When can we get away with using the two-way fixed effects regression?

Apoorva Lal

Netflix

Abstract

The widespread use of the two-way fixed effects regression motivated by folk wisdom that it uncovers the average treatment effect on the treated (ATT) has come under scrutiny recently due to recent results in applied econometrics showing that it fails to uncover meaningful averages of heterogeneous treatment effects in the presence of effect heterogeneity over time and across adoption cohorts. In this paper, we propose simple tests that can be used to test for differences in dynamic treatment effects over cohorts, which allows us to test for when the two-way fixed effects regression is likely to yield biased estimates of the ATT.

Keywords: difference in differences, panel data, heterogeneous treatment effects

1 Introduction

Consider a balanced panel-data setting with $i = 1, ..., N$ individuals observed over $t = 1, ..., T$ time periods. For each unit i, a binary treatment $w_{it} \coloneqq 1(t \geq g_i)$ is assigned at some adoption time $g_i \in \mathcal{G}$ where $\mathcal{G} \coloneqq \{T\} \cup \infty$ is the set of treatment adoption times and $g_i = \infty$ indicates a never-treated unit. We observe a scalar outcome $y_{it} = w_{it}y^1_{it} + (1 - w_{it})y^0_{it}$, where y^1_{it} and y^0_{it} are potential outcomes under treatment and control, respectively.

The following two-way fixed effects regression

$$y_{it} = \tau w_{it} + \alpha_i + \lambda_t + \varepsilon_{it}$$

is a workhorse regression in applied economics and adjacent fields for the estimation of causal effects in such settings. The estimand that researchers typically seek to estimate in panel data settings is the Average Treatment effect on the Treated (ATT) ($\mathbb{E}[y^1_{it} - y^0_{it} \mid w_{it} = 1]$), and researchers often interpret the coefficient on the treatment indicator, τ, as an estimate of the ATT. The above regression's dynamic ('event study') counterpart

$$y_{it} = \sum_{s \neq -1}^{T} \gamma_s \Delta^s_{it} + \alpha_i + \lambda_t + \varepsilon_{it}$$

where Δ^s_{it} is an indicator for the s-th period relative to the adoption time (which in turn is the first-difference of the treatment indicator), is also widely used to estimate the dynamic ATT.

When $g_i \in \{T_0, \infty\}$, the above regressions are unbiased estimates of the ATT under the assumption of parallel trends. However, when $g_i \in \{T_0, ..., T - 1\}$, the above regressions exhibit the 'negative weighting'/'contamination bias' problem (Goodman-Bacon (2021), Chaisemartin and D'Haultfœuille (2020), Goldsmith-Pinkham et al (2024)) the regression coefficient on the treatment indicator, τ, is a weighted average of the ATT over time and across cohorts, where the weights are functions of the treatment timing distribution and the dynamic treatment effect heterogeneity and can be negative for some cohorts. This implies that the two-way fixed effects regression can fail to

uncover meaningful averages of heterogeneous treatment effects over time and across adoption cohorts.

This has prompted a cambrian explosion of new estimators that aim to uncover the ATT in the presence of heterogeneous treatment effects over time and across adoption cohorts (Chaisemartin and D'Haultfœuille (2021), Roth et al (2022), Arkhangelsky and Imbens (2023) for reviews). Such heterogeneity-robust estimators typically involve estimating the ATT separately for each cohort using a more precise comparison between the treated cohort and a never-treated group, and then averaging these estimates to obtain an overall estimate of the ATT. While their consistency properties for the ATT are well understood, they are often computationally expensive and have higher variance than the two-way fixed effects regression. This motivates the primary focus of this paper: to develop simple tests that can be used to test for differences in dynamic treatment effects over cohorts, which allows us to test for when the two-way fixed effects regression is likely to yield biased estimates of the ATT. Heuristically, if the dynamic treatment effects are homogeneous over cohorts, then the two-way fixed effects regression is likely to yield unbiased estimates of the ATT that are considerably more precise than alternative estimators that typically discard more data in order to shut down the negative weighting problem.

To motivate the procedure, consider Figure 1 and Figure 2. In Figure 1, there are three adoption cohorts (plus a never-treated cohort - bottom panel), and all cohorts exhibit the same temporal heterogeneity pattern (the effect function is $\log(t)$ - top panel), and so the 2WFE event study (blue line in panel 2) is consistent for the true dynamic ATT (black line in panel 2). We can also consistently estimate the cohort-level ATTs with an appropriately saturated regression (Abraham and Sun (2020), Wooldridge (2021)) as shown in the third panel. In Figure 2, in contrast, we have the same three adoption cohorts, but the three cohorts exhibit radically different temporal heterogeneity: the first exhibits a linear decay down to zero, the second exhibits a log increase followed by zero, and the third exhibits sinusoidal effects. In this case, the 2WFE event study (blue line in panel 2) is not consistent for the true dynamic ATT (black line in panel 2); in fact, the estimated event study suggests a violation of the parallel trend assumption despite the treatments being randomized and thus parallel trends being true in the DGP. We can still estimate the cohort-level ATTs correctly with a saturated regression. The key insight is that testing for differences between a 'pooled' event study (the blue line in the second panel) and cohort X time interactions (that yield the cohort-level estimates in the hird panel) can help us distinguish between the two scenarios. This can be formulated as a joint F-test on the coefficients of the cohort X time interactions in a saturated regression.

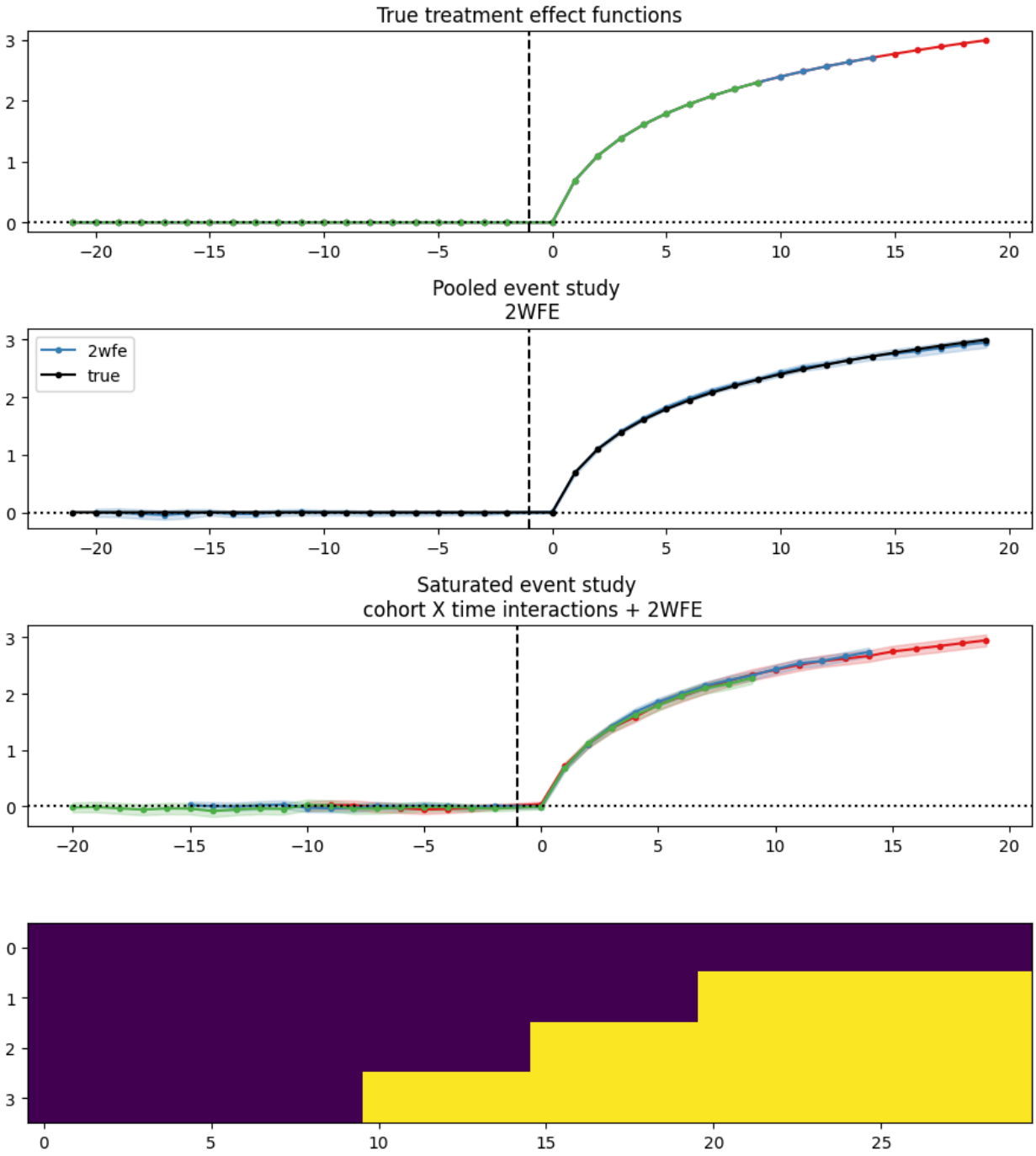

Figure 1: true and estimated effects from pooled and saturated event study regressions with homogeneous treatment effects across three cohorts

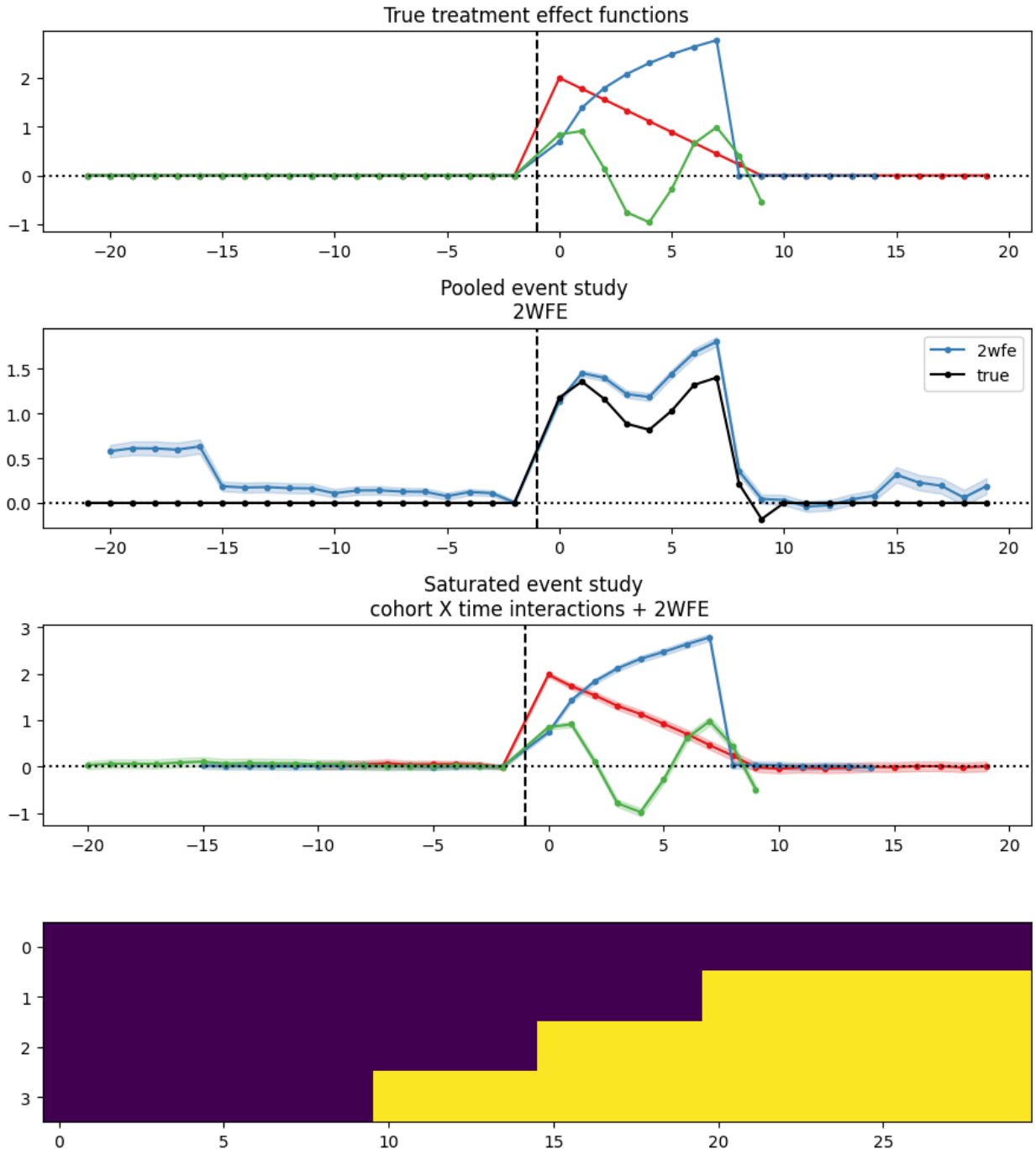

Figure 2: true and estimated effects from pooled and saturated event study regressions in a DGP with heterogeneous treatment effects across three cohorts

2 Methodology

We propose using a joint F-test on the following specification

$$y_{it} = \alpha_i + \lambda_t + \beta W_{it} + \sum_{s \neq -1}^{T} \gamma_s \Delta_{it}^s + \sum_{s \neq -1}^{T} \delta_s W_{it} \Delta_{it}^s + \varepsilon_{it}$$

References

Abraham, S. and Sun, L. (2020) Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects, *Journal of econometrics*

Arkhangelsky, D. and Imbens, G. (2023) Causal Models for Longitudinal and Panel Data: A Survey, *SSRN Electronic Journal*, http://www.nber.org/papers/w31942.pdf

Chaisemartin, C. de and D'Haultfœuille, X. (2020) Two-way fixed effects estimators with heterogeneous treatment effects, *The American economic review*, http://arxiv.org/abs/1803.08807

Chaisemartin, C. de and D'Haultfœuille, X. (2021) *Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey*, https://papers.ssrn.com/abstract=3980758

Goldsmith-Pinkham, P. S., Hull, P. and Kolesár, M. (2024) Contamination bias in linear regressions, *American Economic Review*

Goodman-Bacon, A. (2021) Difference-in-differences with variation in treatment timing, *Journal of econometrics*, https://www.sciencedirect.com/science/article/pii/S0304407621001445

Roth, J., Sant'Anna, P. H. C., Bilinski, A., et al (2022) What's trending in difference-in-differences? A synthesis of the recent econometrics literature, *arXiv [econ.EM]*, https://www.jonathandroth.com/assets/files/DiD_Review_Paper.pdf

Wooldridge, J. M. (2021) Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators, *Working paper*, http://dx.doi.org/