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. These tests are implemented as methods in the pyfixest python library.

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 $w_{it} := 1(t \geq g_i)$ is assigned at some adoption time $g_i \in \mathcal{G}$ where $\mathcal{G} := \{T\} \cup \infty$ 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

$$y_{it} = \tau w_{it} + \alpha_i + \lambda_t + \varepsilon_{it} \tag{1}$$

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, $\hat{\tau}$, as an estimate of the ATT. The above regression's dynamic ('event study') counterpart

$$y_{it} = \sum_{s \neq -1}^{T} \gamma_s \Delta^s_{it} + \alpha_i + \lambda_t + \varepsilon_{it} \tag{2}$$

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, $\hat{\tau}$, 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. The same is true for the event study coefficient vector γ.

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. Additionally, a large re-analysis of published work in political science by Chiu et al (2023) finds that they rarely overturn the conclusions of the two-way fixed effects regression, and are typically have considerably larger variance. 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. We provide a formal statement of this test in the next section, and show through simulation studies that this approach can detect cohort-level temporal heterogeneity in a variety of DGPs.

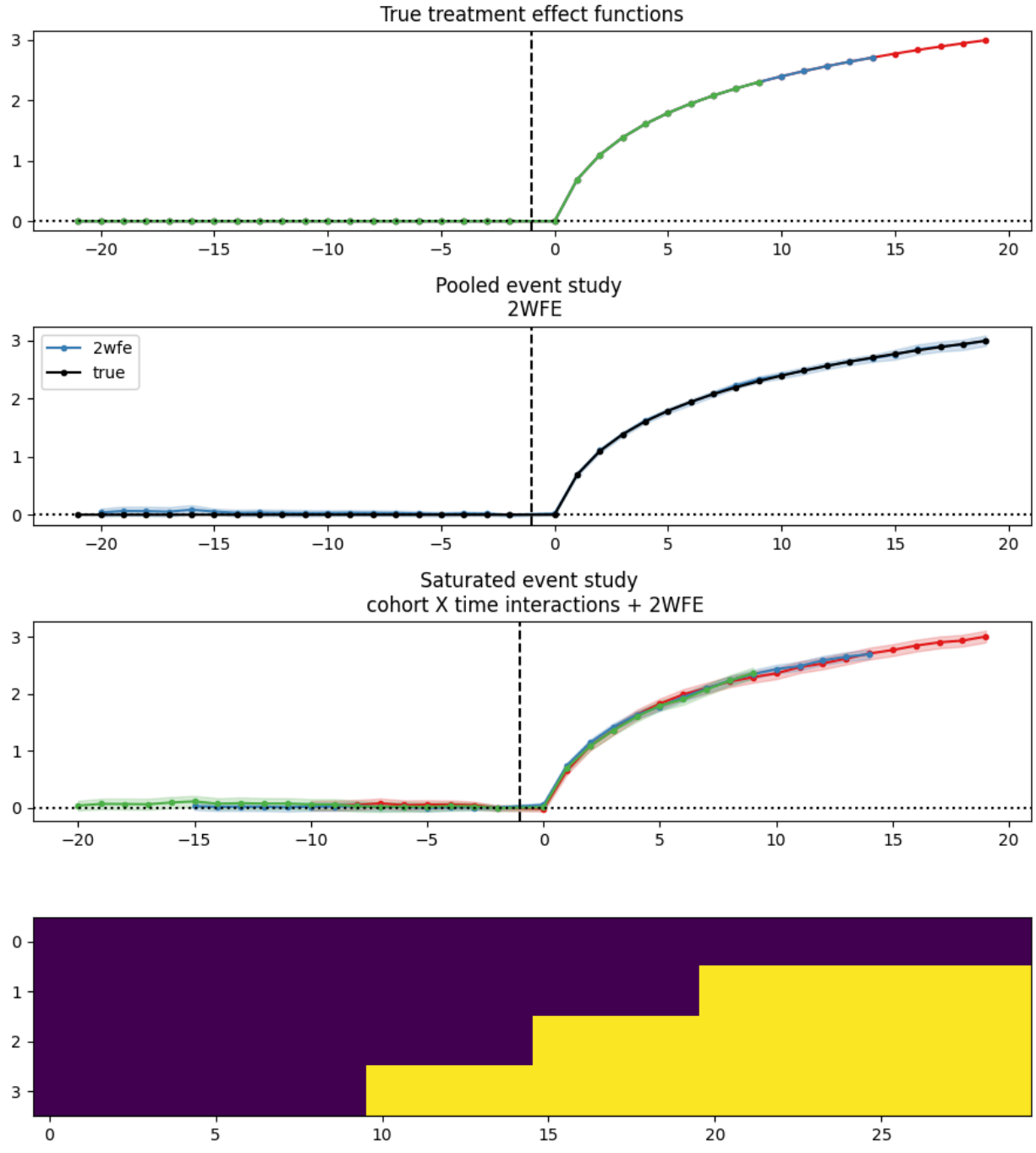

Figure 1: true and estimated effects from pooled and saturated event study regressions with homogeneous treatment effects across three cohorts

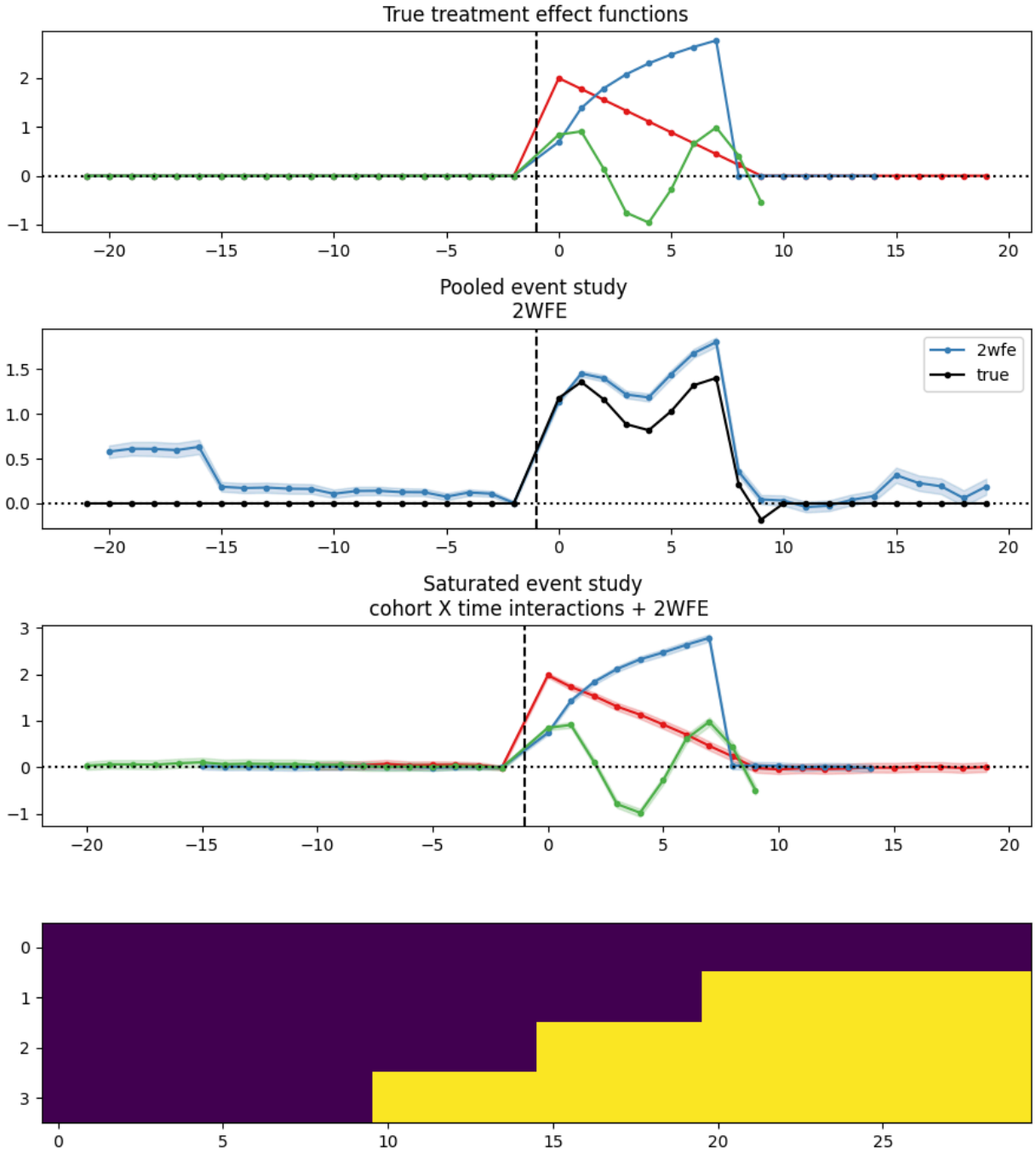

Figure 2: true and estimated effects from pooled and saturated event study regressions in a DGP with heterogeneous treatment effects across three cohorts

2 Methodology

Tests considered in the following section take the form of traditional joint tests of multiple linear restrictions, where the null hypothesis is that $\boldsymbol{R}\beta = \boldsymbol{q}$ where $\boldsymbol{R}$ is a $m \times k$ matrix of linear restrictions, β is a $k \times 1$ vector of coefficients, and q is a $m \times 1$ vector of constants. The test statistic is then

$$F = \frac{\left(\boldsymbol{R}\hat{\beta} - \boldsymbol{q}\right)' \left[\boldsymbol{R}\hat{\mathbb{V}}\right)\boldsymbol{R}']^{-1}\left(\boldsymbol{R}\hat{\beta} - \boldsymbol{q}\right)}{m} \sim F(m, n-k) \text{ under the null hypothesis} \tag{3}$$

where $\hat{\mathbb{V}}$ is the cluster-robust variance-covariance matrix of the coefficient estimates.[1] We consider two tests: one for testing for event study dynamics, and one for testing for heterogeneity in event study dynamics.

2.1 Testing for event study dynamics

As a warmup, consider a simple comparison between (1) and (2). The latter decomposes the ATT across time-periods. For the purposes of testing for event study dynamics, we only care about comparing the equality of the dynamic treatment effects after the treatment is assigned ($\{\gamma_t\}_{t=0}^{T}$) against the common ATT estimate τ. We can test the following null hypothesis

$$H_0 : \{\gamma_t\}_{t=0}^{T} = \hat{\tau} \text{ for all k} > 0 \tag{4}$$

by specifying $\boldsymbol{R} = \boldsymbol{I}_K$ as a $T_1 \times T_1$ identity matrix and $\boldsymbol{q} = (\hat{\tau}, ..., \hat{\tau})'$ as a T_1-vector of the restricted estimate ($\hat{\tau}$ from (1)).[2]

2.2 Testing for across-cohort heterogeneity in dynamic treatment effects

Next, we extend the approach outlined above to construct a test for across-cohort heterogeneity in dynamic treatment effects. A conventional method to estimate the cohort-level ATTs is to estimate the dynamic treatment effects separately for each cohort and then average these estimates to obtain an overall estimate of the ATT (Abraham and Sun (2020), Wooldridge (2021), Lal et al (2024)), which involves specifying the following regression

$$y_{it} = \alpha_i + \lambda_t + \sum_{g_i \in \mathcal{C} \setminus \infty} \sum_{s \neq -1}^{T} \mathbb{1}(g_i = c)\tau^{sc}\Delta_{it}^{s} + \varepsilon_{it} \tag{5}$$

This is a saturated event study that constructs cohort $\times$ time interactions for each adoption cohort (with $g_i = \infty$ never treated cohort) omitted and therefore recovers the cohort-level event studies. These coefficients are reported in the third panel in Figure 1 and Figure 2, and correctly uncover the true cohort-level ATTs in the presence of arbitrary heterogeneous treatment effects across cohorts (top panel). The downside of this approach, however, are twofold. First, these regressions can get unwieldy with many cohorts, and the number of parameters grows linearly with the number of cohorts. Second, the cohort level ATTs are self-contained and therefore constructing a test for equality across multiple cohorts is not straightforward. Instead, one may re-specify the saturated event-study regression (5) as follows:

$$y_{it} = \alpha_i + \lambda_t + \sum_{s \neq -1}^{T} \gamma_s \Delta_{it}^{s} + \sum_{c \in \mathcal{C}} \sum_{s \neq -1}^{T} \delta_s \Delta_{it}^{cs} + \varepsilon_{it} \tag{6}$$

(6) returns numerically identical estimates of the cohort-level dynamic ATT as (5) (illustrated for the Figure 2 dgp in Figure 7), but it allows us to test for differences in dynamic treatment effects over cohorts more easily. This is because unlike (5), (6) contains a common event study coefficient vector, and cohort-level deviations, which in turn can be jointly tested against the null of zero.

[1] This can be implemented using either a χ^2 or F test; the distinction between the two is due to different degrees of freedom that disappear for realistic sample sizes

[2] this can equivalently be formulated by testing for the equality of adjacent elements of $\boldsymbol{\gamma}$, e.g. $\gamma_1 = \gamma_2$ by specifying $\boldsymbol{R}$ that contains rows like $[1, -1, 0, ..., 0]$ and $q = [0, ..., 0]$.

We show in the next section that this test is consistent for the null hypothesis of homogeneous dynamic treatment effects over cohorts, and that it has power against a variety of alternatives. As a concrete example, the joint $p-$value for the cohort $\times$ time interactions in Figure 1 is 0.11, while the joint p-value for the cohort $\times$ time interactions in Figure 2 is 0.000. Thus, we can reject the null hypothesis of homogeneous dynamic treatment effects in Figure 2 but not in Figure 1, which is consistent with the underlying DGP. In the next section, we show through simulation studies that this test has good power to detect across-cohort heterogeneity in dynamic treatment effects in a variety of DGPs.

3 Simulation Studies

3.1 Testing for event study dynamics

To begin, we perform simulation studies based on to study the properties of the testing procedure described in Section 2.1. We consider the simple setting with a single adoption cohort where the treatment effects follow one of the following seven DGPs visualised in Figure 3.

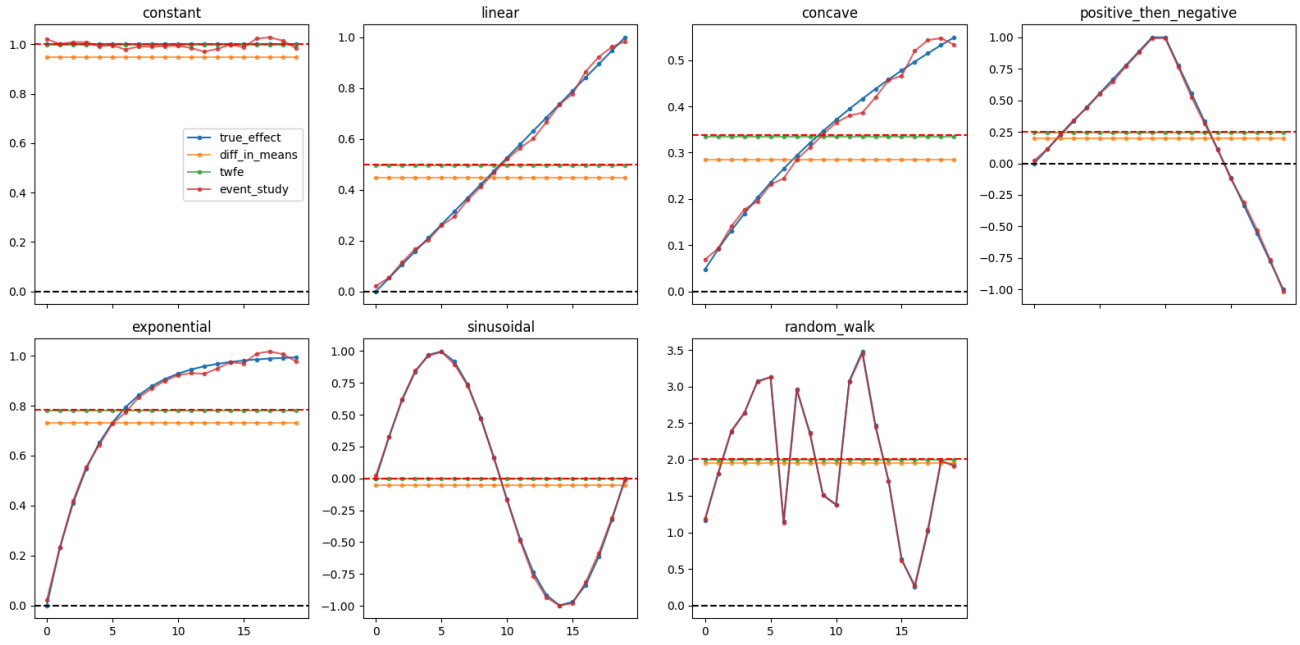

Figure 3: true treatment effect functions and estimates from difference in means, static, and dynamic two-way fixed effects regressions

The first DGP has constant effects over time, while the others have varying degrees of temporal heterogeneity. We simulate 500 replications of the data for each DGP, and compute the rejection rate of the joint test for dynamic treatment effects outlined in the previous section. We report the rejection rate and p-value distribution in Figure 4. We find that the rejection rate for the constant DGP (null) is under the nominal level of $\alpha = 0.05$, while the rejection rates for the other DGPs considerably higher. The rejection rate for concave effects is considerably lower, although this is likely due to the fact that the the treatment effects do actually tail off in later time periods and the static effect captures this well.

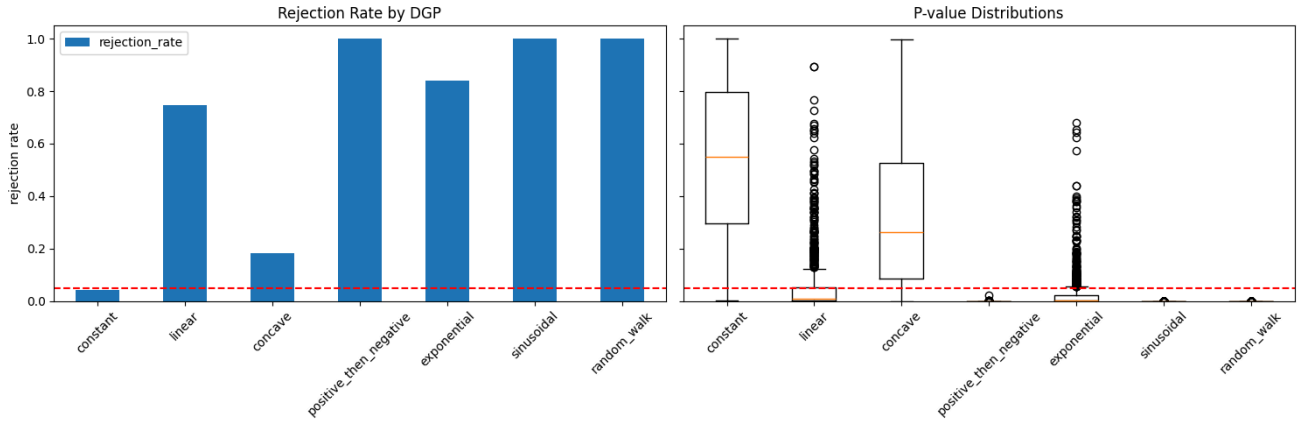

Figure 4: Rejection rates over 500 replications for the joint test of dynamic treatment effects using an F-test in DGPs from Figure 3

3.2 Testing for across-cohort heterogeneity in dynamic treatment effects

Next, we we perform simulation studies based on to study the properties of the testing procedure described in Section 2.2. Here, we consider four different DGPs with homogeneous and heterogeneous treatment effect functions across cohorts as illustrated in Figure 5. In addition to the two DGPs described in the previous section, we also consider DGPs with 'mild' heterogeneity that applies a scaler multiplier to the concave (log) effect function in Figure 1.

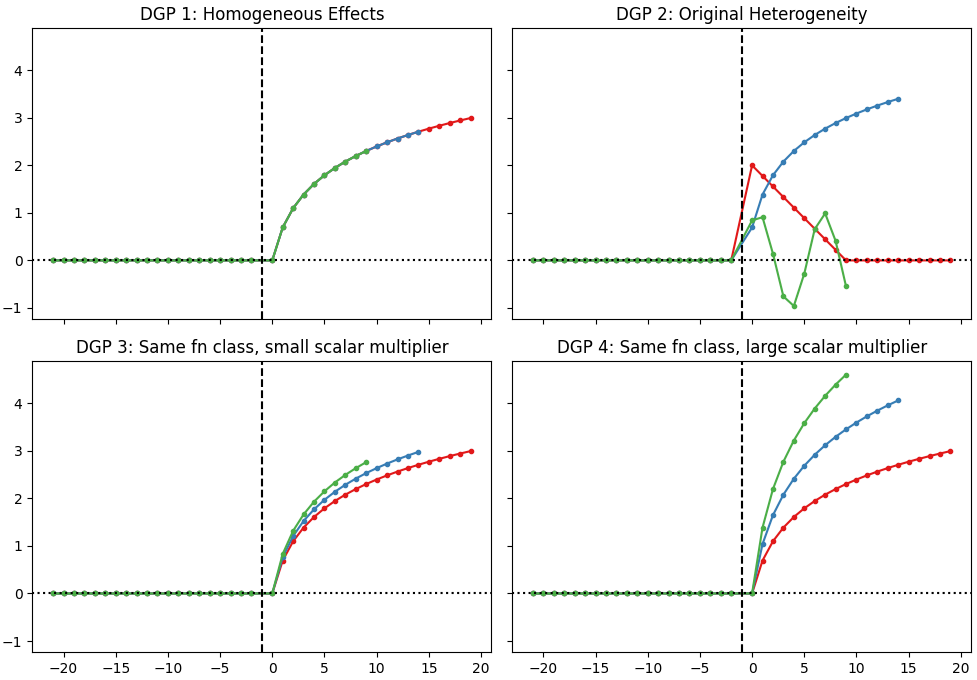

Figure 5: true cohort level effect functions: the first DGP has homogeneous effects across cohorts, while others have heterogeneous effects of varying complexity

For each DGP, we simulate 500 replications of the data, and compute the rejection rate of the joint test for cohort-level coefficients outlined in the previous section. We report the rejection rate and

p-value distribution in Figure 6. We find that the rejection rate for the homogeneous DGP (null) is under the nominal level of $\alpha = 0.05$, while the rejection rates for heterogeneous DGPs are close to 1. This suggests that the test has good power to detect across-cohort heterogeneity in dynamic treatment effects.

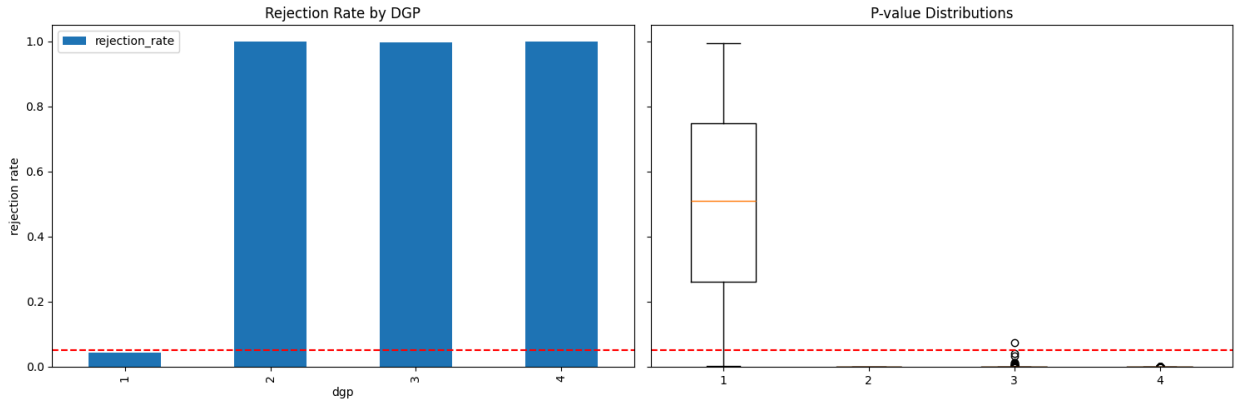

Figure 6: Rejection rates over 500 replications for the joint test of cohort-level coefficients using an F-test in DGPs from Figure 5

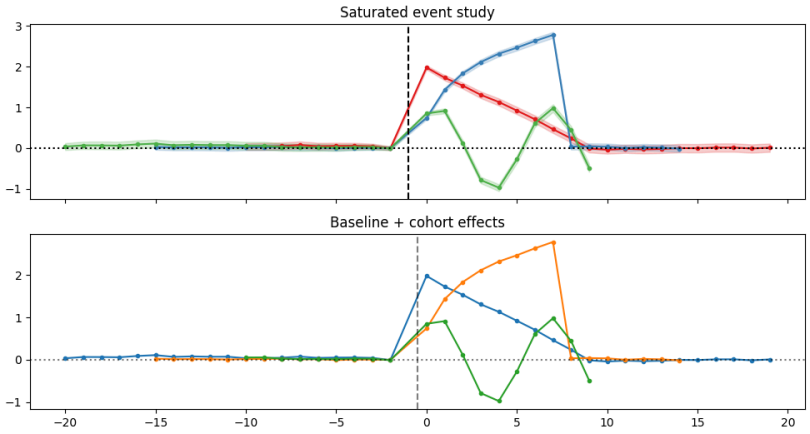

Figure 7: The two event study specifications return identical coefficients in the presence of arbitrary heterogeneity.

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

Chiu, A., Lan, X., Liu, Z., et al (2023) What to do (and not to do) with causal panel analysis under parallel trends: Lessons from a large reanalysis study, *arXiv preprint arXiv:2309.15983*

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

Lal, A., Fischer, A. and Wardrop, M. (2024) Large Scale Longitudinal Experiments: Estimation and Inference, *arXiv preprint arXiv:2410.09952*

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/