Causal Inference using Difference-in-Differences

Lecture 3: Clustering Issues

Pedro H. C. Sant'Anna
Emory University

January 2025

Summary of previous lecture

Summary of Lecture 2

- We have talked about the underlying assumptions in 2x2 DiD:
  - SUTVA;
  - No-Anticipation;
  - Parallel Trends.
- We have talked about identifying the ATT.
- We discussed estimating the ATT "by hand" and using TWFE regressions.
- We have talked about the importance of clustering.

Doing inference with a small number of clusters

Doing inference with a small number of clusters is hard

This discussion is based on Section 5 of Roth, Sant'Anna, Bilinski and Poe (2023).

- In some applications, the number of independent clusters may be small: CLT based on a growing number of clusters may provide a poor approximation
- The CLT may provide a poor approximation with few clusters, even if the number of units within each cluster is large.
  - Reasoning: the standard sampling-based view of clustering allows for arbitrary correlations of the outcome within each cluster
  - But there may be common components at the cluster level (a.k.a. cluster-level "shocks") that do not wash out when averaging over many units within the same cluster.
  - Since we only observe a few observations of the cluster-specific shocks, the average of these shocks will generally not be approximately normally distributed.

Ignoring the problem is not a way forward

- If we ignore this issue and pretend we have many clustered, we may have issues!
- MacKinnon and Webb (2018) have demonstrated using simulations that the cluster wild bootstrap may perform poorly in DiD settings with a small number of treated clusters.
- Canay, Santos and Shaikh (2021) provided a formal analysis of the conditions under which the cluster wild bootstrap procedure would be asymptotically valid in settings with a few large clusters.
- Canay et al. (2021): The reliability of these bootstrap procedures depends on imposing certain homogeneity conditions on treatment effects and the type of estimator used.

Doing inference with a small number of clusters

Model-based approaches

Model-based approaches

- Several papers have made progress on the difficult problem of conducting inference with a small number of clusters by modeling the dependence within clusters.
- These papers typically place some restrictions on the common cluster-level shocks, although the exact restrictions differ across papers.
- Typical starting point is

$$Y_{i,j,t} = \alpha_j + \phi_t + D_{j,t}\beta + (\nu_{j,t} + \epsilon_{i,j,t}), \tag{1}$$

  - $Y_{i,j,t}$ is the (realized) outcome of unit i, in cluster j, at time t;
  - α_j and ϕ_t are cluster and time fixed effects;
  - $D_{j,t}$ is an indicator for whether cluster j is treated in period t;
  - $\nu_{j,t}$ is a common cluster-by-time error term, and $\epsilon_{i,j,t}$ is an idiosyncratic unit-level error term.

Model-based approaches: TWFE approach

$$Y_{i,j,t} = \alpha_j + \phi_t + D_{j,t}\beta + (\nu_{j,t} + \epsilon_{i,j,t}).$$

- "Cluster-level" error term, $\nu_{j,t}$, induces correlation among units within the same cluster.
- It is often assumed that $\epsilon_{i,j,t}$ are *iid* mean-zero across i and j (and sometimes t); see, e.g., Donald and Lang (2007), Conley and Taber (2011), and Ferman and Pinto (2019).
- Letting $Y_{j,t} = n_j^{-1} \sum_{i:j(i)=j} Y_{i,j,t}$ be the average outcome among units in cluster j, where n_j is the number of units in cluster j, we can take averages to obtain

$$Y_{j,t} = \alpha_j + \phi_t + D_{j,t}\beta + \eta_{j,t}, \tag{2}$$

where $\eta_{j,t} = \nu_{j,t} + n_j^{-1} \sum_{i=1}^{n_j} \epsilon_{i,j,t}$.

Model-based approaches: TWFE approach in 2x2 DiD setup

- In the 2x2 setup, we know that the DiD-by-hand-estimator (at the cluster level) is equivalent to the OLS estimated coefficient $\widehat{\beta}$ from (2).
- We can also show that

$$\widehat{\beta} = \beta + \frac{1}{N_1}\sum_{j:D_j=1}\Delta\eta_j - \frac{1}{N_0}\sum_{j:D_j=0}\Delta\eta_j$$

$$= \beta + \frac{1}{N_{cluster,1}}\sum_{j:D_j=1}\left(\Delta\nu_j + n_j^{-1}\sum_{i=1}^{n_j}\Delta\epsilon_{ij}\right) - \frac{1}{N_{cluster,0}}\sum_{j:D_j=0}\left(\Delta\nu_j + n_j^{-1}\sum_{i=1}^{n_j}\Delta\epsilon_{ij}\right), \quad (3)$$

where now $N_{cluster,d}$ corresponds with the number of *clusters* with treatment d, and $\Delta\eta_j = \eta_{j2} - \eta_{j1}$ (and likewise for the other variables).

Model-based approaches: TWFE approach in 2x2 DiD setup

$$\widehat{\beta} = \beta + \frac{1}{N_{cluster,1}} \sum_{j:D_j=1} \left(\Delta \nu_j + n_j^{-1} \sum_{i=1}^{n_j} \Delta \epsilon_{ij} \right) - \frac{1}{N_{cluster,0}} \sum_{j:D_j=0} \left(\Delta \nu_j + n_j^{-1} \sum_{i=1}^{n_j} \Delta \epsilon_{ij} \right),$$

- With few clusters, the averages of the $\Delta \nu_j$ among treated and untreated clusters will tend **not to be approximately normally distributed**, and their variance may be difficult to estimate.
- Essentially, we can't rely on the consistency and asymptotically normality results we usually do!
- Common solutions in the literature: impose assumptions on these "structural error terms" to make inferences.

Model-based approaches: TWFE approach in 2x2 DiD setup

- I am personally not a big fan of these solutions because, implicitly, the assumptions on the errors in the structural model (1) impose (non-transparent) restrictions on the potential outcomes.

- In the Appendix of Roth et al. (2023), we have shown that, in this 2x2 setup, under SUTVA + No anticipation + PT, we have actually shown that this is indeed the case.
- So we need to be careful with all these approaches.

- But, at the same time, recognize that this is a hard problem!!

Model-based approaches: TWFE approach in 2x2 DiD setup

- To be more precise, in the Appendix of Roth et al. (2023), we have shown that, in this 2x2 setup, under SUTVA + No anticipation + PT, we have that that
  - $\beta = \tau_2$ is the ATT at the cluster level (no surprise),
  - $\nu_{j,t} = \nu_{j,t,0} + D_j \nu_{j,t,1}$ (no surprise),
  - $\epsilon_{i,j,t} = \epsilon_{i,j,t,0} + D_j \epsilon_{i,j,t,1}$ (no surprise),

  - $\epsilon_{i,j,t,0} = Y_{i,j,t}(\infty) - \mathbb{E}\left[Y_{i,j,t}(\infty) | j(i) = j\right]$,
  - $\epsilon_{i,j,t,1} = Y_{i,j,t}(2) - Y_{i,j,t}(\infty) - \mathbb{E}\left[Y_{i,j,t}(2) - Y_{i,j,t}(\infty) | j(i) = j\right]$
  - $\nu_{j,t,0} = \mathbb{E}\left[Y_{i,j,t}(\infty) | j(i) = j\right] - \mathbb{E}\left[Y_{i,j,t}(\infty) | D_j\right]$
  - $\nu_{j,t,1} = \mathbb{E}\left[Y_{i,j,t}(2) - Y_{i,j,t}(\infty) | j(i) = j\right] - \tau_t$
- Here, expectations are across units.

Let’s cover some examples

Donald and Lang(2007)

- Donald and Lang (2007): Directly assume that the "cluster-specific" shocks $\nu_{j,t}$ are mean-zero Gaussian, homoskedastic with respect to cluster and treatment status, and independent of other unit-and-time specific shocks.
  - Under these assumptions, if the cluster size is large, you can do inference using critical values from a t-distribution with $J-2$ degrees of freedom, where J is the total number of clusters.
- The key restriction is the assumption that the cluster-specific shocks $\nu_{j,t}$ are *iid* normal.
- The homoskedasticity assumption also rules out many forms of treatment effect heterogeneity.
  - For example, suppose the cluster-level means of $Y_{it}(\infty)$ have the same distribution among treated and untreated clusters. Then, if the average treatment effect at the cluster level is heterogeneous, this will tend to lead $\nu_{j,t}$ to have higher variance among treated clusters, thus violating the homoskedasticity assumption.

Conley and Taber (2011)

- Conley and Taber (2011): consider the setup where the number of treated clusters, J_1, is fixed and potentially equal to one, but there are a large number of untreated clusters, J_0, available.
- The main insight: if the cluster-specific error terms $\eta_{j,t}$ from the untreated group are informative about the cluster-specific error terms for the treated group, one can conduct inference about β using the estimated distribution of the untreated errors.
- To satisfy "informativeness", they impose:
  - $\epsilon_{i,j,t}$ are *iid* across i and independent of clusters and treatment status,
  - the cluster-specific shocks $\nu_{j,t}$ are *iid* across j, independent of treatment status, and have mean zero for all t,
  - all clusters grow at the same rate as J_0.

Conley and Taber (2011) and its variants

- Conley and Taber (2011) assumptions still rule out heterogeneity
- For instance, if average treatment effects differ across clusters, then this will tend to violate the assumption that $\nu_{j,t}$ is *iid* across *j*.
- Another limitation of the Conley and Taber (2011) procedure is that it does not accommodate settings with heterogeneous cluster sizes, a situation that often arises in practice.
  - Ferman and Pinto (2019) build on Conley and Taber (2011) and show how one can use bootstrap-based inference procedures to allow for some types of heteroskedasticity, paying particular attention to the case where heteroskedasticity arises due to variation in cluster sizes.
  - Requires you to estimate the source of heteroskedasticity (so you need to have a good model for it).

Hagemann (2020)

- Hagemann (2020): considers a rearrangement/permutation-based method that is applicable to DiD setups with a single large treated cluster and a fixed number of large untreated clusters.
- The main assumption: the average evolution of the untreated outcomes is the same across all untreated clusters.
  - This is strength parallel trends to the cluster level instead of the treatment level
- Like other proposals, Hagemann (2020) restricts heterogeneity.
  - essentially requires that, as cluster size grows large, any single untreated cluster could be used to infer the counterfactual trend for the treated group
  - This essentially rules out cluster-specific heterogeneity in trends in untreated potential outcomes (and this is testable).

Doing inference with a small number of clusters

Alternative approaches

Alternative approach I: condition on shocks

- All of the “model-based” papers above treat $\nu_{j,t}$ as random.
- An alternative perspective would be to condition on the values of $\nu_{j,t}$ and view the remaining uncertainty as coming from sampling individual units within clusters, constructing standard errors by clustering only at the unit level.
- The problem here is that this can violate parallel trends.
- However, the violation may be relatively small if the cluster-specific shocks are small relative to the idiosyncratic variation.

Alternative approach I: condition on cluster-level shocks

- Let's make this concrete and consider the setting of Card and Krueger (1994) that compares employment in NJ and PA after NJ raised its minimum wage.
- The model-based papers would consider NJ and PA as drawn from a super-population of treated and untreated states, where the state-level shocks are mean-zero.
- The alternative approach we are mentioning here would treat the two states as fixed and view any state-level shocks between NJ and PA as a violation of the parallel trends assumption.
- With two clusters only, this is essentially the only thing you can do.

Alternative approach II: Randomization-based inference

- A large literature in statistics and a growing literature in econometrics has considered Fisher Randomization Tests (FRTs), otherwise known as permutation tests.
- The basic idea is to calculate some statistic of the data (e.g. the t-statistic of the DiD estimator), then recompute this statistic under many permutations of the treatment assignment (at the cluster level).
- We then reject the null hypothesis of no effect if the test statistic using the original data is larger than 95% of the draws of the test statistics under the permuted treatment assignment
- If treatment is randomly assigned, then FRTs have exact finite-sample validity under the **sharp null of no treatment effects for all units**.

Alternative approach II: Randomization-based inference

- The advantage of these FRTs is that they place no restrictions on the values of $Y(\infty)$, and thus allow arbitrary heterogeneity in $Y(\infty)$ across clusters.
- On the other hand, the assumption of random treatment assignment may often be questionable in DiD settings, as it is substantially stronger than parallel trends.
- Moreover, the "sharp" null of no effects for all units may not be as economically interesting as the "weak" null of no average effects.
- Roth and Sant'Anna (2023) extend the idea of FRTs to settings where there is staggered adoption and (quasi-)random timing of treatment, and show that an FRT with a studentized statistic is both finite-sample valid for the sharp null and asymptotically valid (as the number of clusters grows) for the weak null. (We will talk more about this in a later lecture).

At the end, at which level should you cluster?

At what level should you cluster?

What level to cluster

- As we have discussed, choosing the level of clustering depends on different things (and what we can do about it).
- From the sampling perspective, it comes down to how the sample is drawn from the super-populations. You cluster at that level!
- From the model-based perspective, you may need to make some additional assumptions if considering "cluster-level" random shocks and observing few (treated) clusters.
- You can condition on shocks and cluster at unit-level, but that may generate violations of PT.
- Adopt a design-based approach and cluster at the level of treatment assignment.
  - This is justified in DiD (without random assignment) by Rambachan and Roth (2022).

References

Canay, Ivan A., Andres Santos, and Azeem M. Shaikh, "The wild bootstrap with a "small" number of "large" clusters," *Review of Economics and Statistics*, 2021, *103* (2), 346–363.

Card, David and Alan B Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, 1994, *84* (4), 772–793.

Conley, Timothy G. and Christopher R. Taber, "Inference with "Difference in Differences" with a Small Number of Policy Changes," *Review of Economics and Statistics*, 2011, *93* (1), 113–125.

Donald, Stephen G. and Kevin Lang, "Inference with Difference-in-Differences and Other Panel Data," *The Review of Economics and Statistics*, 2007, *89* (2), 221–233.

Ferman, Bruno and Cristine Pinto, "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity," *The Review of Economics and Statistics*, 2019, *101* (3), 452–467.

Hagemann, Andreas, "Inference with a single treated cluster," *arXiv:2010.04076 [econ.EM]*, 2020, pp. 1–23.

MacKinnon, James G. and Matthew D. Webb, "The wild bootstrap for few (treated) clusters," *The Econometrics Journal*, 2018, *21* (2), 114–135. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ectj.12107.

Rambachan, Ashesh and Jonathan Roth, "Design-Based Uncertainty for Quasi-Experiments," *arXiv:2008.00602*, 2022.

Roth, Jonathan and Pedro H.C. Sant'Anna, "Efficient Estimation for Staggered Rollout Designs," *Journal of Political Economy: Microeconomics*, 2023, *1* (4), 669––709.

_ , Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe, "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature," *Journal of Econometrics*, 2023, *235* (2), 2218–2244.