# Clustering Standard Errors at the Session Level

Duk Gyoo Kim

Sungkyunkwan University (SKKU)

July, 2023

# A typical lab experiment

- **Control vs. treatment** just like a medical experiment.
- **Session**: Ppl participated in the experiment at the same time.
- **Between-subject design**: A subject participated in one and only session.
- **No mixture of control and treatment at a session**: The instructions of the experiment must be the same for everyone.
- **Strategic interactions**: A subject is randomly grouped (or paired) with some of the other subjects and plays a game.
- **Repetitions: Subjects play the same game repeatedly.**
- **Random rematch or perfect-stranger match**: Subjects are reshuffled so that each game is regarded as a static game.
- **Anonymity**: Subjects don't have face-to-face interactions.

# Data structure

- A set of observations from an individual is a proper subset of the whole observations from a session.
- A set of observations from a session is a proper subset of the whole observations from the same treatment.
- Adding individual or session fixed effects to the regression does not help to examine a treatment effect due to multicollinearity.

# Questions

Q: When should we cluster standard errors?
Q: When should we cluster standard errors at the "session" level?

Today I will

- persuade that the session-level clusters should be used only in particular situations with proper justification,
- recap that session randomization or counterbalancing cannot be a justification for session-level clustering, and
- point some practical issues regarding cluster-robust standard errors for the laboratory data.

# Motivation: I am not the only one who got

comments that the SEs should be clustered at the session level,

- because of session-level randomization.

  Keith Marzilli Ericson: *"You appear to have randomized at the session level(*), and hence your inference should cluster standard errors at the session level."*

  (*He meant by session randomization a random ordering of the control and treatment sessions.)

- because of the concerns for the 'static' session effects.

  Frechette (2012, ExpEcon) : *"One easily implemented solution [to treat the session effects] in many situations is to use clustering at the session level."*

  (Note that he assumes one observation per subject.)

Those concerns are valid, but the remedy is on weak ground.

# Reasoning for clustering

We should cluster SEs at the session level **when the experiment is designed to render strong correlation among the errors within a session.**

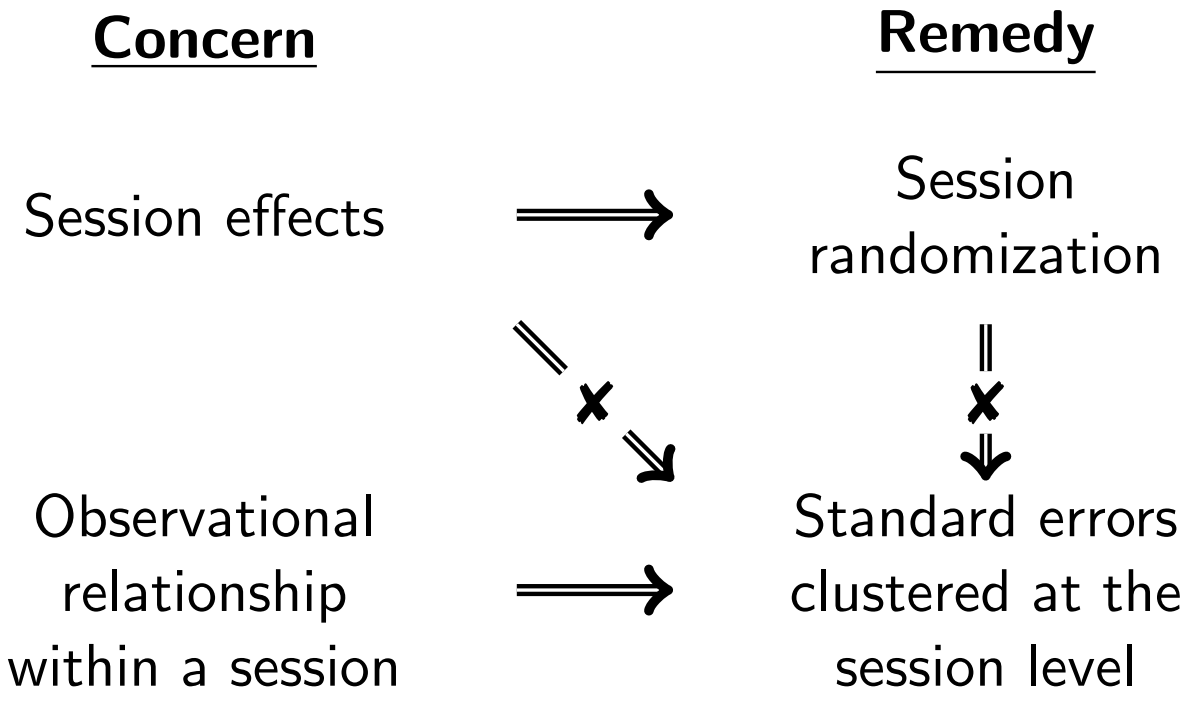



Figure 1: Clustering at the session level is not a remedy for session effects.

# To minimize destructive discussions,

I am not claiming that

- there are no session effects. Session randomization/counterbalancing is crucial.
- clustering SEs at the session level is futile. It is certainly helpful for some cases. I rather discuss its justification.

# Why cluster?

Clustered standard errors should be considered when error terms within a cluster are correlated to each other.

- In words, if the observations within a cluster are similar, then the residuals within a cluster will be more correlated than those of the whole observations.
- Thus, without "penalizing" the observational similarity, we will have downward-biased standard errors, which may lead to false-positive treatment effects more often.

# Cluster-robust inference 101

Suppose

- only one treatment and one control
- the session effects are already controlled properly
- a researcher examines a mean treatment effect
- an econometrician has $N = (S + S) \times I \times R$ observations in total, where $S$ is #controlled and treated sessions, $I$ is #subjects per session, and $R$ is #repetitions of the game.
- wlog, observations are lexicographically ordered.
  - 1=1st session, 1st subject, 1st observation
  - 2=1st session, 1st subject, 2nd observation...
  - R+1=1st session, 2nd subject, 1st observation...

# Cluster-robust inference 101

- Subtract the control-group mean so that the dependent variable as the deviation from the control-group mean.
- The treatment effect is captured by $\beta$ in

$$y_i = \beta T_i + \varepsilon_i,$$

  where $E[\varepsilon_i] = 0$. $T_i$ is a treatment dummy.
- With a slight abuse of notation, $T$ is a set of treated observations such that for $i \in T$, $T_i = 1$.
- The OLS estimator is

$$\hat{\beta} = \frac{\sum_i T_i y_i}{\sum_i T_i^2} = \frac{\sum_{i \in T} y_i}{SIR},$$

- and the variance of the estimator is

$$V[\hat{\beta}] = E[(\hat{\beta} - \beta)^2] = \frac{V[\sum_{i \in T} \varepsilon_i]}{S^2 I^2 R^2}$$

- Our main interest is on $V[\sum_{i \in T} \varepsilon_i]$.

# Cluster-robust inference 101

- $V[\sum_{i\in T}\varepsilon_i] = \sum_{i\in T}\sum_{j\in T} Cov[\varepsilon_i,\varepsilon_j] = \sum_{i\in T}\sum_{j\in T} E[\varepsilon_i\varepsilon_j]$
- If errors are uncorrelated, that is, $E[\varepsilon_i\varepsilon_j] = 0$ for $i \neq j$, then $\sum_{i\in T}\sum_{j\in T} E[\varepsilon_i\varepsilon_j] = \sum_{i\in T} E[\varepsilon_i^2]$. Heteroskedasticity-robust SE would be sufficient: $\sum_{i\in T} u_i^2$, where $u_i = y_i - \hat{\beta}\, T_i$.
- Let $C_i$ be the cluster that $i$ belongs to. If $E[\varepsilon_i\varepsilon_j] \neq 0$ for $i$ and $j \in C_i$, then

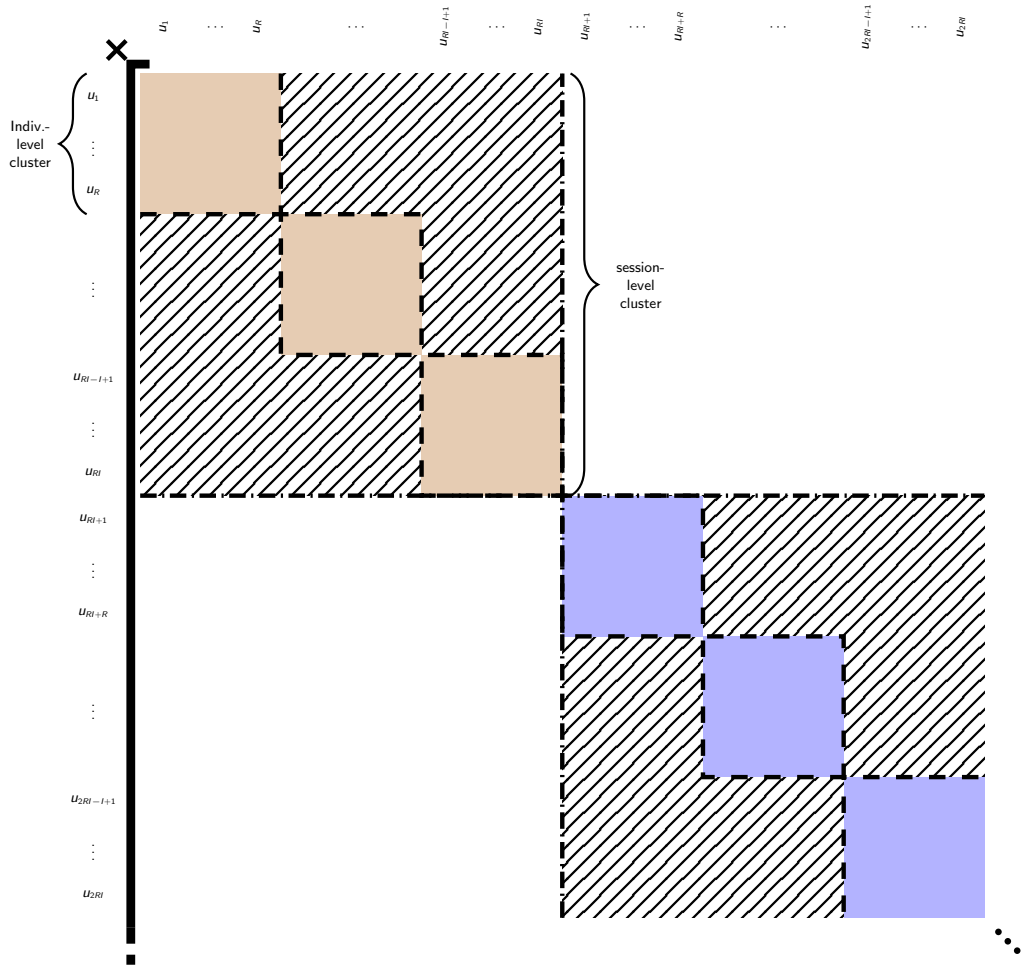
$$V_{clu}[\hat{\beta}] = \frac{\sum_{i\in T}\sum_{j\in T} E[\varepsilon_i\varepsilon_j]\mathbf{1}_{j\in C_i}}{S^2 I^2 R^2}$$

$$\hat{V}_{clu}[\hat{\beta}] = \frac{\sum_{i\in T}\sum_{j\in T} u_i u_j \mathbf{1}_{j\in C_i}}{S^2 I^2 R^2}$$

  so the key is the multiplication of residuals within a cluster.
- Note that if the entire set is a cluster, then $\sum_{i\in T}\sum_{j\in T} u_i u_j \mathbf{1}_{j\in C_i} = \sum_{i\in T}\sum_{j\in T} u_i u_j = 0$.

# Cluster-robust inference 101

We ordered the observations in a way that $\sum_{i \in T} \sum_{j \in T} u_i u_j \mathbf{1}_{j \in C_i}$ is the sum of entities on the block diagonal matrices of $u_i u_j$. Now the question becomes to figure out "the right size of block."



# What would be the proper cluster level?

- The **experimental design** should primarily determine the clustering level.

  (Abadie, Athey, Imbens, and Wooldridge (2017, 2020) made the same point for field experiments.)
- If the lab experiment asks the participants to make decisions in a similar environment repeatedly, *clustering at the participant level* is intended from the experimental design.
- If the cluster is determined in a purely statistical manner, then it begs questions: Are ALL potential clusters considered? Do you ditch your design?

# When to cluster at the session level?

!!! When the experimental design intends a strong correlation among error terms within the entire session data (e.g., consider a market experiment)

?? When the design intends observational relationship within an individual, but the $CRSE_{session}$ happen to be larger than $CRSE_{ind}$

I focus on the latter. First, it is less likely. Second, if it indeed is, then something might be wrong.

# 1. Will it be likely?

Imagine a controlled lab experiment on a group decision making where a between-subject design, random rematch, anonymity, and no communications are adopted.

| | Control-Session-01 | | | | | | Treatment-Session-01 | | | | | | Control-Session-02 | | | | | | Treatment-Session-02 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| r01 | 6 | 36 | 21 | 17 | 32 | 17 | 29 | 22 | 13 | 40 | 19 | 29 | 37 | 7 | 16 | 23 | 26 | 16 | 8 | 41 | 29 | 16 | 33 | 16 |
| r02 | 12 | 42 | 23 | 20 | 27 | 15 | 27 | 16 | 15 | 39 | 13 | 23 | 39 | 8 | 19 | 23 | 31 | 11 | 11 | 39 | 24 | 20 | 29 | 14 |
| r03 | 6 | 36 | 28 | 18 | 31 | 18 | 29 | 16 | 14 | 43 | 17 | 31 | 39 | 5 | 13 | 23 | 29 | 11 | 13 | 39 | 30 | 22 | 34 | 14 |
| r04 | 12 | 37 | 26 | 14 | 31 | 17 | 27 | 18 | 13 | 42 | 19 | 30 | 35 | 6 | 12 | 26 | 25 | 14 | 9 | 38 | 25 | 16 | 31 | 15 |
| r05 | 6 | 43 | 22 | 19 | 27 | 17 | 34 | 17 | 10 | 42 | 16 | 25 | 42 | 8 | 16 | 21 | 30 | 12 | 8 | 44 | 22 | 20 | 28 | 20 |
| r06 | 7 | 40 | 22 | 15 | 29 | 12 | 31 | 17 | 9 | 36 | 16 | 30 | 40 | 8 | 16 | 24 | 27 | 15 | 9 | 42 | 25 | 17 | 32 | 18 |
| r07 | 9 | 36 | 22 | 20 | 32 | 17 | 27 | 19 | 13 | 38 | 14 | 29 | 39 | 11 | 18 | 23 | 27 | 14 | 14 | 38 | 28 | 17 | 34 | 17 |
| r08 | 9 | 37 | 23 | 19 | 33 | 15 | 32 | 21 | 12 | 41 | 13 | 23 | 37 | 6 | 14 | 25 | 29 | 11 | 9 | 44 | 25 | 20 | 33 | 20 |
| r09 | 11 | 39 | 21 | 17 | 29 | 16 | 35 | 17 | 11 | 39 | 19 | 27 | 38 | 5 | 14 | 20 | 26 | 14 | 14 | 40 | 24 | 22 | 32 | 18 |
| r10 | 10 | 41 | 20 | 16 | 27 | 16 | 31 | 19 | 10 | 43 | 19 | 28 | 41 | 12 | 12 | 18 | 27 | 16 | 15 | 41 | 27 | 20 | 31 | 17 |
| Std. | 2.4 | 2.7 | 2.4 | 2.1 | 2.3 | 1.7 | 2.9 | 2.0 | 1.9 | 2.3 | 2.5 | 2.9 | 2.1 | 2.4 | 2.4 | 2.4 | 1.9 | 2.0 | 2.7 | 2.2 | 2.5 | 2.3 | 2.0 | 2.2 |
| | Std.(session)=11.2 | | | | | | Std.(session)=11.4 | | | | | | Std.(session)=11.4 | | | | | | Std.(session)=11.6 | | | | | |
| | Std.(whole)=11.4 | | | | | | | | | | | | | | | | | | | | | | | |

| | (1) | (2) | (3) |
|---|---|---|---|
| Treatment | 2.600** | 2.600 | 2.600*** |
| | (1.98) | (0.63) | (4.43) |
| _cons | 21.55*** | 21.55*** | 21.55*** |
| | (23.22) | (7.26) | (36.75) |
| Cluster SE | – | ID | Session |
| *N* | 240 | 240 | 240 |

*t* statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

False-positive effects when SEs are clustered at a session level.

# 2. If the session-level clusters yield larger SEs, how come?

- It is possible if there are dynamic session effects. If, say, the residual of person 1's round 12 decision is strongly correlated to the residual of person 7's round 1 decision, then the $CRSE_{ss}$ can be larger than $CRSE_{ind}$.
- Perhaps someone's initial choice profoundly affects other's later choices so that those observations are related.
- Wait. Isn't it then demonstrating that your random rematch doesn't work? Is it telling us that your samples may not be balanced?

# Discussions: Practical issues

1. Let's say we consider $CRSE_{ss}$ simply because it generates larger SE: why not considering at the date-of-session level, at the experimenter level, and at the experiment level for meta analysis?
2. An ad-hoc definition of a session: Suppose the experimenter splits 24 subjects into two, but the experiment is identical. Would a session consist of 24 subjects, or would two sessions consist of 12 subjects each?
3. Unequal size implies unequal weight: Suppose 24 subjects in session 1 and 12 subjects in session 2. If a fixed group of 6 played a game repeatedly, the observations from session 1 are likely less related because there are more (potentially heterogeneous) groups. Then $CRSE_{ss}$ will be more affected by session 2. Is it justified?

# Practical issues

4. Session size and #sessions: Given that the number of total participants is limited, considering session-level clusters encourages to have more sessions with a smaller number of subjects per session. It is problematic.
   - random rematch: Playing 10 games with a random partner. If 40 subjects per session, Pr(whole new match)=29.3%. If 15 subjects per session, Pr=1.9%. The fundamental reason for adopting a random rematch—minimizing the strategic interactions between the games—is compromised.
   - small sample$\not\Rightarrow$balanced: Suppose 4 subjects per session, and #female varies from 0 to 4. How can a researcher be sure whether the session effects are controlled by randomizing sessions, and if not, how does she distinguish the gender ratio effect from others unless having more sessions with sufficiently large variations of the gender ratio? A vicious cycle of demanding more sessions to control issues with a small-size session may be established.

# Takeaway messages

1. Session randomization is crucial.
2. A "typical" experimental design intends clusters at the participant (or independent decision group) level.
3. Clustering standard errors at the participant level is likely to yield more conservative statistical results.
4. (Did not cover in this talk) If there are several ways of clustering observations and those are equally justifiable, then a researcher would choose a cluster within which observations are more related.
5. Although not justifiable by the experimental design, clustering standard errors at the session level may be considered. It begs further questions of why a session, among several other potential levels, should be a level for clustering and of whether the sessions have balanced samples.