{i:C_i=c}\underbrace{(\boldsymbol{Y_i} - \hat{\beta}'\boldsymbol{X_i})\boldsymbol{X_i}'}_{\hat{\varepsilon}X_i}\Big)\right.$$

$$\left.\Big(\sum_{i:C_i=c}\underbrace{(\boldsymbol{Y_i} - \hat{\beta}'\boldsymbol{X_i})\boldsymbol{X_i}}_{\hat{\varepsilon}X_i}\Big)'\right)(\boldsymbol{X'X})^{-1} \tag{4}$$

- These are the main estimators used by applied researchers between which one has to choose.

The almost forgotten reason for clustering

"How were your data collected ?"

- "Textbook cases" discussed before are what one may call "model-based" cases for clustering
- These examples implicitly assume that data are collected randomly, or randomly enough.
- However, surveys often use more sophisticated sampling methods with nested structures (e.g. sampling cities, then neighborhoods, then households), stratification and/or weightings.

The first clustering issue should be survey design effect
$\Rightarrow$ Clustering at the primary survey unit (PSU) at the minimum.

Conventional wisdom about standard errors

When to cluster according to Colin Cameron and Miller 2015

- Equation (1) while restrictive shows that the inflation factor increases in:
  - The within-cluster correlation of the regressors ρ_X
  - The within-cluster correlation of the error ρ_ϵ
  - The number of observations in each cluster
- Consequently one could think clustering does not change a thing if either $\rho_X = 0$ or $\rho_\epsilon = 0$
- It has been shown by Moulton 1990 that the inflation factor can be large despite very small correlation.
- Colin Cameron and Miller 2015 basically say that whenever there is a reason to believe that there is some correlation within some groups, one should cluster.
- "The consensus is to be conservative and avoid bias and to use bigger and more aggregate clusters when possible". (p. 333)

Conventional wisdom about standard errors

When to cluster according to Colin Cameron and Miller 2015

> *"There are settings where one may not need to use cluster-robust standard errors. We outline several though note that in all these cases it is always possible to still obtain cluster-robust standard errors and contrast them to default standard errors. If there is an appreciable difference, then use cluster robust standard errors". (p.334)*

Outline

Clustering matters, yes, so what ?

One misconception according to Abadie et al. 2017

- Because of the formula (1), it is often thought that clustering "does not matter" if either $\rho_X = 0$ or $\rho_\varepsilon = 0$
- Thus, adjusting for cluster wouldn't change a thing in situation such as:
  - Individual randomized control trials
  - Adding cluster fixed effects to the regression
- Using simulated data, they show that clustering **does** affect estimated standard errors in this setting.

Clustering matters, yes, so what ?

Example data (not knowing the DGP)

- Sample: N= 100 323 with 100 clusters and ≈ 1 000 units per cluster.
- Observe Y_{ic}, $W_{ic} = \mathbb{1}(\text{treated})$, C_{ic}
- Estimate linear regression $Y_i = \alpha + \tau W_i + \epsilon_i$ by OLS.

First result

$$\hat{\rho}_{\hat{\varepsilon}} = 0.001 \quad \hat{\rho}_{\hat{W}} = 0.001$$

- Correlations are essentially 0, hence, the correction should not have an impact.

Clustering matters, yes, so what ?

Cluster adjustment matters (1)

$$\hat{\tau}_{LS} = -0.120 \quad \hat{SE}_{EHW} = 0.004 \quad \hat{SE}_{LZ} = 0.100$$

- Adjusting for cluster matters a lot. $\Rightarrow$ Inspecting within cluster correlation is not enough to determine whether adjusting SE would matter.
- Indeed, the LZ adjustment relies on something else:

$$\hat{\mathbb{V}}_{LZ}(\hat{\beta}) = (\boldsymbol{X'X})^{-1}\Big(\sum_{c=1}^{C}\big(\sum_{i:C_i=c}\underbrace{(\boldsymbol{Y_i} - \hat{\beta}'\boldsymbol{X_i})\boldsymbol{X_i}'}_{\hat{\varepsilon}X_i}\big)\big(\sum_{i:C_i=c}\underbrace{(\boldsymbol{Y_i} - \hat{\beta}'\boldsymbol{X_i})\boldsymbol{X_i}}_{\hat{\varepsilon}X_i}{}'\big)'\Big)(\boldsymbol{X'X})^{-1}$$

- What matters for the adjustment is the **within-cluster correlation** of the product of the residuals and the regressors.
- Here, $\rho_{\hat{\varepsilon}W} = 0.500$

Clustering matters, yes, so what ?

Cluster adjustment matters (2)

Estimating the fixed effect model: $Y_i = \tau W_i + \sum_{c=1}^{C} \alpha_c C_{ic} + \varepsilon_i$

$$\hat{\tau}_{FE} = -0.120 \quad \hat{SE}_{EHW} = 0.003 \quad \hat{SE}_{LZ} = 0.243$$

- Adding fixed effect did not change the point estimate, but increased precision (as one would expect) of the EHW robust SE.
- Clustering however matters a lot here too $\Rightarrow$ Adding fixed effect does not necessary *fix* the clustering issue.

So it's not because you can cluster (and it matters) that you should cluster

If we were to follow Colin Cameron and Miller 2015

- We would cluster everything in the previous example.
- Abadie et al. 2017 disagree and illustrate with another example

Data generating process

- General population of 10 million units, 100 clusters of 10 000 units in each.
- Here, W_i is assigned at random with probability p=1/2.
- Treatment effect is heterogenous w.r.t. clusters such that:

$$\tau_c = \begin{cases} -1 & \text{for half of the clusters} \\ 1 & \text{for the other half} \end{cases}$$

- Error term $\sim \mathcal{N}0, 1)$ and ATE=0.

So it's not because you can cluster (and it matters) that you should cluster

Which SE is correct ?

- Draw random samples 10 000 times with sampling probability = 1 % and estimate the models.

Table: Standard errors and coverage rates of random samplings

Simple OLS				Fixed effect			
EHW Variance		LZ Variance		EHW Variance		LZ Variance	
(SE)	cov rate	(SE)	cov rate	(SE)	cov rate	(SE)	cov rate
0.007	0.950	0.051	1.000	0.007	0.950	0.131	0.986

- The correct standard error is EHW as it rejects the appropriate proportion of type 1 error.

So it's not because you can cluster (and it matters) that you should cluster

Why the differences ? How to choose ?

- Given random assignment, both errors are corrects but for different estimands.
- EHW assumes that the population is randomly selected from the relevant population (which is the case here)
- LZ assumes that the clusters here are a sample of more clusters in the main population.
- This assumption is often implicit in the textbook cases but has important consequences.
- More obvious in the sampling design literature (e.g. French Labor force survey)

$\Rightarrow$ One cannot tell from the data itself whether other clusters exist in the full population.

Outline

Conceptual framework

Sequence of population

- Sequence of populations defined by M_n units and C_n clusters ; M_n is strictly increasing and C_n is weakly increasing in n.
- Rubin's causal framework with 2 potential outcomes for each individual: $Y_{in}(1)$; $Y_{in}(0)$.
- 2 treatment specific errors:

$$\varepsilon_{in}(W) = Y_{in}(W) - \frac{1}{n}\sum_{j=1}^{n} Y_{jn}(W) \text{ for } W = 0, 1.$$

- Main interest lies in the n-population's average treatment effect:

$$\tau_n = \frac{1}{M_n}\sum_{i=1}^{n}\Big(Y_{in}(1) - Y_{in}(0)\Big) = \bar{Y}_n(1) - \bar{Y}_n(0) \qquad (5)$$

Sampling and assignment

Sampling process

- Define a variable $R_{in} = \mathbb{1}(\text{sampled})$ such that we observe the triplet (Y_{in}, W_{in}, C_{in}) if $R_{in} = 1$ and nothing otherwise.
- For a population M_n, we observe a sample of size $N = \sum_{i=1}^{M_n} R_{in}$.
- $R_{in} \perp Y_{in}(1) \ ; \ Y_{in}(0)$
- 2 stages design:
  - Clusters are sampled with probability P_{C_n}
  - Individuals are sampled in the selected clusters with P_{U_n}
- Probability of person *in* being sampled is $P_{C_n} P_{U_n}$.
- Both probability may be equal to 1, or close to 0:

	$P_{C_n} = 1$	$P_{C_n} \approx 0$
$P_{U_n} = 1$	full population	sample everyone in few clusters
$P_{U_n} = 0$	random sample	few units from few clusters

Sampling and assignment

Treatment assignment

- Treatment assignment is also a 2 stages process:
- **First Stage** For cluster c in population n, an assignment probability is drawn randomly from a distrubtion $f(\cdot)$ with mean $\mu_n = \frac{1}{2}$ and variance $\sigma_n^2 \geq 0$.
  - If $\sigma_n^2 = 0$, we have pure random assignment.
  - If $\sigma_n^2 > 0$, we have correlated assignment **within the cluster**.
  - Special case: $\sigma_n^2 = \frac{1}{4}$ then $q_{cn} \in \{0, 1\}$ all units within a cluster have identical assignments.
- **Second stage:** each individual within a cluster c is assigned to treatment independently with cluster-specific probability q_{cn}
- **Translation**: If $\sigma_n^2 > 0$, individuals from a cluster are all either more likely or less likely to be treated than average. Thus, there is a correlation between treatment assignment and being in a cluster.

Estimators

OLS Estimator of the treatment effect

- The least square estimator of τ_n is:

$$\hat{\tau} = \frac{\sum_{i=1}^{n} R_{in}(W_{in} - \bar{W}_n) Y_{in}}{\sum_{i=1}^{n} R_{in}(W_{in} - \bar{W}_n)^2} = \bar{Y}_n(1) - \bar{Y}_n(2)$$

- What matters is the estimation of the variance of $\hat{\tau}$
- By definition, the true variance is $\sqrt{N_n}(\hat{\tau} - \tau_n)$
- Using large sample properties, the authors show:

$$\sqrt{N_n}(\hat{\tau} - \tau_n) - \underbrace{\frac{2}{\sqrt{M_n P_{C_n} P_{U_n}}} \sum_{i=1}^{M_n} R_{in}(2W_{in} - 1)\varepsilon_{in}}_{\text{linear approximation}} = o_p(1) \tag{6}$$

Estimators

Properties of the linear approximation of the variance

$$\eta_n = \frac{2}{\sqrt{M_n P_{C_n} P_{U_n}}} \sum_{i=1}^{M_n} \eta_{in} \text{ With } \eta_{in} = R_{in}(2W_{in} - 1)\varepsilon_{in} \quad (7)$$

- They calculate the exact variance of η_n for various values of the parameters and the corresponding EHW and LZ estimator.

Estimators

Proposition 1-i

$$\mathbb{V}[\eta_n] = \frac{1}{M_n}\sum_{i=1}^{M_n}\left[2\left(\varepsilon_{in}(1)^2 + \varepsilon_{in}(0)^2\right) - P_{U_n}(\varepsilon_{in}(1) - \varepsilon_{in}(0))^2 + 4P_{U_n}\sigma_n^2(\varepsilon_{in}(1) - \varepsilon_{in}(0))^2\right]$$
$$+ \frac{P_{U_n}}{M_n}\sum_{c=1}^{C_n} M_{cn}^2\left[(1 - P_{C_n})(\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0))^2 + 4\sigma_n^2(\bar{\varepsilon}_{cn}(1) + \bar{\varepsilon}_{cn}(0))^2\right] \quad (8)$$

- The first sum in this formula is approximately the EHW Variance.
- if $P_{U_n} \approx 0$, the firt term simplifies to $\mathbb{V}_{EHW} = \sum_{i=1}^{N}\left(\frac{\varepsilon_{in}(1)^2 + \varepsilon_{in}(0)^2}{M_n}\right)$
- The second term captures the effects of clustered sampling and assignment on variance.

Estimators

Proposition 1-i

$$\mathbb{V}[\eta_n] = \frac{1}{M_n}\sum_{i=1}^{M_n}\Big[2\left(\varepsilon_{in}(1)^2 + \varepsilon_{in}(0)^2\right) - P_{U_n}(\varepsilon_{in}(1) - \varepsilon_{in}(0))^2 + 4P_{U_n}\sigma_n^2(\varepsilon_{in}(1) - \varepsilon_{in}(0))^2\Big]$$
$$+ \frac{P_{U_n}}{M_n}\sum_{c=1}^{C_n} M_{cn}^2\Big[\underbrace{(1 - P_{C_n})}_{\text{sampling}}(\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0))^2 + 4\underbrace{\sigma_n^2}_{\text{assignment}}(\bar{\varepsilon}_{cn}(1) + \bar{\varepsilon}_{cn}(0))^2\Big] \quad (9)$$

- First part of the second sum disapears if $P_{C_n} = 1$, that is, if we have all clusters in the sample (e.g. in a pure random assignment)
- Second part of the second sum disapears if $\sigma_n^2 = 0$ if there is no correlation between assignment to treatment and clustering.

Estimators

Proposition 1-ii

- The difference between the correct variance and the limit of the normalized LZ variance estimator is:

$$\mathbb{V}_{LZ} - \mathbb{V}[\eta_n] = \frac{P_{C_u} P_{U_n}}{M_n} \sum_{c=1}^{C_n} M_{cn}^2 \big(\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0)\big)^2 \geq 0 \tag{10}$$

- LZ variance captures correctly the component due to cluster assignment but performs poorly for the clustering due to sampling design unless $P_{C_n} \approx 0$
- Due to the assumption that the sampled cluster are a small proportion of the population of clusters which explain why the LZ estimator and the true variance are proportional to P_{C_n}.

Estimators

Proposition 1-iii

- The difference between the limit of the normalized LZ and the EHW variance estimator is:

$$\mathbb{V}_{LZ} - \mathbb{V}_{EHW} = \frac{-2P_{U_n}}{M_n} \sum_{i=1}^{M_n} \left[\Big(\varepsilon_{in}(1) - \varepsilon_{in}(0) \Big)^2 + 4\sigma_n^2 \Big(\varepsilon_{in}(1) + \varepsilon_{in}(0) \Big)^2 \right]$$
$$+ \frac{P_{U_n}}{M_n} \sum_{c=1}^{C_n} M_{cn}^2 \left[\big(\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0) \big)^2 + 4\sigma_n^2 \big(\bar{\varepsilon}_{cn}(1) + \bar{\varepsilon}_{cn}(0) \big)^2 \right] \quad (11)$$

- This part show when adjusting with LZ makes a difference with EHW
- First sum is small relative to the second part if there is a large number of unit per cluster relative to the number of cluster.
- If the number of unit per cluster is constant (M_n/C_n) and large compared to the number of clusters, the second sum is proportional to $\frac{M_n}{C_n^2}$ and large relative to the first sum.

⋆ This is how the generated data in the 1st example.

Estimators

Proposition 1-iii

- In the case where the number of individuals in each cluster is large relative to the number of clusters, the clustering matters if there is **heterogeneity of treatment accross clusters** or if there is cluster assignment.
- This comes from the fact that $\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0) = \tau_{cn} - \tau_n$

  Proof
- Looking at the second sum only:

$$\frac{P_{U_n}}{M_n}\sum_{c=1}^{C_n} M_{cn}^2\Big[\underbrace{\big(\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0)\big)^2}_{\text{Heterogeneity}} + \underbrace{4\sigma_n^2}_{\text{assignment}} \big(\bar{\varepsilon}_{cn}(1)+\bar{\varepsilon}_{cn}(0)\big)^2$$

Estimators

Corollary 1: When we don't need to cluster

- There is no need for clustering in two situations:
  1. There is no clustering in the sampling ($P_{C_n} = 1 \quad \forall\ n$) **and** there is no clustering in the assignment ($\sigma_n^2 = 0$)
  2. There is no heterogenity of treatment ($Y_{in}(1) - Y_{in}(0) = \tau \quad \forall\ i$) **and** there is no clustering assignment ($\sigma^2 = 0$)
- Corollary 1 is a special case of Proposition 1-i.

Estimators

Corollary 2: When LZ correction is correct

- One can use LZ variance estimation to adjust clustering if:
  1. There is no heterogenity of treatment ($Y_{in}(1) - Y_{in}(0) = \tau \quad \forall\ i$)
  2. ($P_{C_n} \approx 0 \quad \forall\ n$) i.e. We only observe few clusters from the total population.
  3. P_{U_n} is close to 0 so that there is at most one sampled unit per cluster (in which case clustering adjustment do not matter but the PSU is a level higher)
- Corollary 2 emerges from P1-ii with important restrictions.
  - 1) is not likely to hold in general
  - 2) cannot be assessed using the actual data. One has to know the sampling conditions.
  - If one concludes that all clusters are included, then LZ is in general too conservative.

Estimators

Clever idea: Using heterogeneity

- In a situation where all clusters are included, LZ is too conservative
- If the assignment is perfectly correlated within the cluster, there is nothing much to do.
- If there is variation in the treatment within clusters, one can estimate $\mathbb{V}_{LZ} - \mathbb{V}[\eta_n]$ and substract that from $\mathbb{V}_{LZ}$ using again that $\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0) = \tau_{cn} - \tau_n$.
- The proposed cluster-adjusted variance estimator is then:

$$\hat{\mathbb{V}}_{CA}(\hat{\tau}) = \hat{\mathbb{V}_{LZ}}(\hat{\tau}) - \frac{1}{N^2}\sum_{c=1}^{C} N_c^2(\hat{\tau}_{cn} - \hat{\tau}_n) \tag{12}$$

Outline

Conclusions

- Adjusting SE for clustering effect is often misunderstood
- Usual recommandations are often too conservatives
- We should cluster:
  - In the presence of heterogenous treatment effect and small number of clusters compared to the overall population
  - when there is correlation between treatment and clusters (cluster assignment)
- We should not cluster:
  - In pure randomized control trial (or any situation without sampling clustering or assignment clustering)
  - when there is constant treatment effect and no clustering in the assignment.

♭ Convincing model but specific to the stated configurations.

♭ Less usefull for less RCT-like designs (e.g. the infamous serial correlation in DID)

▸ Back to intro examples

Bibliography I

Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge. 2017. *When Should You Adjust Standard Errors for Clustering?* Working paper. October 8. http://arxiv.org/abs/1710.02926.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics* 119 (1): 249–275.

Cameron, A Colin, and Pravin K Trivedi. 2005. *Microeconometrics : Methods and Applications.* Cambridge University Press.

Colin Cameron, A., and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50 (2): 317–372. ISSN: 0022-166X, 1548-8004. doi:10.3368/jhr.50.2.317. http://jhr.uwpress.org/lookup/doi/10.3368/jhr.50.2.317.

Kloek, Tuenis. 1981. "OLS Estimation in a Model Where a Microvariable Is Explained by Aggregates and Contemporaneous Disturbances Are Equicorrelated." *Econometrica* 49 (1): 205–207.

Liang, Kung-Yee, and Scott L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73 (1): 13–22.

Bibliography II

Moulton, Brent. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32 (3): 385–397.

Moulton, Brent R. 1990. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units Vol. 72, No. 2 (May, 1990), Pp. 334-338." *The review of Economics and Statistics* 72 (2): 334–338.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55, no. 3 (May): 703. ISSN: 00129682. doi:10.2307/1913610. https://www.jstor.org/stable/1913610?origin=crossref.

White, Albert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Discret Test for Heteroskedasticiy." *Econometrica* 48 (4): 817–838.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* MIT Press.

Wooldridge, Jeffrey M. 2012. "Introductory Econometrics: A Modern Approach": 910.

Outline

Outline

Errors and residuals

We sometimes get confused...

- **Errors** are the vertical distances between observations and the unknown Conditional Expectation Function (CEF). Therefore, they are **unknown**.
- **Residuals** are the vertical distances between observations and the estimated regression function. Therefore, they are **known**.
- Errors comme from the CEF decomposition property[3]:

$$Y_i = \mathbb{E}[Y_i|\boldsymbol{X_i}] + \varepsilon_i$$

where ε_i is mean independent of $\boldsymbol{X_i}$ and is therefore uncorrelated with any function of $\boldsymbol{X_i}$

3. Angrist and Pischke 2008, Theorem 3.1.1

Errors and residuals

We sometime get confused...

- Errors represent the difference between the outcome and the true conditional mean.
- In matrix notation:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$
$$\varepsilon = \boldsymbol{Y} - \boldsymbol{X}\beta$$

- Residuals represent the difference between the outcome and the estimated average.
- In matrix notation:

$$\boldsymbol{Y} = \boldsymbol{X}\hat{\beta} + \hat{\epsilon}$$
$$\hat{\epsilon} = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta}$$

Estimations depend on error !

OLS Estimand as seen in class

$$\begin{aligned}\hat{\beta}_{OLS} &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'X}\beta + \varepsilon \\ &= \boldsymbol{\beta} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\varepsilon\end{aligned}$$

$\hat{\beta}_{OLS}$ is known to be unbiased but its variance **depends on the unknown error**.

Estimations depend on error !

Variance of $\hat{\beta}$ as seen in class

$$\begin{aligned}
\mathbb{V}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] &= \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'|\boldsymbol{X}] \\
&= \mathbb{E}\Big[[\boldsymbol{X}'\boldsymbol{X}]^{-1}[\boldsymbol{X}'\varepsilon]([\boldsymbol{X}'\boldsymbol{X}]^{-1}[\boldsymbol{X}'\varepsilon])'|\boldsymbol{X}\Big] \\
&= \mathbb{E}\Big[[\boldsymbol{X}'\boldsymbol{X}]^{-1}\boldsymbol{X}'\varepsilon\varepsilon'\boldsymbol{X}[\boldsymbol{X}'\boldsymbol{X}]^{-1}|\boldsymbol{X}\Big]
\end{aligned}$$

Which give us the variance covariance matrix of the betas:

$$\mathbb{V}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = [\boldsymbol{X}'\boldsymbol{X}]^{-1}\mathbb{E}\Big[\boldsymbol{X}'\varepsilon\varepsilon'\boldsymbol{X}|\boldsymbol{X}\Big][\boldsymbol{X}'\boldsymbol{X}]^{-1} \tag{13}$$

Back

Estimations depend on error !

What does this matrix looks like

Without clusters:

$$\mathbb{V}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = \begin{pmatrix} \mathbb{V}[\hat{\beta}_0|\boldsymbol{X}] & cov(\hat{\beta}_0, \hat{\beta}_1|\boldsymbol{X}) & \cdots & cov(\hat{\beta}_0, \hat{\beta}_P|\boldsymbol{X}) \\ cov(\hat{\beta}_1, \hat{\beta}_0|\boldsymbol{X}) & \mathbb{V}[\hat{\beta}_1|\boldsymbol{X}] & \cdots & cov(\hat{\beta}_1, \hat{\beta}_P|\boldsymbol{X}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\hat{\beta}_P, \hat{\beta}_0|\boldsymbol{X}) & cov(\hat{\beta}_P, \hat{\beta}_1|\boldsymbol{X}) & \cdots & \mathbb{V}[\hat{\beta}_P|\boldsymbol{X}] \end{pmatrix}$$

This matrix is not identified and we need either some extra assumptions such as homoskedasticity and/or no serial correlation to estimate it.

Estimating the variance-covariance matrix of $\hat{\beta}$

Under homoskedasticity and no serial correlation

If we assume that the correlation between errors is null and that the errors' variance is constant, that is:

$$\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}\right] = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \boldsymbol{\sigma^2 I_{[N.N]}}$$

Then the variance-covariance matrix of betas simplifies a lot:

$$\begin{aligned} \mathbb{V}_{homo}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] &= [\boldsymbol{X}'\boldsymbol{X}]^{-1}\boldsymbol{\sigma}^2 I[\boldsymbol{X}'\boldsymbol{X}]^{-1} \\ &= \boldsymbol{\sigma}^2[\boldsymbol{X}'\boldsymbol{X}]^{-1}[\boldsymbol{X}'\boldsymbol{X}]^{-1} \\ &= \boldsymbol{\sigma}^2[\boldsymbol{X}'\boldsymbol{X}]^{-1} \end{aligned}$$

The estimated variance of the error term:

$$\hat{\sigma^2} = \frac{1}{n-p}\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}' = \frac{\sum_{i=1}^{N} \hat{\epsilon}_i^2}{N-p}$$

Estimating the variance-covariance matrix of $\hat{\beta}$

Allowing Heteroskedasticity

If we assume that the correlation between errors is null but that the errors' variance is heterogenous, that is:

$$\mathbb{E}\left[\varepsilon\varepsilon|\boldsymbol{X}\right] = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

There is now n different variances and the variance of the coefficient simplifies:

$$\begin{aligned} \mathbb{V}_{EHW}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\mathbb{E}[diag(\sigma_i^2)|\boldsymbol{X}]\hat{\beta}|\boldsymbol{XX}(\boldsymbol{X'X})^{-1} \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\frac{1}{n}\sum_{i=1}^{n}\left(\sigma_i^2\boldsymbol{X_iX_i'}\right)(\boldsymbol{X'X})^{-1} \\ &\equiv (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\frac{1}{n}\sum_{i=1}^{n}\left(\Omega_{ii}\boldsymbol{X_iX_i'}\right)(\boldsymbol{X'X})^{-1} \end{aligned}$$

Estimating the variance-covariance matrix of $\hat{\beta}$

Allowing Heteroskedasticity

$$\mathbb{V}_{EHW}(\hat{\beta}|\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\frac{1}{n}\sum_{i=1}^{n}\left(\Omega_{ii}\boldsymbol{X}_i\boldsymbol{X}_i'\right)(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{14}$$

$$\tag{15}$$

The estimated version is:

$$\mathbb{V}_{EHW}(\hat{\beta}|\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\frac{1}{n}\sum_{i=1}^{n}\Big(\underbrace{(Y_i - \hat{\beta}'X_i)^2}_{=\hat{\epsilon_i^2}}\boldsymbol{X}_i\boldsymbol{X}_i'\Big)(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Outline

Proof that $\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0) = \tau_{cn} - \tau_n$

Start with writing the difference and substitute with the expressions page 8:

$$
\begin{aligned}
\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0) &= \frac{1}{M_{cn}}\Big(\sum_{i\in c_{in}=c} C_{inc}\varepsilon_{in}(1) - C_{inc}\varepsilon_{in}(0)\Big) \\
&= \frac{1}{M_{cn}}\Big[\sum_{i\in c_{in}=c} C_{inc}\Big(Y_{in}(1) - \frac{1}{n}\sum_{j=1}^{n}(Y_{jn}(1)) - Y_{in}(0) + \frac{1}{n}\sum_{j=1}^{n}(Y_{jn}(0))\Big)\Big] \\
&= \frac{1}{M_{cn}}\Big[\sum_{i\in c_{in}=c} C_{inc}\Big(Y_{in}(1) - Y_{in}(0)\Big) - \frac{C_{inc}}{n}\Big(\sum_{j=1}^{n}(Y_{jn}(1) - Y_{jn}(0)\Big)\Big]
\end{aligned}
$$

Since $C_{inc} = \mathbb{1}(c_{in} = c)$, then

$$
\frac{1}{M_{cn}}\sum_{i\in c_{in}=c} C_{inc}\Big(Y_{in}(1) - Y_{in}(0)\Big) = \frac{1}{M_{cn}}\sum_{i\in c_{in}=c} 1\Big(Y_{in}(1) - Y_{in}(0)\Big) \equiv \tau_{cn}
$$

$$
\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0) = \tau_{cn} - \frac{1}{M_{cn}}\sum_{i\in c_{in}=c}\frac{C_{inc}}{n}\Big(\sum_{j=1}^{n}(Y_{jn}(1) - Y_{jn}(0)\Big)
$$

Proof that $\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0) = \tau_{cn} - \tau_n$

On the right hand side of the sum, the only thing that depends on c is C_{inc}. Thus,

$$\sum_{i \in c_{in}=c} C_{inc} = M_{cn} = n = \sum_{c=1}^{C_n} M_c n$$

$$\begin{aligned}
\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0) &= \tau_{cn} - \frac{M_{cn}}{M_{cn}} \frac{1}{n} \Big(\sum_{j=1}^{n} (Y_{jn}(1) - Y_{jn}(0) \Big) \\
&= \tau_{cn} - \underbrace{\frac{1}{M_n} \Big(\sum_{j=1}^{n} (Y_{jn}(1) - Y_{jn}(0) \Big)}_{=\tau_n} \\
&= \tau_{cn} - \tau_n && (Q.E.D)
\end{aligned}$$

Back