DiD with heterogeneous treatment effects
Overview part II

Manuel Bagues

University of Warwick

References and links to relevant material

1. De Chaisemartin, Clément, and Xavier d'Haultfoeuille (2020). "Two-way fixed effects estimators with heterogeneous treatment effects." American Economic Review.
2. Rambachan, Ashesh, and Jonathan Roth (2019). "An honest approach to parallel trends."
3. Roth, Jonathan (2019). "Pre-test with caution: Event-study estimates after testing for parallel trends."
4. Kahn-Lang, Ariella, and Kevin Lang (2020). "The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications." Journal of Business Economic Statistics.
5. Manski, Charles F., and John V. Pepper (2018). "How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions." Review of Economics and Statistics.
6. Cengiz, Doruk, et al.(2019) "The effect of minimum wages on low-wage jobs." The Quarterly Journal of Economics.
7. Silvia Vannutelli Differences-in-differences summary

Agenda for today

1. Negative Weights, diagnostics and solution (Chaisemartin and d'Haultfoeuille)
2. Solution II (Cengiz at al. stacked diff-in-diff)
3. Pre-Trends
    - Weakening the parallel trends assumption: Rambachan & Roth 2019, Pepper & Manski 2018.
    - Power issues: Roth 2019

Weights - Recap

From de Chaisemartin and d'Haultfoeuille

- Main result: without assuming constant TE,

$$E\left[\widehat{\beta}\right] = E\left[\sum_{g,t} W_{g,t}\Delta_{g,t}\right], \tag{1}$$

  where $W_{g,t}$: weights summing to 1, and $\Delta_{g,t}$ =ATE in group g at time t.

- $W_{g,t} \neq$ to proportion of units in (g,t), so $\beta \neq ATE$.
- But even worse, often times, many weights $W_{g,t}$ are < 0.
- Then, $E\left[\widehat{\beta}\right]$ could be < 0 even if all the $\Delta_{g,t}$ are > 0.
- Estimating weights = diagnostic of β's robustness to heterogeneous TE.

Groups and time periods

- One considers observations that can be divided into G groups and T periods.
- For every $(g, t) \in \{1, ..., G\} \times \{1, ..., T\}$: $N_{g,t}$ = number of observations in group g at period t , and $N = \sum_{g,t} N_{g,t}$ = total number of observations.
- Data may be:
  - individual-level panel or repeated cross-section data set where groups are, e.g., individuals' county of birth;
  - cross-section data set where cohort of birth plays the role of time.
- One may have $N_{g,t} = 1$, e.g. because a group is actually an individual or a firm.

Notation

- We assume binary treatment (results extend to non-binary treatments as well)
- $D_{i,g,t}$: treatment of observation i in group g and at period t.
- $(Y_{i,g,t}(0), Y_{i,g,t}(1))$: potential outcomes.
- $Y_{i,g,t} = Y_{i,g,t}(D_{i,g,t})$: observed outcome.
- For any X, we let $X_{g,t} = \sum_{i=1}^{N_{g,t}} X_{i,g,t}/N_{g,t}$.
- We also let $D_{g,.}$ (resp. $D_{.,t}$, $D_{.,.}$) be the average value of the treatment in group g (resp. in period t, over all g,t).
- $\widehat{\beta}_{fe}$= OLS coeff. of $D_{g,t}$ in a reg. of $Y_{i,g,t}$ on group and time FE and $D_{g,t}$.
- We then let $\beta_{fe} = E\left[\widehat{\beta}_{fe}\right]$.

Assumptions

1. (Balanced panel of groups) For all $(g,t) \in \{1,...,G\} \times \{1,...,T\}$, $N_{g,t} > 0$.
2. (Sharp design) For all $(g,t) \in \{1,...,G\} \times \{1,...,T\}$ and $i \in \{1,...,N_{g,t}\}$, $D_{i,g,t} = D_{g,t}$.
3. (Independent groups) The vectors $(Y_{1,g}(0), Y_{1,g}(1), D_{1,g}, ..., Y_{T,g}(0), Y_{T,g}(1), D_{T,g})$ are mutually independent.
4. (Strong exogeneity) For all $g \in \{1,...,G\}$, $E(Y_{g,t}(0) - Y_{g,t-1}(0) | D_{g,1}, ..., D_{g,T}) = E(Y_{g,t}(0) - Y_{g,t-1}(0))$ (shocks independent of her past, present and future treatments)
5. (Common trends) For all $t \geq 2$, $E(Y_{g,t}(0) - Y_{g,t-1}(0))$ does not vary across g.

Parameters of interest

Let $\Delta^{TR} = \frac{1}{N_1}\sum_{i,g,t}\left(Y_{i,g,t}(1) - Y_{i,g,t}(0)\right)$, with $N_1 = \sum_{(g,t):D_{g,t}=1} N_{g,t}$.

Let $\delta^{TR} = E\left[\Delta^{TR}\right]$: δ^{TR} is the ATT.

Let $\Delta_{g,t}$ denote the ATE in cell (g,t):

$$\Delta_{g,t} = \frac{1}{N_{g,t}}\sum_{i=1}^{N_{g,t}}\left(Y_{i,g,t}(1) - Y_{i,g,t}(0)\right).$$

Then δ^{TR} satisfies

$$\delta^{TR} = E\left[\sum_{g,t:D_{g,t}=1}\frac{N_{g,t}}{N_1}\Delta_{g,t}\right].$$

We now show a similar result on β_{fe}, but with additional, possibly < 0 weights.

β_{fe} = weighted sum of ATEs under common trends

Let $\epsilon_{fe,g,t}$= residual of observations in cell (g,t) in regression of $D_{g,t}$ on a constant, group FEs, and time FEs.

We define the weights $w_{fe,g,t}$ as:

$$w_{fe,g,t} = \frac{\epsilon_{fe,g,t}}{\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \epsilon_{fe,g,t}}.$$

If assumptions maintained above hold, then,

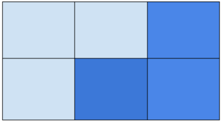
$$\beta_{fe} = E\left[\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} w_{fe,g,t} \Delta_{g,t}\right].$$

Therefore, in general $\beta_{fe} \neq \delta^{TR}$

Example

- FE regression with 2 groups 3 periods, and group 1 only treated at period 3, group 2 treated at periods 2 and 3. N obs. the same in each (g, t). Then, $\epsilon_{fe,g,t} = D_{g,t} - D_{g,.} - D_{.,t} + D_{.,.}$, so:

$$
\begin{aligned}
{}_{fe,1,3} &= 1 - 1/3 - 1 + 1/2 = 1/6 \\
{}_{fe,2,2} &= 1 - 2/3 - 1/2 + 1/2 = 1/3 \\
{}_{fe,2,3} &= 1 - 2/3 - 1 + 1/2 = -1/6.
\end{aligned}
$$



- Weight definition and some algebra imply:

$$\beta_{fe} = 1/2E(\Delta_{1,3}) + E(\Delta_{2,2}) - 1/2E(\Delta_{2,3}).$$

$$\beta_{fe} \neq \delta^{TR} = 1/3E(\Delta_{1,3}) + 1/3E(\Delta_{2,2}) + 1/3E(\Delta_{2,3}).$$

β_{fe} may be of opposite sign than the $\Delta_{g,t}$s

$$\beta_{fe} = 1/2E(\Delta_{1,3}) + E(\Delta_{2,2}) - 1/2E(\Delta_{2,3}).$$

- The weight assigned to group 2 in period 3 is < 0.
- Then, β_{fe} may be very misleading measure of treatment effect.
- E.g., assume $E(\Delta_{1,3}) = E(\Delta_{2,2}) = 1$ and $E(\Delta_{2,3}) = 4$. Then,

$$\beta_{fe} = 1/2 \times 1 + 1 - 1/2 \times 4 = -1/2.$$

- $\beta_{fe} < 0$ while $E(\Delta_{1,3})$, $E(\Delta_{2,2})$, and $E(\Delta_{2,3})$ are all > 0.
- Negative weights are an issue only if $E(\Delta_{g,t})$s heterogeneous. If $E(\Delta_{1,3}) = E(\Delta_{2,2}) = E(\Delta_{2,3}) = 1$, then $\beta_{fe} = 1$.

Intuition for the negative weights

- In this simple example, one can show that $\beta_{fe} = (DID_1 + DID_2)/2$, with

$$DID_1 = E(Y_{2,2}) - E(Y_{2,1}) - (E(Y_{1,2}) - E(Y_{1,1}))$$
$$DID_2 = E(Y_{1,3}) - E(Y_{1,2}) - (E(Y_{2,3}) - E(Y_{2,2}))\,.$$

- Control group in DID_2, group 2, is treated both in the pre and in the post period. Therefore, under common trends, one can show that $DID_1 = \Delta^{TR}_{2,2}$, but $DID_2 = \Delta^{TR}_{1,3} - (\Delta^{TR}_{2,3} - \Delta^{TR}_{2,2})$.
- DID_2 is equal to average treatment effect in group 1 period 3, minus change in average treatment effect of group 2 between periods 2 and 3 (see also Chaisemartin, 2011, Borusyak and Jaravel, 2017, and Goodman-Bacon, 2018).
- Intuitively, mean outcome of groups 1 and 2 may follow different trends from period 2 to 3 either because group 1 becomes treated, or because treatment effect changes in group 2.

Characterizing (g, t) cells weighted negatively by β_{fe}

- β_{fe} more likely to assign negative weight to periods where a large fraction of observations treated, and to groups treated for many periods.
- Negative weights = concern when treatment effects may differ at periods when many / few groups treated, or across groups treated for many periods / few periods.
- In staggered designs (where $D_{g,t} \geq D_{g,t-1}$ for all g, t):
  - $w_{g,t}$ is decreasing in t (also Borusyak and Jaravel, 2017)
  - groups adopting treatment earlier more likely to have < 0 weights.

Summary

Chaisemartin and d'Haultfœuille

- $\hat{\beta}_{twfe}$ is a weighted average of ATE in each treated cell, but weights can be negative
- Weights are the product of sample share and residuals from a regression of treatment indicator on group and period FE.
- Negative weights are a concern only when treatment effects are heterogenous.
- Aside / note: New paper with multiple treatments. With multiple treatments, not only negative weights, but also contamination from other treatments in ATT (see Chaisemartin and d'Haultfœuille (2021).

Solution II

Chaisemartin and d'Haultfœuille

1. Estimate weights as diagnostic measure of β's robustness to heterogeneous TE. Test for negative weights: ratio between $\left|\hat{\beta}_{\text{twfe}}\right|$ and S.D. of the weights.
2. Intuitively, if ratio is close to 0 , the $\hat{\beta}_{\text{twfe}}$ and ATT can be of opposite signs, even if amount of TE heterogeneity is small.
3. Alternative estimand: average of the ATEs of switching cells (joiners' TE and leavers' TE), weighted by sample shares, consequently, different estimator
4. Notice that this will capture only instantaneous effects, no long term (for long-term, use "long differences" from Callaway Sant'Anna).
5. For staggered adoption: average of the treatment effect at the time when a group starts receiving the treatment (joiners' TE), using only treated-untreated comparisons.
6. Placebo estimator for pre-trends.

Stata commands: *twowayfeweights, fuzzydid, did multiplegt*

Stacked differences-in-differences: Steps

Cengiz, Dube, Lindner, and Zipperer (2019)

1. Create separate datasets for each treatment-cohort g.
2. Keep all units treated in that cohort, and all units that are not treated by year $g + k$ where g is the cohort-treatment year and k is the outermost relative year that you want to test (e.g. if you do an event study plot from -5 to 5 , would equal 5).
3. Keep only observations within years $g - k$ and $g + k$ for each cohort-specific dataset, and then stack them in relative time.
4. Append all cohort-specific datasets together.
5. Run the same TWFE estimates as in standard DiD but include interactions for the cohort-specific dataset with all of the fixed effects, controls, and clusters.

Stacked differences-in-differences: Application

Cengiz, Dube, Lindner, and Zipperer (2019)

- Impact of minimum wage changes in US on low-wage jobs across a series of 138 state-level minimum wage changes between 1979-2016.
- 138 event h-specific datasets including the outcome variable and controls for the treated state h and all other “clean controls states” in timeframe (-3 to +4)
- For each event, run a ”single treatment” diff-in-diff:
- Comparing only switchers to not (yet) treated units (drop already treated states).
- Prevents negative weighting but less statistical power (less observations included).

Pre-trends: Levels and trajectories

The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications, Kahn-Lang and Lang 2019

1. Similarity in levels, not only trends, makes common trends assumption more plausible: *why* do levels differ, and can the same mechanism affect trends?
2. If levels (or distribution) differs, functional form matters, and implies a different counterfactual - should be theoretically justified.
   - Example: levels vs. log.
3. Pre-trends tests are not sufficient to establish "parallel trends", e.g. because of false negatives (more later, Roth 2019)
4. Test sensitivity to range of assumptions on trends (next up).

Pre-trends and the parallel trends assumption

Researchers usually seek reassurance for the parallel trends assumption by looking at pre-trends for treatment and control groups, e.g. significant coefficient on "leads".
Main issues:

1. Parallel trends may not hold exactly.
2. Statistical power in testing for pre-trends.

Relaxation of parallel trends

An Honest Approach to Differences-in-Differences, Rambachan and Roth 2019

- Classical parallel trend assumption requires no difference in trend between treatment and control. $\delta=0$
- Instead, new method allows δ to lie in a set of trend differences Δ, specified by the researcher. The common parallel trends assumption $\delta=0$ is then a "special case" in this framework.

Relaxation of parallel trends

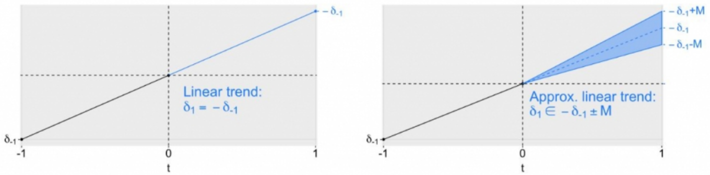
An Honest Approach to Differences-in-Differences, Rambachan and Roth 2019

Deviation from a linear trend bounded above by M

$$\Delta^{SD}(M) := \{\delta : |(\delta_{t+1} - \delta_t) - (\delta_t - \delta_{t-1})| \leqslant M, \forall t\} \qquad (2)$$

where for $t > 0, \delta_t$ refers to the t-th element of δ_{post} , δ_{-t} refers to the t-th element of δ_{pre} , and we adopt the convention that $\delta_0 = 0$.[8] The parameter $M \geqslant 0$ governs the amount by which the slope of δ can change between consecutive periods. [9] In the special case where $M = 0$, $\Delta^{SD}(0)$ requires that the difference in trends be exactly linear.

Figure 1: Linear and Approximately Linear Trends



Pre-trends: bounds

Manski and Pepper 2018 (Special case of Rambachan and Roth) How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions.

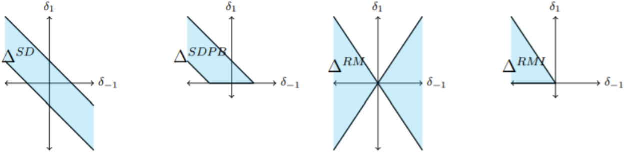
- Bounds informed by pre-treatment trend differences.
- Look at pre-treatment values of outcomes in $\top \& C$
- Calculate all the changes btw $T \& C$ across consecutive years in the pre-treatment periods

$$[Y_{T,t-1} - Y_{C,t-1}] - [Y_{T,t-2} - Y_{C,t-2}] = \delta_{t,t-1}$$

- Standard parallel pre-trend assumption assumes $\delta_{t,t-1} = 0 \forall$t before treatment
- Bound Parameter = maximum value across all $\delta_{t,t-1}$

Choices of Δ

Figure 2: Example choices for Δ



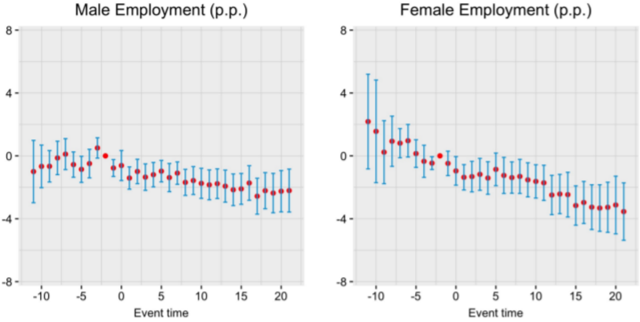
Note: Diagrams of potential restrictions Δ on the set of possible violations of parallel trends in the three-period difference-in-differences model. See discussion in Section 2 for further details on each example.

$$\text{Linear: } \Delta = \{\delta : \delta_1 = -\delta_{-1}\}$$
$$\text{Linear approx.: } \Delta^{SD}(M) = [-\delta_{-1} - M, -\delta_{-1} + M].$$
$$\text{Based on pre-trend diff: } \Delta^{RM}(\bar{M}) = \left\{(\delta_{-1}, \delta_1)' : |\delta_1| \leqslant \bar{M}\,|\delta_{-1}|\right\}.$$

Example application

Rambachan and Roth 2019 based on Lovenheim and Willen 2019

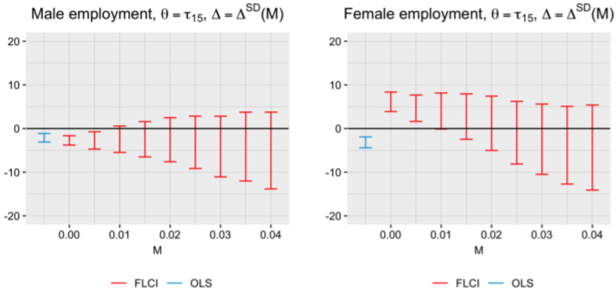
- Long run effect of collective bargaining on employment. Impact of state-level public sector duty-to-bargain (DTB) laws on student labor market outcomes.
- Outcome considered: employment.



Example application

Rambachan and Roth 2019 based on Lovenheim and Willen 2019

Figure 7: Sensitivity analysis for $\theta = \tau_{15}$ using $\Delta = \Delta^{SD}(M)$



- For $M < 0.01$, opposite sign by gender.
- For $M > 0.01$, cannot reject null effects.

Summary

Rambachan and Roth 2019

- Possible differences in trends are restricted to some set Δ, instead of assuming δ=0.
- Partial (set) identification of treatment effect, given M.
    - Choice of M depends on the underlying economic mechanism that leads to violation - benchmark M using knowledge of the likely magnitudes of those mechanisms.
- It is possible to back out the breakdown value of M at which treatment effects are no longer significant.
- R Code: *HonestDID*

Pre-trends: Power issues

Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends. Roth 2019

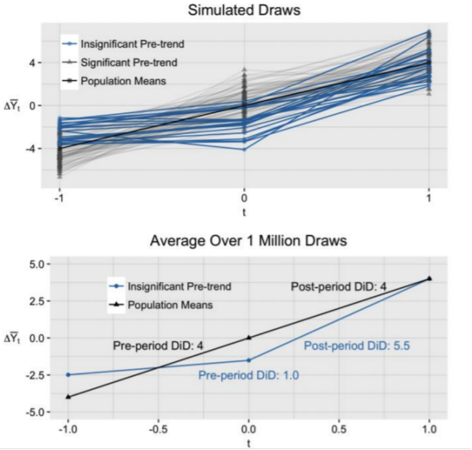
1. **Failure to reject the null of parallel trends does not imply absence of non-common trends $\neq$ existence of parallel trends (false negative) in case of underpowered test.**
2. This may introduce bias, exacerbated by the rejection of the pre-trends.

Pre-trends: Power issues - Roth 2019

True causal effect is 0 ($y_{it}(1) = y_{it}(0)$), and true model is:

$$y_{it}(0) = \alpha_i + \phi_t + D_i \times g(t) + \epsilon_{it} \tag{3}$$

With underlying upward trend $g(t) = \gamma t$



Pre-trends: Power issues, take-aways from simulation

Roth 2019

- When there is an underlying trend, pre-trends testing exacerbates bias.
- Statistical noise in finite sample may prevent detecting trend
- Blue draws would not detect a pre-trend
- True slope between -1 and 0 would be $-\beta_{-1}$, and β between 0 and 1, but in the blue ones $\beta = 0$
- If we get these draws (the cases where we fail to detect the underlying trend), we will produce large treatment estimates because of this failure.
- $\rightarrow$ "Passing" the pre-trends test, paradoxically leads to more biased estimates.

Pre-trends: Power issues

Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends. Roth 2019

- Pre-trends tests often underpowered.
- Overstatement of the treatment effect follows from rejection of parallel trends assumption due to "noise".
- Reporting DiD effects conditional on surviving a pre-trend test of introduces a pre-testing problem, which can exacerbate the bias from an underlying trend, and lead to wrong CI.
- Additionally: pre-trends testing is a special case of "pre-testing" (proceed only conditional on "passing" the test) → standard errors need to be corrected (Roth 2019)
- Parametric approaches: impose a structure for differential trends (e.g. linear), control parametrically for it without pre-testing.
- Alternative relaxations of parallel trends assumptions: e.g. Rambarachan & Roth (2019), Manski & Pepper (2018)
- Code: *Shiny app*

Conclusion

- Intuition for negative weights
  - de Chaisemartin & d'Hauftfoeille diagnostics and solution + stacked diff-in-diff solution.
- Problems with parallel trends → "Pre-test" honestly + with caution!
  - May not hold in general → weaker assumption + structure → bounds.
  - Pre-trend tests underpowered: may lead to biased estimates.