Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey

Clément de Chaisemartin Xavier D'Haultfœuille

Sciences Po CREST-ENSAE

2022 French Stata Conference in applied econometrics

Marseille, June 2022

The pervasiveness of two-way fixed effect regressions

- To estimate effect of a treatment/policy on an outcome, researchers often consider two-way fixed effects (TWFE) models of the kind:

$$Y_{g,t} = \alpha_g + \gamma_t + \beta_{fe} D_{g,t} + \epsilon_{g,t}.$$

- E.g.: employment in county g and year t regressed on county FEs, year FEs, and minimum wage in county g year t.
- Extremely pervasive in economics: **26 of 100 most cited 2015-2019 AER papers estimate TWFE** (dCDH, 2021).
- Also commonly used in political science, sociology, and environmental sciences.

Researchers have long thought that TWFE = DID

- With 2 groups (s and n) and 2 periods (1 and 2), DID estimator is:

$$\text{DID} = Y_{s,2} - Y_{s,1} - (Y_{n,2} - Y_{n,1}). \quad (1)$$

where s switches from no treatment to treatment;
n remains untreated.

- Let $Y_{g,t}(d)$= potential outcome for group g at t under treatment value d.
- DID relies on // trends assumption: without treatment, both groups would have experienced same outcome evolution:

$$E[Y_{s,2}(0) - Y_{s,1}(0)] = E[Y_{n,2}(0) - Y_{n,1}(0)].$$

- Under // trends, DID unbiased for ATE in group s at period 2:

$$E[\text{DID}] = E[Y_{s,2}(1) - Y_{s,2}(0)].$$

Unlike DID, TWFE relies on constant effect assumption

- Recent research has shown that unlike DID, TWFE estimators generally unbiased for an ATE only if:
  1. // trends holds;
  2. **Treatment effects are constant, between groups and over time**.
- Point 2 often implausible. E.g.: effect of minimum wage on employment likely to differ in counties with highly vs less educated workers.
- Realization that most commonly used method in quantitative social sciences relies on implausible assumption has spurred flurry of papers:
  1. diagnosing the seriousness of the issue;
  2. proposing alternative estimators.
- This survey provides an overview of this recent literature.

Some of the papers discussed in this survey

- Borusyak K., Xavier Jaravel X. and Spiess J. (2021), Revisiting event study designs: Robust and efficient estimation, *Working paper*.
- Callaway B. and Sant'Anna P. (2021), Difference-in-differences with multiple time periods (2021), *Journal of Econometrics*.
- de Chaisemartin C. and D'Haultfœuille X. (2018), Fuzzy difference-in-differences, *Review of Economic Studies*.
- de Chaisemartin C. and D'Haultfœuille X. (2020), Two-way fixed effect estimators with heterogeneous treatment effects, *American Economic Review*.
- de Chaisemartin C. and D'Haultfœuille X. (2021a), Difference-in-differences estimators of intertemporal treatment effects, *Working paper*.
- de Chaisemartin C. and D'Haultfœuille X. (2021b), Two-way fixed effects regressions with several treatments, *Working paper*.
- Goodman-Bacon A. (2021), Difference-in-differences with variation in treatment timing, *Journal of Econometrics*.
- Sun L. and Abraham S. (2021), Estimating dynamic treatment effects in event studies with heterogeneous treatment effects, *Journal of Econometrics*.

Outline

TWFE may not estimate convex combination of effects

- dCDH (2020) show that under // trends:

$$E\left[\widehat{\beta}_{fe}\right] = E\left[\sum_{(g,t):D_{g,t}\neq 0} W_{g,t} TE_{g,t}\right]. \tag{2}$$

  $TE_{g,t}$= treatment effect in g at t and $W_{g,t}$= weights summing to 1.

- $W_{g,t} \neq$ proportional to population of cell (g,t), so $\widehat{\beta}_{fe}$ may be biased for the average treatment effect across all treated (g,t) cells.
- Some $W_{g,t}$s may be <0. Then, $\widehat{\beta}_{fe}$ doesn't satisfy "no-sign-reversal": $E\left[\widehat{\beta}_{fe}\right]$ may be, say, <0 even if $TE_{g,t}>0$ for all (g,t).
- Issue more likely with non-binary than with binary treatment.
- The `twowayfeweights` Stata and R commands compute weights $W_{g,t}$.

Origin: $\widehat{\beta}_{fe}$ may compare switchers to always treated

- When D binary and design staggered ($D_{g,t} \geq D_{g,t-1}$), Goodman-Bacon (2021) shows that $\widehat{\beta}_{fe}$ = weighted average of two types of DIDs:
  - DID_1, comparing group s switching from untreated to treated to group n untreated at both dates.
  - DID_2, comparing switching group s to group a treated at both dates.
- `bacondecomp` Stata and R packages compute the DIDs and their corresponding weights entering in $\widehat{\beta}_{fe}$.
- **Negative weights in** (2) **originate from second type of DIDs.**
- Example: group e treated at $t = 2$, group ℓ treated at $t = 3$. Then:

$$\widehat{\beta}_{fe} = \frac{1}{2} \times \underbrace{\text{DID}^{1-2}_{e-\ell}}_{\text{DID}_1} + \frac{1}{2} \times \underbrace{\text{DID}^{2-3}_{\ell-e}}_{\text{DID}_2}.$$

Origin: $\widehat{\beta}_{fe}$ may compare switchers to always treated

- At periods 2 and 3, e's outcome = treated potential outcome, so

$$Y_{e,3} - Y_{e,2} = Y_{e,3}(1) - Y_{e,2}(1) = Y_{e,3}(0) + TE_{e,3} - (Y_{e,2}(0) + TE_{e,2}).$$

- On the other hand, group ℓ only treated at period 3, so

$$Y_{\ell,3} - Y_{\ell,2} = Y_{\ell,3}(0) + TE_{\ell,3} - Y_{\ell,2}(0).$$

Thus,
$$E\left[\text{DID}_{\ell-e}^{2-3}\right] = E\left[Y_{\ell,3} - Y_{\ell,2} - (Y_{e,3} - Y_{e,2})\right] = E\left[TE_{\ell,3} + TE_{e,2} - TE_{e,3}\right],$$

so $TE_{e,3}$ enters with negative weight in (2).

- Note: if $TE_{e,2} = TE_{e,3}$, $E[\text{DID}_{\ell-e}^{2-3}] = E\left[TE_{\ell,3}\right]$. More generally, if $TE_{g,t} = TE_{g,t'}$, no negative weights attached to $\widehat{\beta}_{fe}$. But restrictive!

$\widehat{\beta}_{fe}$ may compare "switching more" to "switching less"

- Suppose the treatment D is not binary or the design not staggered.
- Then, $\widehat{\beta}_{fe}$ may leverage DIDs comparing group m whose D increases more to group ℓ whose D increases less.
- In fact, with two groups m and ℓ and two periods,

$$\widehat{\beta}_{fe} = \frac{Y_{m,2} - Y_{m,1} - (Y_{\ell,2} - Y_{\ell,1})}{D_{m,2} - D_{m,1} - (D_{\ell,2} - D_{\ell,1})}. \quad (3)$$

- dCDH (2018) show that this "Wald-DID" estimator may not estimate convex combination effects, even if TE constant over time.

$\widehat{\beta}_{fe}$ may compare "switching more" to "switching less"

- E.g.: assume m goes from 0 to 2 units of treatment while ℓ goes from 0 to 1, and potential outcomes linear in treatment:

$$Y_{m,t}(d) = Y_{m,t}(0) + \delta_m d$$
$$Y_{\ell,t}(d) = Y_{m,t}(0) + \delta_\ell d,$$

with $\delta_\ell = 3\delta_m > 0$.

- Treatment effect constant over time, heterogeneous across groups, and no variation in treatment timing.

- Then, under // trends,

$$\begin{aligned} E[\widehat{\beta}_{fe}] &= E[Y_{m,2} - Y_{m,1} - (Y_{\ell,2} - Y_{\ell,1})] \\ &= E[Y_{m,2}(2) - Y_{m,1}(0) - (Y_{\ell,2}(1) - Y_{\ell,1}(0))] \\ &= E[Y_{m,2}(0) + 2\delta_m - Y_{m,1}(0) - (Y_{\ell,2}(0) + \delta_\ell - Y_{\ell,1}(0))] \\ &= -\delta_m < 0. \end{aligned}$$

Example: effect of newspapers on electoral turnout?

- Gentzkow et al. (2011) answer that question with 1868 to 1928 US data.
- Reg change in turnout from presidential election $t-1$ to t in county g on change in # newspapers and state-year FE. $\widehat{\beta}_{fd} = 0.0026$ (s.e.=0.0009).
- We estimate FE reg, and find $\widehat{\beta}_{fe} = -0.0011$ (s.e.=0.0011).
- $\widehat{\beta}_{fe}$ and $\widehat{\beta}_{fd}$ significantly different (t-stat=2.86), so under common trends, we reject constant treatment effect.
- 45.7% of weights attached to $\widehat{\beta}_{fd}$ negative, negative weights sum to -1.43.
- 40.1% of weights attached to $\widehat{\beta}_{fe}$ negative, negative weights sum to -0.53.
- Weights attached to $\widehat{\beta}_{fd}$ negatively correlated with the election year.

$\Rightarrow$ $\widehat{\beta}_{fd}$ biased if treatment effect changes over time.

Dynamic TWFE also not robust to heterogeneous effects

- With binary D and stagg. design, researchers estimate dynamic TWFE:

$$Y_{g,t} = \gamma_g + \lambda_t + \sum_{\ell=-K,\ell\neq-1}^{L} \beta_\ell 1\{F_g = t-\ell\} + \varepsilon_{g,t},$$

where F_g =period at which g becomes treated.

- For $\ell \geq 0$, β_ℓ supposed to estimate cumulative effect of $\ell+1$ treatment periods. For $\ell \leq -2$, $\beta_\ell =$ placebo.

Dynamic TWFE also not robust to heterogeneous effects

- Sun and Abraham (2021) show that under // trends,

$$E\left[\widehat{\beta}_{\ell}\right] = E\left[\sum_{g} w_{g,\ell} TE_g(\ell) + \sum_{\ell' \neq \ell} \sum_{g} w_{g,\ell'} TE_g(\ell')\right], \tag{4}$$

where $TE_g(\ell) =$ effect of $\ell + 1$ treatment periods in group g.

- 1st sum: weighted sum across groups of effect of $\ell + 1$ treatment periods, with possibly < 0 weights $\Rightarrow \widehat{\beta}_{\ell}$ **not robust to heterogeneous effects.**
- 2nd sum: weighted sum, across $\ell' \neq \ell$, of effects of $\ell' + 1$ treatment periods. $\Rightarrow \widehat{\beta}_{\ell}$ **contaminated by effects of $\ell' + 1$ treatment periods.**
- For $\ell \leq -2$, placebo coeffs $\widehat{\beta}_{\ell}$ also not robust to het. effects.
- `eventstudyweights` Stata package computes weights in (4).

More general results with dynamic TWFE

- Many applications do not have binary D and staggered designs.
- dCDH (2021a) consider two cases of interest:
  1. Distributed-lag regression for a binary $D_{g,t}$ with non-staggered designs.
  2. Regressions of $Y_{g,t+\ell}$ on g, t FE and $D_{g,t}$ (local projections, inspired by Jordà, 2005).
- In both cases, similar decompositions as above, with <0 weights in general.
- Actually, "local projections" may produce biased estimators even if treatment effects are homogenous!

TWFE with several treatments

- Consider TWFE with several binary treatments:

$$Y_{g,t} = \gamma_g + \lambda_t + \sum_{\ell=1}^{L} \beta_\ell D^\ell_{g,t} + \varepsilon_{g,t}.$$

- E.g.: $D^1_{g,t}$: whether US state g has a medical marijuana law in year t, $D^2_{g,t}$: whether US state g has a recreational marijuana law.

- dCDH (2021b) show that under // trends,

$$E\left[\widehat{\beta}_1\right] = E\left[\sum_{\substack{(g,t): \\ D^1_{g,t}=1}} w_{g,t}\, TE_{g,t}(1) + \sum_{\substack{(g,t): \\ D^{-1}_{g,t} \neq \mathbf{0}}} w_{g,t}\, TE_{g,t}(-1)\right], \qquad (5)$$

where $TE_{g,t}(1) = E[Y_{g,t}(1, D^{-1}_{g,t}) - Y_{g,t}(0, D^{-1}_{g,t})]$,
$TE_{g,t}(-1) = E[Y_{g,t}(0, D^{-1}_{g,t}) - Y_{g,t}(0, \mathbf{0})]$.

TWFE with several treatments

- In the 1st sum, $\sum w_{g,t} = 1$ but possibly $w_{g,t} < 0$, as in dCDH (2020).
- 2nd sum=contamination term, as in SA (2021).
- However, $\sum w_{g,t} \neq 0$ in general. We get $\sum w_{g,t} = 0$, as in SA (2021), if $L = 2$ or if treatments are mutually exclusive.
- `twowayfeweights` Stata and R package computes the weights in (5).
- Often adding more treatments exacerbate the issue of < 0 weights.
- Example (from Hotz & Xiao, 2011): effect of state center-based daycare regulations on the demand for family home daycare?
- Two treatments: minimum staff-child ratio and minimum years of schooling required to be the director of a center-based care.
- For the minimum years of schooling treatment,

$$\sum_{(g,t):D^{\ell}_{g,t}=1} w_{g,t} \mathbb{1}\left\{w_{g,t} < 0\right\} \simeq -9.02!$$

Outline

Estimators ruling out dynamic effects (dCDH, 2020)

- With a binary treatment, dCDH (2020) focus on the effect on switchers:

$$\delta^S = E\left[\frac{1}{N_S}\sum_{(g,t):D_{g,t}\neq D_{g,t-1}} N_{g,t}\left[Y_{g,t}(1)-Y_{g,t}(0)\right]\right].$$

- They propose to estimate δ^S by DID_M, a weighted average, across t, of two types of DIDs:
  - DID_+ compares the $t-1$ to t outcome evolution of groups going from untreated to treated and of groups untreated at both dates.
  - DID_- compares the $t-1$ to t outcome evolution of groups treated at both dates, and of groups going from treated to untreated.
- DID_+ relies on // trends assumption on untreated outcome $Y_{g,t}(0)$. DID_- relies on // trends assumption on treated outcome $Y_{g,t}(1)$.

Estimators ruling out dynamic effects (dCDH, 2020)

- DID_M computed by `did_multiplegt` Stata and R commands.
- dCDH also propose placebo tests of the two // trends assumptions.
- DID_M can easily be extended to discrete treatments.
- Example (Gentzkow et al., 2011, cont'd): $DID_M = 0.0043$ (s.e.=0.0015).
- 66% larger and significantly different from $\widehat{\beta}_{fd}$ at the 10% level (t-stat=1.77), has an opposite sign to $\widehat{\beta}_{fe}$.

No dynamic effects but several treatments (dCDH, 2021b)

- As above, DID_M aggregates DIDs comparing carefully chosen "treated" and "control" groups:
  - "Treated" g satisfy $D^{\ell}_{g,t} = 1 - D^{\ell}_{g,t-1}$ and $D^{-\ell}_{g,t} = D^{-\ell}_{g,t-1} = \mathbf{0}$;
  - "Control" g' satisfy $D^{\ell}_{g',t} = D^{\ell}_{g',t-1} = D^{\ell}_{g,t-1}$ and $D^{-\ell}_{g',t} = D^{-\ell}_{g',t-1} = \mathbf{0}$.
- DID_M can be computed by `did_multiplegt` Stata and R commands.
- Example (Hotz and Xiao, 2011, cont'd): for the minimum years of schooling treatment, $\text{DID}_\text{M} = -0.066$ (se=0.136),
- Significantly $\neq$ (t-test=2.25) from the TWFE coeff =-0.445 (se=0.167).

Dynamic effects, with a binary and staggered treatment

- With dynamic effects, group g's outcome at time t is allowed to depend on her past treatments.
- E.g., $Y_{g,t}(\mathbf{0}_{t-1},1)$: potential outcome if untreated until $t-1$, then treated at t.
- Callaway & Sant'Anna (2021) and SA (2021) replace the // trends assumption on $Y_{g,t}(0)$ by // trends assumption on $Y_{g,t}(\mathbf{0}_t)$.

CSA (2021) and SA (2021)

- With binary D and stagg. design, groups can be aggregated into cohorts that start receiving the treatment at the same period.
- CSA (2021) define parameters of interest as $TE_{c,c+\ell}$, ATE at period $c+\ell$ of cohort that started receiving treatment at period c.
- To estimate $TE_{c,c+\ell}$, they propose DID comparing $c-1$ to $c+\ell$ outcome evolution in cohort c and in never-treated groups.
- CSA (2021) also propose estimators of more aggregated effects: average effect of having been treated for $\ell+1$ periods.
- They also propose estimators using not-yet-treated as controls, and estimators relying on conditional parallel trends.
- Estimators computed by the `csdid` and `did` Stata and R commands.

Borusyak, Jaravel, and Spiess (2021)

- Borusyak, Jaravel, and Spiess (2021) have proposed alternative estimators.
- Obtained from TWFE regression of outcome on group and time FE, and dummies for every treated (g,t).
- Estimator of TE in treated cell (g,t): coeff on that cell's dummy.
- Under // trends and the assumptions of Gauss-Markov thm, linear unbiased estimator of TE in treated cell (g,t) with lowest variance.

$\Rightarrow$ More efficient than estimators of Callaway and Sant'Anna (2021).

⚠ The result requires in particular $\text{cov}(Y_{g,s}(0), Y_{g,t}(0)) = 0$ for any $s \neq t$.

- Not realistic in many cases, but BJS provide simulations that still show efficiency gains with modest serial correlation.
- Estimators computed by the `did_imputation` Stata package.

Understanding the difference between the two estimators

- With only one treated group s, which starts to receive treatment at period t_s, CSA's estimator of that group's effect at $t_s+\ell$ is:

$$Y_{s,t_s+\ell}-Y_{s,t_s-1}-\frac{1}{G-1}\sum_{g\neq s}\left(Y_{g,t_s+\ell}-Y_{g,t_s-1}\right),$$

while BJS' estimator is:

$$Y_{s,t_s+\ell}-\frac{1}{t_s-1}\sum_{k=1}^{t_s-1}Y_{s,k}-\frac{1}{G-1}\sum_{g\neq s}\left(Y_{g,t_s+\ell}-\frac{1}{t_s-1}\sum_{k=1}^{t_s-1}Y_{g,k}\right).$$

- CSA's estimator use groups' t_s-1 outcome as the baseline.
- BJS' estimator instead uses average outcome from period 1 to t_s-1, which is why it is more precise if $\text{cov}(Y_{g,s}(0),Y_{g,t}(0))=0$.

Estimators' exhibit $\neq$ biases if trends not exactly //

- If trends not exactly //, BJS' estimator is:
  - more biased if differential trends widen over time, as would happen with group-specific trends;
  - less biased if // trends fails due to anticipation effects just before t_s.

$\Rightarrow$ Overall, which estimator to use may depend on:

1. serial correlation (affecting the relative s.e.);
2. one's degree of confidence in // trends;
3. the type of violations of this assumption likely to arise.

Illustration: effects of unilateral divorce laws

- Between 1968 and 1988, 29 US states adopted a unilateral divorce law.
- Building upon Friedberg (1998), Wolfers (2006) studies the effects of those laws on divorce rates.
- He uses a parsimonious event-study regression.
- We revisit this application, considering a standard event-study regression and the new methods.

Illustration: effects of unilateral divorce laws

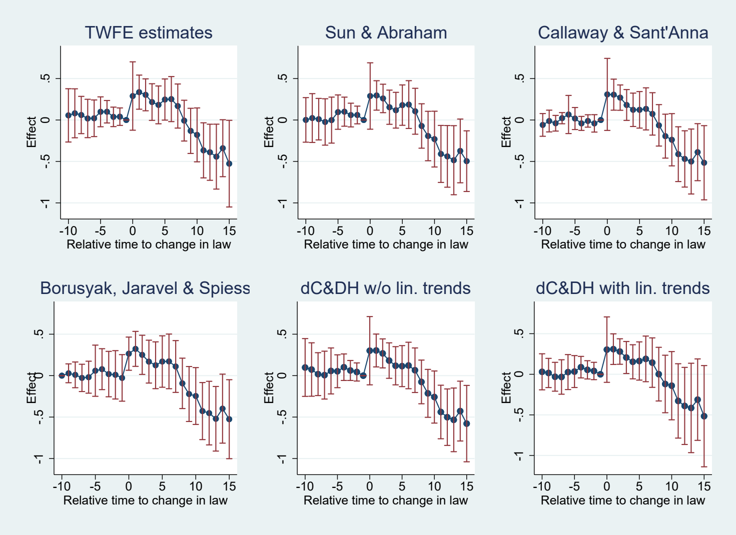

Illustration: effects of unilateral divorce laws

- Few differences between the different estimates.
- Adding linear time trends, as suggested by Friedberg (1998), does not affect the results.
- Differences on standard errors: BJS more (resp. less) precise for short- (resp. long-) run treatment effects.
- Summary:

Table 1: Average effect from 0 to 7 years after the law change

Wolfers (2006)	0.200 (0.056)
Event-study without binning pairs of years	0.249 (0.106)
BJS	0.198 (0.129)
dCDH, no linear trends	0.185 (0.107)
dCDH, linear trends	0.219 (0.096)

Dynamic effects with general $D_{g,t}$ (dCDH, 2021a)

- Focus on binary D below, but the idea extends to discrete, ordered D.
- We extend event-study approach, by redefining event as period F_g where a group's treatment changes for the first time.
- Let $\delta_{g,\ell} = E(Y_{g,F_g+\ell} - Y_{g,F_g+\ell}(D_{g,1},...,D_{g,1}))$.
- Difference b/w group g's actual outcome at $F_g + \ell$ and the counterfactual "status quo" outcome if treatment had remained equal to $D_{g,1}$.
- To estimate $\delta_{g,\ell}$, $\text{DID}_{g,\ell}$ compares $F_g - 1$-to-$F_g + \ell$ outcome evolution between group g and proper "control groups".
- In such groups g', $D_{g',1} = ... = D_{g',F_g+\ell}$ and $D_{g',1} = D_{g,1}$.

Dynamic effects with general $D_{g,t}$ (dCDH, 2021a)

- We aggregate the $\delta_{g,\ell}$ into δ_ℓ: effect of having experienced weakly higher treatment for $\ell+1$ periods.
- Leads to event-study graph, with distance to first treatment change on x-axis, δ_ℓ on the y-axis to the right of zero, placebos to the left.
- Magnitude of δ_ℓ may be hard to interpret, as the number of treatments for ℓ periods may vary.
- One can complement it with a "first-stage", by computing δ_ℓ^D.
- We can also define δ =weighted avg of δ_ℓ/ weighted avg of 1st-stage effects δ_ℓ^D.
- May be used to conduct a cost-benefit analysis comparing groups' actual treatments to the "status quo" scenario.
- Computed by the `did_multiplegt` Stata and R commands.

Illustration: banking deregulation & housing market

- In 1994, the Interstate Banking and Branching Efficiency Act allowed US banks to operate across states without formal authorization.
- 42 states lifted at least one restriction over 1993-2005.
- Favara and Imbs (2015) measure effect of banking deregulation on mortgages originated by banks and housing prices.
- They use 1993- 2005 county×year-level data and rely on a TWFE local projection.
- Treatment: number of regulations lifted in state s and year t.
- Outcomes: loan volume and housing prices.
- We compare our estimators with TWFE regressions.

Results: effects after ℓ periods.

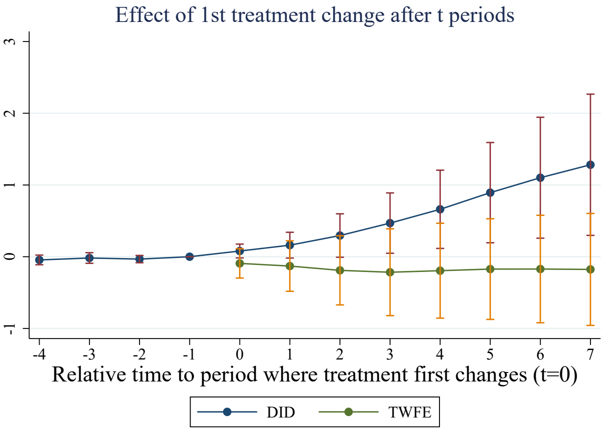

Figure 1: Effect of banking deregulations on loan volume.

Outline

Avenues for future research

- TWFE coeffs may not always estimate convex combination of effects. Then, could for instance be of different sign than every unit's effect.
- Literature has mostly focused on providing alternative estimators for binary and staggered treatments.

⇒ developing more estimators for non-binary and/or non-staggered treatments is a promising avenue.

- Also unclear whether researchers should completely abandon TWFE regs.
- Sometimes they estimate a convex combination of effects, and often have lower variance than heterogeneity-robust DID estimators.

⇒ A comparison of the MSE of TWFE and heterogeneity-robust DID in broad set of applications is another promising avenue.

Thank you!