Differences in Differences

Compass-Lexecon

Raquel Carrasco

Universidad Carlos III de Madrid

May-June 2024

Outline Session 5

- DD with multiple time periods and differential timing
  - Solutions
    - Callaway and Sant'Anna (2021)
    - Sun and Abraham (2021)
  - Other solutions: Matrix completion
- Concluding remarks on DD

Lessons from Goodman-Bacon (2021)

- While conventional TWFE specifications make sensible comparisons of treated and untreated units in the canonical two-period DD setting, in the staggered case they typically make "forbidden comparisons" between already-treated units.
- As a result, treatment effects for some units and time periods receive negative weights in the TWFE estimand.
- In extreme cases, this can lead the TWFE estimand to have the "wrong sign" - e.g., the estimand may be negative even if all the treatment effects are positive.
- Even if the weights are not so extreme as to create sign reversals, it may nevertheless be difficult to interpret which comparisons the TWFE estimator is making, as the "control group" is not transparent, and the weights it chooses are unlikely to be those most relevant for economic policy.

Lessons from Goodman-Bacon (2021)

- Goodman-Bacon decomposition can be used to report the weights placed on the different TWFE estimates from each 2-period, 2-group estimate. This allows us to evaluate how much weight is being placed on "forbidden" comparisons of already-treated units and how removing the comparisons would change the estimate.
- Stata commands: `bacondecomposition` or `ddtiming`.
- Several recent papers have proposed alternative estimators that more sensibly aggregate heterogeneous treatment effects in settings with staggered treatment timing:
  - Callaway and Sant'Anna (2021) (CS) (Stata command: `csdid`)
  - Sun and Abraham (2021) (SA) (Stata command: `eventstudyinteract`)

Solutions to the problem: Callaway and Sant'Anna (2021)

- Idea: estimate the ATT separately for each group and time
  - Use as control group only groups that have not yet been treated
  - Aggregate the group-time ATTs into a (weighted) ATT using non-negative weights
- Parameter of interest:
  - Group-Time ATT: ATT for a specific group and time

$$ATT(g,t) = E(Y_t^1 - Y_t^0 \mid G_g = 1), \text{for } t > g$$

  - ATT at time t of starting treatment at time g, among the units that indeed started treatment at time g
  - G_g are indicators for groups treated at different times
  - CS identifies all feasible $ATT(g,t)$ and calculate an ATT per group/time
  - Provides a way to aggregate over these to get a single ATT

Solutions to the problem: Callaway and Sant'Anna (2021)

- Assumptions:
- T periods $(t = 1, ...T)$
- Irreversibility of the Treatment:
  - $D_1 = 0$ (i.e., everyone remains untreated in the first period). And for $t = 2, ...T$, $D_{t-1} = 0 \Longrightarrow D_t = 0$ (once a unit becomes treated, that unit will remain treated)
- Conditional parallel trends assumption is generalized using two types of "control" units: never treated:

$$E(Y_t^0 - Y_{t-1}^0 \mid X, G_g = 1) = E(Y_t^0 - Y_{t-1}^0 \mid X, C = 1),$$

where $C = 1$ for if "never treated". And not yet treated: for each (s, t) such that $t \geq g$, $s \geq t$

$$E(Y_t^0 - Y_{t-1}^0 \mid X, G_g = 1) = E(Y_t^0 - Y_{t-1}^0 \mid X, D_s = 0, G_g = 0)$$

Solutions to the problem: Callaway and Sant'Anna (2021)

$$ATT(g,t) = E(Y_t^1 - Y_t^0 \mid G_g = 1), \text{for } t > g$$

- They show that the family of group-time average treatment effects are identified under the previous assumptions.
- They can obtain different ATT's using the corresponding units as controls:
  - Group-Time ATT using the Never-Treated Units as Controls.
  - Group-Time ATT using the Not-Yet-Treated Units as Controls.
  - One can use outcome regression (OR), inverse probability weighting (IPW), or doubly robust (DR) estimators to obtain the different $ATT(g,t)$.

Solutions to the problem: Callaway and Sant'Anna (2021)

- For notational simplicity, we will focus on the IPW Estimators:

$$ATT(g,t) = E\left[\left(\frac{G_g}{E(G_g)}\frac{\frac{p_g(x)C}{1-p_g(x)}}{E\left(\frac{p(x)C}{1-p(x)}\right)}\right)(Y_t - Y_{g-1})\right]$$

- $(Y_t - Y_{g-1})$: Long differences between outcomes in period t and the period before group g was treated
- Propensity score $p_g(x) = \Pr(G_g = 1 \mid x, G_g + C = 1)$
- This is the IPW estimator. Alternatively, there is an outcome regression approach and a doubly robust.
  - Sant'Anna recommends DR.
- CS uses the never-treated or the not-yet-treated as controls but never the already-treated (there is a similar expression for the not-yet-treated).

Solutions to the problem: Callaway and Sant'Anna (2021)

- Next step: Aggregate the $ATT(g,t)$ across time and groups
  - Aggregate the $ATT(g,t)$ into fewer interpretable causal effect parameters, which makes interpretation easier, and also increases statistical power and reduces estimation uncertainty.
  - Part of the value of this paper is that they are able to carefully justify and make explicit the weights and comparisons they use to aggregate the parameters of interest.
- In general, the aggregation schemes are of the form:

$$\theta = \sum_{g=2}^{T} \sum_{t=2}^{T} \mathbf{1}(g \leq t) w(g,t) ATT(g,t)$$

- Different choices of $w(g,t)$ allow researchers to highlight different types of treatment effects.

Solutions to the problem: Callaway and Sant'Anna (2021)

- CS focus their discussion on how to answer three particular questions:

(1) How do average treatment effects vary across groups? (Heterogeneity by cohort)

$$\theta_s(g) = \frac{1}{T-g+1} \sum_{t=2}^{T} \mathbf{1}(g \leq t) ATT(g,t)$$

- These parameters are useful for understanding if the effect average of participating in the treatment was larger for groups that are treated earlier relative to groups that are treated later.
- Example: The effect of graduating during a recession on labor market outcomes are related to heterogeneous effects across groups
  - Ex: $T = 5; g = 3; g = 4;$
  - $\theta_s(3) = \frac{1}{3}\left[ATT(3,3) + ATT(3,4) + ATT(3,5)\right];$
  - $\theta_s(4) = \frac{1}{2}\left[ATT(4,4) + ATT(4,5)\right]$

Solutions to the problem: Callaway and Sant'Anna (2021)

(2) How does the effect of participating in the treatment vary with length of exposure to the treatment? (Dynamic effects)

$$\theta_D(e) = \sum_{g=2}^{T} \mathbf{1}(g + e \leq T) ATT(g, g+e) \Pr(G = g \mid C \neq 1, G + e \leq T)$$

- Average effect of participating in the treatment for the group of units that have been exposed to the treatment for exactly e time periods
- $e = t - g$ denotes the time elapsed since treatment was adopted

- This is perhaps the most popular summary measure used by practitioners.

Solutions to the problem: Callaway and Sant'Anna (2021)

$$\theta_D(e) = \sum_{g=2}^{T} \mathbf{1}(g + e \leq T) ATT(g, g+e) \Pr(G = g \mid C \neq 1, G + e \leq T)$$

- This is the average effect of participating in the treatment e time periods after the treatment was adopted across all groups that are ever observed to have participated in the treatment for exactly e time periods.
- Example:
  - $T = 5;\ e = 0; g = 3; g = 4;$

    $$\theta_D(0) = ATT(3,3) \Pr(G = 3 \mid C \neq 1, G \leq T) + ATT(4,4) \Pr(G = 4 \mid C \neq 1, G$$

  - $T = 5;\ e = 2; g = 3; g = 4;$

    $$\theta_D(2) = ATT(3,5)$$

- This is the natural target for event study regressions that are common in applied work.

Solutions to the problem: Callaway and Sant'Anna (2021)

(3) What is the cumulative average treatment effect of the policy across all groups until some particular point in time? (Heterogeneity by calendar time)

- In some applications, researchers may want to construct an aggregated parameter to highlight treatment effect heterogeneity with respect to calendar time.
  - For example, we want to study heterogeneous treatment effects across the business cycle.
- The average effect of participating in the treatment in time period t (across groups that have adopted the treatment by period t) is given by

$$\theta_c(t) = \sum_{g \in G}^{T} \mathbf{1}(t \geq g) ATT(g,t) \Pr(G = g \mid G \leq t)$$

Solutions to the problem: Callaway and Sant'Anna (2021)

$$\theta_c(t) = \sum_{g \in G}^{T} \mathbf{1}(t \geq g) ATT(g,t) \Pr(G = g \mid G \leq t)$$

- Example
  - $T = 5; g = 3; g = 4; t = 3;$

$$\theta_c(3) = ATT(3,3)$$

  - $T = 5; g = 3; g = 4; t = 4;$

$$\theta_c(4) = ATT(3,4) \Pr(G = 3 \mid G \leq 4) + ATT(4,4) \Pr(G = 4 \mid G \leq 4)$$

Solutions to the problem: Callaway and Sant'Anna (2021)

- An extension to $\theta_c(t)$ is to think about the cumulative effect of participating in the treatment up to some particular time period.
- For instance, in active labor market applications, policy makers may want to know the cumulative average effect of a given training program on earnings from the year that the first group of people were trained until year $\widetilde{t}$.
  - This would provide a measure of the cumulative earnings gains induced by the training program.

$$\theta_c^{cum}(\widetilde{t}) = \sum_{t=2}^{\widetilde{t}} \theta_c(t)$$

$\theta_c^{cum}(\widetilde{t})$ can be interpreted as the cumulative average treatment effect among the units that have been treated by time $\widetilde{t}$

- CS are able to carefully reason and make explicit which weights and comparisons they use to aggregate the parameters of interest.

Issues with dynamic TWFE: Sun and Abraham (2021)

- What happens when we consider a TWFE event-study specification?
- Sun and Abraham (2021) (SA) show that similar issues arise¡
- Goodman-Bacon (2021) focused on the decomposition of the TWFE to show bias under differential timing within a static specification.
- Callaway and Sant'Anna (2021) present alternative estimator that yields unbiased estimates of group time ATT which can be aggregated or put into event studies plots.
- SA is a "combination" of the two papers:
  - SA is a decomposition of the population regression coefficients on event studies leads and lags with differential timing estimated within TWFE.
  - SA proposed estimator coincides with CS when there are no covariates and use the never-treated/last-treated cohort as a comparison group.

Issues with dynamic TWFE: Sun and Abraham (2021)

- Once again they are bringing us bad news!
- They show that:
  - Even when we impose the strong unconditional parallel trends and the no-anticipation assumption, the OLS coefficients of the TWFE specification are, in general, very hard to interpret.
  - Coefficient on a given lead or lag can be contaminated by effects from other periods
  - Under treatment effect heterogeneity, leads and lags are biased (that is, pre-trends can arise solely from treatment effects heterogeneity!)

Issues with dynamic TWFE: Sun and Abraham (2021)

- SA propose an interacted TWFE estimator which addresses the problems they find.
- It coincides with CS when there are no covariates and use the never-treated/last-treated cohort as a comparison group.
- However, CS has many other results about the pitfalls of TWFE that are not in CS.

Issues with dynamic TWFE: Sun and Abraham (2021)

- The dynamic specification regresses the outcome on individual and period fixed effects, as well as dummies for time relative to treatment
- SA show that if all units have the same treatment effect β_s in the *sth* period after treatment, regardless of what year the unit enter into treatment, then the dynamic specification yields a sensible causal estimand, under suitable generalizations of the parallel trends and no anticipation assumptions.
- However, when there are heterogeneous dynamic treatment effects across adoption cohorts, the coefficients become difficult to interpret.
    - For example, problems may arise if the average treatment effect in the first year after adoption is different for states entering into treatment in 2014 as it is for states that adopted he policy 2015.

Issues with dynamic TWFE: Sun and Abraham (2021)

- In this case there are two issues:
  - First, as with the "static" regression specification, the coefficient β_s may put negative weight on the treatment effect s periods after treatment for some units.
  - Second, there could be is cross-lag "contamination". Thus, for example, the coefficient β_2 may be influenced by the treatment effect for some states three periods after.
  - Like the static specification, the dynamic specification thus fails to yield sensible estimates of dynamic causal effects under heterogeneity across cohorts.
  - The derivation of this result is mathematically more complex, and so we do not pursue it here.
  - The intuition is that, as in the static case, the dynamic OLS specification does not aggregate natural comparisons of units and includes "forbidden comparisons" between sets of units both of which have already been treated.

Issues with dynamic TWFE: Sun and Abraham (2021)

- An important implication of the results derived by SA is that if treatment effects are heterogeneous, the "treatment lead" coefficients are not guaranteed to be zero even if parallel trends is satisfied in all periods (and vice versa), and thus evaluation of pre-trends based on these coefficients can be very misleading.
  - SA propose a fully interacted regression to recover estimates of group-specific ATTs that avoid the issues related to TWFE regressions.
- SA proposal: It is an extension of the standard event-study TWFE model but saturating the model:
  - By including interactions between the relative time indicators (i.e. $t = -2, -1, ...$) with indicators for the treatment initiation year group, and then aggregate them to overall aggregate relative time indicators by cohort size.
- In the case of no covariates, this gives the same estimate as Callaway & Sant'Anna if you fully saturate the model with time indicators

Other solutions: Matrix completion

- There might be other approaches different from the DD framework to address the problem.
- An example of these alternative approaches is Athey et al. (JASA, 2021), who see the causal inference problem as a challenge of "missing data".
  - The central concept here is that, perhaps, with improved data quality and advanced algorithms, there exist alternative methods to reconstruct the counterfactual information essential for causal inference.
    - Using Machine Learning techniques, they attempt to retrieve potential outcomes by applying algorithms for "matrix completion".

Other solutions: Matrix completion

- What is matrix completion?
- Completing a matrix means guessing at the correct values that are missing
- In causal inference, if the matrix is a matrix of potential outcomes (e.g., Y^0), then missingness is caused by treatment assignment
- Example: A synthetic control design with a single (unit i) is missing Y^0 from the $3rd$ periods:

$$Y^0 = \begin{pmatrix} Y^0_{11} & Y^0_{12} & \cdots & Y^0_{1t} \\ Y^0_{21} & Y^0_{22} & \cdots & Y^0_{2t} \\ \cdots & \cdots & \cdots & \cdots \\ Y^0_{i1} & Y^0_{i2} & ? & ? \end{pmatrix}$$

- Matrix completion methods will impute the missing values (using regularized regression)
- And once you have those, you can calculate individual level treatment effects that can be used to aggregate to the ATT

Summary

- The "recent" DD literature has
  - pointed out a number of limitations of traditional regression-based approaches for implementing DD identification strategies in the presence of treatment effect heterogeneity, and
  - proposed a set of alternative approaches that do not suffer from the same set of limitations.
- Different groups of solutions take different approaches:
  - some carefully outline the control group and exclude the forbidden group from being used as a control group (CS)
  - another class of solutions estimates regressions with interaction terms to capture the heterogeneity in treatment (SA)
  - other solutions: Wooldridge (2021), similar to SA....

Summary

- These new methods may initially appear complicated, although there currently exist several well-developed software packages implementing them.
- However, while traditional TWFE regressions are easy to specify, they are actually quite difficult to interpret, since they make complicated and unintuitive comparisons across groups.
- By contrast, the methods that have a simple interpretation using a proper comparison group.
- And they can be viewed as simple aggregations of comparisons of group means.

Concluding remarks on DD

- To conclude, there are additional topics and extensions not covered in these sessions:
  - More complicated treatment regimes (e.g., moving into and out of treatment)
  - Models with continuous treatments
  - Fuzzy DD (all groups are treated in both time periods, but the proportion of units exposed to treatment increases in one group but not in the other)
    - Ref: de Chaisemartin and D'Haultfoeuille (2021; 2018); Callaway, Goodman-Bacon, and Sant'Anna (2021)
  - Distributional treatment effects (the effect of a treatment on the entire distribution of an outcome)
    - Ref: Athey and Imbens (2006)
  - Spillover effects. The vast majority of the DD literature imposes the SUTVA assumption, which rules out spillover effects. However, spillover effects may be important in many economic applications, such as when policy in one area affects neighboring areas, or when individuals are connected in a network.
    - Ref: Butts (2021); Huber and Steinmayr (2021)
  - etc.....

Concluding remarks on DD

- DD is a comparison of means and we should know which means we're comparing
- In some cases, using the "new" DD methods will not lead to a big change in your results (empirically, treatment effect heterogeneity is not that large in most cases)
    - The exceptions are cases where there are many periods with very few treated units – this is when "forbidden comparisons" get the most weight
- The most important thing is to be precise about who you want the comparison group to be and to choose a method that only uses these "clean comparisons".

Limitations of treatment effect approach

- Main goal: evaluate ex-post the impact of an existing policy.
- Compare distribution of a chosen outcome variable for individuals affected by the policy (the treatment group), with the distribution of unaffected individuals (control group).
- Main challenge: comparison so that the distribution of outcome for the control group serves as a good counterfactual for the distribution of the outcome for the treated group in the absence of treatment.
- Main advantage: given its focus on internal validity, the exercise gives transparent and credible identification.

Limitations of treatment effect approach

- Main limitations:
  - Estimated parameters are not useful for welfare analysis because they are not deep parameters (they are reduced-forms instead), and as a result, they are not policy-invariant (Lucas, 1976; Heckman and Vytlacil 2005).
  - It does not address general equilibrium effects.
  - Causal mechanisms remain a "black box"

Structural Estimation

- Another possibility to quantitative policy evaluation is to use a structural approach:
  - specifies a class of theory-based models of individual choice
  - chooses the one within the class that best fits the data
  - and uses it to evaluate policies through simulation.
- Main advantage:
  - it allows evaluating different variations of a similar policy without need to change the structure of the model or reestimate it
- Main critique:
  - host of untestable functional form assumptions which have unknown implications for the results (too much discretion and sensitivity of estimates to assumptions about functional forms and distributions of unobservables)
  - too much emphasis on external validity at the expense of a more basic internal validity
  - complexity & computational cost (replication and sensitivity analyses are often more difficult in this approach than in the program evaluation approach).

Structural Estimation

- The differences between the two approaches have divided the economics profession into two groups whose research programmes have developed almost independently, although they focus on similar issues.
  - Two papers, Deaton (2009), and Heckman and Urzua (2009), argue against what they see as an excessive and inappropriate use of experimental and quasi-experimental methods in empirical work in economics.
  - Imbens (JEL, 2010) in the paper "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)" argues that empirical work is much more credible as a result of the natural experiments revolution started by Card, Angrist, Krueger, and others in the late 1980s.

Structural Estimation

- However, researchers have realized about the important complementarity between the two: Heckman (2010): "Building bridges between structural and program evaluation"; Whited (2022)...
- The two approaches can be reconciled by noting that for many policy questions, it is not necessary to identify fully specified models to answer a range of policy questions.
  - It is often sufficient to identify policy-invariant combinations of structural parameters.
  - These combinations are often much easier to identify (i.e., require fewer and weaker assumptions),