brief memo: the problem with estimating change

Two Way Fixed Effects (TWFE) estimator is often widely used in classic Difference-in-Difference (DiD) setups. To estimate causal effects, social scientists often rely on DiD. Observing change between two groups and two time periods: in outcomes before and after a treatment, over the *same period,* and in a *control group* that did not receive the treatment.

Recent literature found the TWFE model, once thought to produce consistent average treatment effects, can lead to biased estimates. This is especially so when treatment effects change over time (heterogeneity) or have a lag (dynamics). In staggered designs, traditional two-way fixed-effects (TWFE) regression can assign negative weights to treatment effects, thereby obscuring their dynamic and heterogeneous patterns. In some cases, TWFE regression models can be biased even when treatment effect dynamics are homogeneous across different treatment groups/cohorts. Because DiD is commonly used to assess the impact of policy changes, DiD employing traditional TWFE estimators may not reflect the true changes accurately.

This is because the model confounds factors and mixes "early vs. late," "treated vs. never-treated," and "treated vs. treated" comparisons. The coefficient on a given lead or lag can be contaminated by effects from other periods, using already-treated units as invalid controls. Apparent pre-trends can arise solely from treatment effects heterogeneity, resulting in an uninterpretable average. To counter this, several alternative estimators have been proposed for robust treatment effect heterogeneity and staggered treatment designs. Some methods use disaggregation or re-weighting to focus on clean comparisons, thus avoiding negative weights that bias to TWFE model (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021). Others use an imputation-based estimator that construct counterfactuals for treated units (Borusyak et al., 2024). Other matrix completion approaches combine DiD with synthetic control methods, connecting the extended TWFE estimator with the difference-in-differences estimator.

To summarise, there is no one superior TWFE estimator. Each involves trade-offs. Personally I'd choose the Goodman - Bacon Decomposition to decompose the TWFE into four interpretable, distinct components. 1. Early vs Late, 2. Treated vs Untreated, 3. Late vs Early (bad controls), 4. Never Treated vs Untreated. Running the decomposition allows us to first assess the reliability of TWFE results and diagnose a next appropriate method. Thereafter we can use more robust estimators.

REFERENCES

1. *“A comparative analysis of two-way fixed effects estimators in staggered treatment designs”* (2025),Jhordano Aguilar-Loyo, Journal of Econometrics, Volume 251, September 2025, 106059

2. *“When Can We Use Two-Way Fixed-Effects (TWFE): A Comparison of TWFE and Novel Dynamic Difference-in-Differences Estimators”* (2025), Tobias Rüttenauer, Ozan Aksoy, UCL Social Research Institute, University College London, 55-59 Gordon Square, London WC1H 0NU, UK)

3. "Problems with two-way fixed-effects event-study regressions”, Brantly Callaway and Pedro H. C. Sant’Anna (2021), https://bcallaway11.github.io/did/articles/TWFE.html

4. “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects”, Sun & Abraham (2021), Journal of Econometrics, Volume 225, Issue 2, December 2021, Pages 175-199

5. “Difference-in-differences with variation in treatment timing” (2021), Andrew Goodman-Bacon, Journal of Econometrics, Volume 225, Issue 2, December 2021, Pages 254-277

6. https://chenxing.space/blog/notes-on-callaway-sant-anna-2021-staggered-adoption-did/

7. “How much should we trust staggered difference-in-differences estimates?”, Journal of Financial Economics, Andrew C. Baker, David F. Larcker, Charles C.Y. Wang, Volume 144, Issue 2, May 2022, Pages 370-395