University of Münster
Centre for Applied Economic Research Münster
Institute for Public and Regional Economics

Master Thesis

Difference-in-Differences Estimation with Variation in Treatment Timing and Heterogeneous Treatment Effects

Submitted by: Niels Wich
Field of Study: Economics M.Sc.

Professor: Prof. Dr. Nadine Riedel
Advisor: Dr. Tobias Böhm

Issue date: 17/05/2021
Closing date: 19/10/2021

Contents

List of Figures

List of Tables

1 Introduction

Difference-in-Differences (DiD) is arguably one of the most widely applicable quasi-experimental research designs for causal inference from observational data. The common principle behind all versions of DiD designs is to compare the evolution of an outcome over time between units with different treatment histories. It's then feasible to uncover causal parameters under relatively weaker assumptions compared to standard methods.

The simplest case where the DiD design is applicable is a setting with binary treatment (e.g. a change in law) and two groups of units observed in two periods, where both groups are untreated in the first period and only one group is treated in the second period. Under the so called 'common trends assumption', the average treatment effect of the treated can be uncovered fully non-parametrically from an estimand of four group-time population averages, the '2x2 DiD estimand'. In settings with more than two periods and/ or treatment groups, the DiD design is usually implemented by a 'two-way fixed effects' (TWFE) regression, a linear regression specification that includes a single indicator for each unit/ group, each time period, and a set of treatment variables. The TWFE regression can be seen as a natural extension of the 2x2 DiD estimand and is numerically identical to the 2x2 DiD estimand in the two groups and two periods setting. In more general settings the TWFE regression however implicitly imposes additional parametric assumptions on treatment effects. Researchers usually assumed that TWFE regressions would be robust to this kind of misspecification or imposed the necessary assumptions. A recently emerging literature shows however, that this practice might be more critical than previously thought (Borusyak et al. (2021), de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2021), Sun and Abraham (2020), Strezhnev (2018), Imai and Kim (2021), Gardner (2021)). Coefficients from TWFE regressions generally do not uncover 'reasonably' weighted averages of treatment effects, if necessary homogeneity assumptions are violated. In response, there have been several attempts to develop robust estimators (de Chaisemartin and D'Haultfœuille (2020), Borusyak et al. (2021), Gardner (2021), Sun and Abraham (2020), Callaway and Sant'Anna (2020), Wooldridge (2021)). This quickly evolving literature has accordingly received a lot of attention from applied researchers.

Many of these existing research papers are still unpublished or have been very recently accepted for publication and new papers appear regularly. Up to date it's therefore not fully clear to what extent the theoretical limitations of TWFE regressions translate into noticeable differences in practice. There is also no consensus yet, which of the many alternative approaches will be the future 'go-to' approach for applied researchers as a stand-alone method or robustness check. The goal of this thesis is therefore to summarize, compare, and evaluate the existing research about this topic. In particular, I also want to consider the perspective of an applied researcher that usually has to deal with different challenges than a theoretical econometrician.

The structure of this thesis is aligned to this goal: Section 2 introduces and motivates the DiD research design. I start with the canonical two groups and two periods scenario and generalize from there to more complex settings. This is mostly meant as a gentle introduction to the topic and helps to build some foundations I will heavily rely on later. Section 3 then deals with the shortcomings of TWFE regressions if treatment

effects are heterogeneous. I address three widely used regression specifications: 'Static' TWFE specifications include a single indicator variable for a binary treatment. I derive a result that decomposes the treatment coefficient into a weighted sum of treatment effects and also try to give some intuition for the sign and magnitude of treatment effect weights. If the treatment is binary and units can only join the treatment but can't leave it, researchers often replace the single treatment indicator by a set of indicator variables for the relative time to the initial treatment period. These 'dynamic' TWFE specifications are usually motivated by the desire to study treatment dynamics over time. I present an analogous decomposition result following Sun and Abraham (2020) together with an intuition for the arising problems. Last but not least, I also characterize the problems that additionally emerge in settings with non-binary treatment variables. To conclude this section, I present some simulation results that visually illustrate how the level of treatment effect heterogeneity and other parameters affect the regression coefficient(s). Section 4 introduces, compares, and evaluates alternative estimators which are robust to heterogeneous treatment effects. In particular I consider the estimators of de Chaisemartin and D'Haultfœuille (2020), Borusyak et al. (2021), Gardner (2021), Sun and Abraham (2020), and Callaway and Sant'Anna (2020). In Section 5 I apply the alternative estimators and the TWFE estimator to real world data and compare the results. Specifically, I use data from Gallagher (2014) who studies the effect of regional flood events on flood insurance take-up in the US with the dynamic TWFE regression. This exercise also serves as a way to explore practical limitations of the alternative approaches that might not be obvious beforehand. Section 6 concludes.

2 Difference-in-Differences Design

2.1 General Setting, Notation, and Terminology

I consider a large population of N units $i \in \{1, ..., N\}$ observed in T time periods $t \in \{1, ..., T\}$, i.e. a panel balanced in calendar time. Hence, there are NT unit-time observations $(i, t) \in \{1, ..., N\} \times \{1, ..., T\}$. For each (i, t) the quantitative outcome variable Y_{it} and the treatment variable D_{it} can be observed. Furthermore I partition the units into G groups $g \in \{1, ..., G\}$, where all units within a group experience the same treatment history $(D_{i1}, ..., D_{iT})$. Group membership of unit i is denoted by g_i. Because of this, I use the notation D_{g_it} equivalently to refer to the treatment status of unit i that belongs to group g in period t. The treatment variable can be either binary or quantitative. In the binary treatment case, I set $D_{g_it} \in \{0, 1\}$ for all (i, t): $D_{g_it} = 1$ if observation (i, t) is treated and $D_{g_it} = 0$ if the observation is untreated. A binary treatment history is called 'staggered', if for a given group g it holds that for all $t \in \{2, ..., T\} : D_{g_it} \geq D_{g_it-1}$. Put differently, a group which receives treatment once remains treated until the last period. Binary treatment settings without this restriction, the treatment can switch on and off arbitrarily, may be called 'non-staggered'.

Throughout this section and the section thereafter I rely on the potential outcomes framework introduced by Rubin (1974) and modifications of it to formalize treatment effects.

2.2 The Generic 2x2 Difference-in-Differences Design

The simplest case where a DiD design is applicable is the setting with $T = G = 2$ and a binary treatment variable. Units in group $g = 1$ are untreated in both periods and units in group $g = 2$ are untreated in the first but treated in the second period. Following Rubin (1974), I define two potential outcomes for each (i, t): A potential outcome under no treatment $Y_{it}(0)$ and a potential outcome under treatment $Y_{it}(1)$. The natural object of interest for causal analysis is the treatment effect, i.e. the difference between the treated and untreated potential outcomes: $\rho_{it} := Y_{it}(1) - Y_{it}(0)$. Potential outcomes are mapped to the observed outcome by: $Y_{it} = D_{it}Y_{it}(1) + (1 - D_{it})Y_{it}(0)$. Put simply, for treated observations the treated potential outcome is observed and for untreated observations the untreated potential outcome is observed. Because only one of the potential outcomes is observable for each observation, it's impossible to directly compute treatment effects. This constitutes the 'fundamental problem of causal inference'. However, under additional assumptions it may be still possible to uncover aggregates of treatment effects, if units' unobserved potential outcomes can be imputed by observed outcomes of observations with the opposing treatment status. In the setting at hand, one could be interested in the average treatment effects of group 2 in period 2: $E(Y_{i2}(1) - Y_{i2}(0)|g_i = 2)$. Note that $E(Y_{i2}(1)|g_i = 2)$ is observed since group 2 is treated in period 2 and $E(Y_{i2}(0)|g_i = 2)$ is subsequently unobserved. Therefore, it's now necessary to impute the unobserved average to find the quantity of interest.

A first approach could be to simply replace it by the average observed outcome of the untreated units in period 2, $E(Y_{i2}|g_i = 1)$:

$$\begin{aligned} &E(Y_{i2}|g_i = 2) - E(Y_{i2}|g_i = 1) \\ &= E(Y_{i2}(1) - Y_{i2}(0)|g_i = 2) + [E(Y_{i2}(0)|g_i = 2) - E(Y_{i2}(0)|g_i = 1)] \end{aligned}$$

The first term on the right-hand side is the parameter of interest and the second term is the so called 'selection bias' term. This strategy is valid, as long as on average the untreated potential outcomes in period 2 are identical in both groups. Often times this identification strategy will however fail, e.g. if the groups systematically differ in unobserved characteristics that influence the outcome variable ('selection on unobservables').

A second approach could be to exploit the longitudinal structure instead, by replacing $E(Y_{i2}(0)|g_i = 2)$ by $E(Y_{i1}|g_i = 2)$:

$$E(Y_{i2}|g_i = 2) - E(Y_{i1}|g_i = 2) = E(Y_{i2}(1) - Y_{i2}(0)|g_i = 2) + E(Y_{i2}(0) - Y_{i1}(0)|g_i = 2)$$

Similarly, the approach only identifies the parameter of interest, if on average the second group's untreated potential outcome does not change over time. This condition could be violated, if unobserved characteristics that influence the outcome variable systematically differ over time in group 2.

The first strategy only exploits cross-sectional variation, but not the panel structure of the data and the second strategy only exploits the longitudinal dimension but not the variation between groups. By combining both approaches, one can finally impute the unobserved average by $E(Y_{i1}|g_i = 2) + E(Y_{i2} - Y_{i1}|g_i = 1)$ instead:

$$\begin{aligned} &E(Y_{i2} - Y_{i1}|g_i = 2) - E(Y_{i2} - Y_{i1}|g_i = 1) \qquad (1) \\ &= E(Y_{i2}(1) - Y_{i2}(0)|g_i = 2) + [E(Y_{i2}(0) - Y_{i1}(0)|g_i = 2)) - E(Y_{i2}(0) - Y_{i1}(0)|g_i = 1)] \end{aligned}$$

Now the parameter of interest is identified, if the average change of the untreated potential outcome is identical in both groups, i.e. $E(Y_{i2}(0) - Y_{i1}(0)|g_i = 2) = E(Y_{i2}(0) - Y_{i1}(0)|g_i = 1)$. This condition is known as the 'common trends assumption under no treatment' (CTA), which is the key identifying assumption of the DiD design. Estimand (1) is known as the '2x2 DiD estimand'. The 2x2 DiD estimand relaxes the identification conditions of the cross-sectional and the longitudinal approaches: Differences between groups in the untreated potential outcomes are allowed, as long as they are time-invariant, since differencing eliminates level differences. Note that the cross-sectional approach would already fail in this scenario. Both groups are also allowed to experience changes of the untreated potential outcome over time which are group-invariant. In this case the longitudinal approach would fail. Put differently, the 2x2 DiD estimand allows to identify the parameter of interest, if there are additive time and group effects but rules out effects that depend on group and time simultaneously.

Instead of computing the 2x2 DiD estimand directly from four conditional averages, it's also feasible to use an Ordinary Least Squares (OLS) population regression instead. Generally, the conditional expectation function (CEF) is identical to the population regression function if it is linear (cf. Theorem 3.1.4 in Angrist and Pischke (2009)). When there are only discrete variables in the conditioning set, the CEF can be always thought of as linear by treating each unique combination of variable values as a single variable. By specifying a 'saturated' regression, where the number of coefficients is identical to the number of unique variable value combinations in the conditioning set of the CEF, it's then feasible to identify the CEF fully non-parametrically. Here $(g, t) \in \{1, 2\}^2$ and therefore it's necessary to have four coefficients in the population regression. It turns out that δ_3 from the following saturated population regression identifies the 2x2 DiD estimand (1):

$$Y_{it} = \delta_0 + \delta_1 I(g_i = 2) + \delta_2 I(t = 2) + \delta_3 I(g_i = 2) I(t = 2) + \epsilon_{it} \tag{2}$$

To see a bit more clearly why exactly this is the case, consider the underlying OLS minimization problem:

$$(\delta_0, \delta_1, \delta_2, \delta_3)' = \arg\min \sum_{(i,t)} [Y_{it} - d_0 - d_1 I(g_i = 2) - d_2 I(t = 2) - d_3 I(g_i = 2) I(t = 2)]^2$$

$$= \arg\min \{ \sum_{i:g_i=1} (Y_{i1} - d_0)^2 + \sum_{i:g_i=2} (Y_{i1} - d_0 - d_1)^2 + \sum_{i:g_i=1} (Y_{i2} - d_0 - d_2)^2 + \sum_{i:g_i=2} (Y_{i2} - d_0 - d_1 - d_2 - d_3)^2 \}$$

It's easy to verify that the first term of the second line becomes minimal, if $d_0 = E(Y_{i1}|g_i = 1)$. Even though d_0 also appears in the other three terms, it's not necessary to take them into account for the minimization problem, because they all have at least one additional parameter to fully offset any d_0 value. By the same principle the second and the third terms become minimal if $d_1 = E(Y_{i1}|g_i = 2) - E(Y_{i1}|g_i = 1)$ and $d_2 = E(Y_{i2} - Y_{i1}|g_i = 1)$. Finally, the last term becomes minimal if d_3 is equal to the 2x2 DiD estimand (1).

The population regression representation (2) of the CEF is particularly attractive to transfer the specification to a random sample setting, since it allows to obtain standard error estimates easily. Note however that specification (2) is not the only way to identify the CEF non-parametrically by a population regression. One could for instance specify the following population regression:

$$Y_{it} = \gamma_0 I(t = 1, g_i = 1) + \gamma_1 I(t = 1, g_i = 2) + \gamma_3 I(t = 2, g_i = 1) + \gamma_4 I(t = 2, g_i = 2) + \epsilon_{it}$$

Because the regression specification includes a unique indicator for each $(g,t) \in \{1,2\}^2$, the regression coefficients are by construction identical to the respective values of the CEF. While it's a bit more straightforward to understand the saturated regression principle based on this specification, it's no longer possible to identify the 2x2 DiD estimand by a single coefficient.

2.3 Two-Way Fixed Effects Regression

Often times researchers want to uncover treatment effects in settings with more than two time periods and with several treatment groups, i.e. $T > 2$ and $G \geq 2$. To do so, it's common to use the 'two-way fixed effects' (TWFE) regression, a regression specification that includes a set of treatment variables and additionally a single indicator for each group/ unit and each time period. The TWFE population regression with group indicators and a single treatment variable takes the following form:

$$Y_{it} = \hat{\alpha}_{g_i} + \hat{\beta}_t + \mu D_{g_i t} + \epsilon_{it} \tag{3}$$

If the treatment variable is binary, the specification is often times called a 'static' TWFE regression specification. Strictly speaking it's not feasible to include a single indicator for each group and for each time period, since all group and time indicators sum to one respectively, what causes perfect multicollinearity. This problem can be circumvented, by excluding one group and one time indicator and adding an intercept instead. However, for the ease of notation one usually writes down the specification as in (3) and only denotes the indicator coefficients which are 'active' for a given (i,t).

Note that the interaction term in the 2x2 DiD regression (2) can be replaced by the treatment indicator, since only units in group 2 and period 2 receive treatment in that setting. By acknowledging this, it's clear that the 2x2 DiD regression is identical to the TWFE regression (3) in the binary treatment setting with two groups and two periods ('2x2 setting').

In practice, it's however more common to use unit instead of group indicators in the TWFE regression specification. It turns out that in the 2x2 setting, the treatment coefficient from the TWFE specification is still numerically identical to the 2x2 DiD estimand. By the Frisch-Waugh-Lovell (FWL) theorem, δ_3 from the 2x2 DiD regression (2) can be obtained by the following two-step procedure:[1]

1. Regress $D_{g_i t}$ and $I(t=2)$ on an intercept and $I(g_i = 2)$ respectively and obtain the residuals $\check{D}_{g_i t}$ and $\check{I}(t=2)$.

   Since both variables are binary, the regressions are saturated and hence they match the associated CEFs: $E(D_{g_i t}|g_i = 1) = 0$ and $E(D_{g_i t}|g_i = 2) = 0.5$. $E(I(t=2)|g_i = 1) = E(I(t=2)|g_i = 2) = 0.5$, since all units are observed in

[1] See e.g. Chapter 2.4 of Davidson and MacKinnon (2003) for a linear algebra based proof of the FWL theorem.

both periods. Hence:

$$\check{D}_{g_it} = \begin{cases} 0, \text{ if } g_i = 1 \\ -0.5, \text{ if } g_i = 2 \text{ and } t = 1 \\ 0.5, \text{ if } g_i = 2 \text{ and } t = 2 \end{cases}$$

$$\check{I}(t=2) = \begin{cases} -0.5, \text{ if } t = 1 \\ 0.5, \text{ if } t = 2 \end{cases}$$

2. Regress Y_{it} on $\check{D}_{g_it}$ and $\check{I}(t=2)$. The coefficient of $\check{D}_{g_it}$ is numerically identical to δ_3 from the 2x2 DiD regression (2).

Equivalently for the TWFE regression, partial out the unit indicators from D_{g_it} and $I(t=2)$ to obtain the residuals $\hat{D}_{g_it}$ and $\hat{I}(t=2)$. The TWFE coefficient of D_{g_it} is then identical to the coefficient of $\hat{D}_{g_it}$ in a regression of Y_{it} on $\hat{D}_{g_it}$ and $\hat{I}(t=2)$. Since treatment status only varies by group, it holds for all observations within a group that $\hat{D}_{g_it} = \check{D}_{g_it}$. Because each unit is observed in both periods it also holds that $\hat{I}(t=2) = \check{I}(t=2)$. The FWL representations of both regressions are therefore identical and the treatment coefficient of the TWFE regression with unit indicators subsequently also identifies the 2x2 DiD estimand. By the same line of argumentation one can show that TWFE regressions with unit indicators always yield the same treatment coefficients as TWFE regressions with group indicators (3), as long as the treatment trajectories of all units within a group are identical.

3 TWFE Regressions with Heterogeneous Treatment Effects

Until now I have addressed TWFE regressions (beyond the 2x2 setting) from a purely mechanical angle, without stating assumptions about the underlying potential outcomes and without giving causal interpretations to regression coefficients. This section addresses the causal interpretation of coefficients in different TWFE specifications, when treatment effects are heterogeneous.

3.1 Static TWFE Regression

Consider the TWFE regression specification (3) in a setting with binary treatment, where the panel length and the number of treatment groups is flexible.

Potential Outcomes and CTA

The definitions of potential outcomes and treatment effects in this section carry over from the previous section. It's also straightforward to generalize the CTA from the 2x2 setting. For all $(\bar{g}, \bar{t}) \in \{1, ..., G\} \times \{2, ..., T\}$ assume that:

$$E[Y_{i\bar{t}}(0) - Y_{i\bar{t}-1}(0) | g_i = \bar{g}] = E[Y_{i\bar{t}}(0) - Y_{i\bar{t}-1}(0)]$$

In plain words, the average change in the untreated potential outcome between all pairs of consecutive periods is assumed to be identical in all groups.[2] Note that this assumption also implies common trends for all pairs of periods which are further apart than one period. If $G = T = 2$, the formulation coincides with the CTA introduced in the previous section.

The CEF and the Population Regression Function

In the 2x2 setting there is, by assumption, only one group that receives treatment in the second period. $E(\rho_{i2}|g_i = 2)$ is therefore identical to the average treatment effect of all treated observations (ATT), $ATT := E(\rho_{it}|D_{g_it} = 1)$. In the generalized setting there are however 2^T feasible treatment histories with a binary treatment. The ATT is therefore potentially an average of treatment effects from multiple groups and time periods. I refer to the average treatment effect of a group $\bar{g}$ in a period $\bar{t}$ as a 'group-time specific average treatment effect' (GATE): $GATE_{\bar{g},\bar{t}} := E(\rho_{i\bar{t}}|g_i = \bar{g})$. The question arises, whether the TWFE regression (3) still identifies the ATT in the generalized setting as it is the case in the 2x2 setting.

Without making further assumptions the CTA implies the following CEF:

$$\begin{aligned} E(Y_{i\bar{t}}|g_i = \bar{g}) &= \alpha_{\bar{g}} + \beta_{\bar{t}} + GATE_{\bar{g},\bar{t}} D_{\bar{g}\bar{t}} \\ &= \alpha_{\bar{g}} + \beta_{\bar{t}} + ATT\ D_{g_it} + [GATE_{\bar{g},\bar{t}} - ATT] D_{g_it} \end{aligned} \tag{4}$$

$\alpha_{\bar{g}}$ and $\beta_{\bar{t}}$ represent the additive group and time effects of the CEF, which are implied by the CTA. The second line follows from adding and subtracting $(ATT\ D_{g_it})$.

As already mentioned in the previous section, it generally holds that the population regression function coincides with the CEF if it's linear. The CEF (4) is however generally not linear in group indicators, time indicators and a single treatment indicator because $GATE_{\bar{g},\bar{t}}$ can vary across different $(\bar{g}, \bar{t})$. A TWFE regression specification with a single treatment indicator therefore only identifies the CEF, if $GATE_{\bar{g},\bar{t}} = ATT$ for all $(\bar{g}, \bar{t})$ with $D_{\bar{g},\bar{t}} = 1$. This is illustrated by the second line in (4) (Gardner, 2021). In the generalized setting, identification of the ATT therefore requires homogeneity of GATEs across all $(\bar{g}, \bar{t})$ what is, by construction, not necessary in the 2x2 setting. If there is only one treated $(\bar{g}, \bar{t})$, this condition is also satisfied in the generalized setting. Without this assumption, it's only known that the population regression function is the best linear predictor of the CEF in a mean squared error sense (cf. Theorem 3.1.5 in Angrist and Pischke (2009)). Unfortunately, this does not allow to make a general statement about how different treatment effects influence the treatment coefficient and whether it's still a 'reasonable' summary measure of the treatment effects in the treated stratum of the panel.

A Decomposition Result

In order to answer this question, I now derive a way to decompose the treatment coefficient μ from the (static) TWFE population regression (3) into a weighted sum of

[2] I generally use indices $\bar{g}$ and $\bar{t}$ instead of g and t to refer to a specific group and time period and thereby to avoid ambiguous notations. E.g.: $E(Y_{it})$ denotes the average of Y_{it} over all (i, t), whereas $E(Y_{i\bar{t}}) = E(Y_{it}|t = \bar{t})$ denotes the average in a specific period $t = \bar{t}$.

GATEs. de Chaisemartin and D'Haultfœuille (2020, Theorem 1) and Borusyak et al. (2021, Proposition 2) present identical decomposition results.

By the 'regression CEF theorem' (cf. Theorem 3.1.6 in Angrist and Pischke (2009)), replacing Y_{it} by the CEF in the TWFE regression (3) yields the same population regression function:

$$E(Y_{i\bar{t}}|g_i = \bar{g}) = \hat{\alpha}_{\bar{g}} + \hat{\beta}_{\bar{t}} + \mu D_{\bar{g}\bar{t}} + \hat{\epsilon}_{\bar{g}\bar{t}}$$

Note that the residual in this regression is identical for all observations in a given $(\bar{g}, \bar{t})$. From the FWL theorem, it follows that μ in the above regression can be equivalently obtained by first regressing D_{g_it} on the group and time indicators,

$$D_{g_it} = \tilde{\alpha}_{g_i} + \tilde{\beta}_t + \tilde{D}_{g_it}$$

, and then regressing the CEF on the residuals from this auxiliary regression $\tilde{D}_{g_it}$:

$$E(Y_{i\bar{t}}|g_i = \bar{g}) = \mu \tilde{D}_{\bar{g}\bar{t}} + \tilde{\epsilon}_{\bar{g}\bar{t}}$$

Multiplying both sides by $\tilde{D}_{\bar{g}\bar{t}}$ and building the sum over all observations yields:

$$\sum_{(\bar{g},\bar{t})} |\bar{g}| \tilde{D}_{\bar{g}\bar{t}} [E(Y_{i\bar{t}}|g_i = \bar{g})] = \sum_{(\bar{g},\bar{t})} |\bar{g}| \tilde{D}_{\bar{g}\bar{t}} [\mu \tilde{D}_{\bar{g}\bar{t}} + \tilde{\epsilon}_{\bar{g}\bar{t}}]$$

, where $E(Y_{i\bar{t}}|g_i = \bar{g})$ is defined as in (4) and $|\bar{g}|$ is the number of units in group $\bar{g}$. Since $\tilde{\epsilon}_{g_it}$ is the residual in a regression where $\tilde{D}_{g_it}$ is a regressor, the variables are by construction uncorrelated. Analogously, the group and time indicators are orthogonal to $\tilde{D}_{g_it}$, since it's the residual of the FWL auxiliary regression from above. After some rearrangement μ can be expressed as:

$$\begin{aligned} \mu &= \frac{\sum_{(\bar{g},\bar{t})} |\bar{g}| \tilde{D}_{\bar{g}\bar{t}} GATE_{\bar{g},\bar{t}} D_{\bar{g}\bar{t}}}{\sum_{(\bar{g},\bar{t})} |\bar{g}| \tilde{D}^2_{\bar{g}\bar{t}}} \\ &= \sum_{(\bar{g},\bar{t}):D_{\bar{g}\bar{t}}=1} w_{\bar{g}\bar{t}} GATE_{\bar{g},\bar{t}} \end{aligned} \tag{5}$$

, where $w_{\bar{g}\bar{t}} := \frac{|\bar{g}| \tilde{D}_{\bar{g}\bar{t}}}{\sum_{(\bar{g},\bar{t})} |\bar{g}| \tilde{D}^2_{\bar{g}\bar{t}}}$. μ is therefore a weighted sum of the GATEs of all treated (g, t). Furthermore the weights sum up to one, since:

$$\begin{aligned} \sum_{(g,t):D_{gt}=1} w_{gt} &= \frac{\sum_{(i,t)} \tilde{D}_{g_it} D_{g_it}}{\sum_{(i,t)} \tilde{D}^2_{g_it}} \\ &= \frac{\sum_{(i,t)} \tilde{D}_{g_it} (\tilde{\alpha}_{g_i} + \tilde{\beta}_t + \tilde{D}_{g_it})}{\sum_{(i,t)} \tilde{D}^2_{g_it}} \\ &= \frac{\sum_{(i,t)} \tilde{D}^2_{g_it}}{\sum_{(i,t)} \tilde{D}^2_{g_it}} \end{aligned}$$

The second line results from replacing D_{g_it} by its auxiliary regression representation. The third line follows from the fact that $\tilde{D}_{g_it}$ is the residual in the aforementioned

auxiliary regression representation and is subsequently uncorrelated with the group and time indicators. Due to this property, the treatment coefficient in the TWFE regression (3) with a binary treatment can be interpreted as a weighted average of GATEs of all treated (g, t).

An alternative, more general, decomposition result in terms of average observed outcomes instead of average treatment effects is given by:

$$\mu = \sum_{(\bar{g},\bar{t})} w_{\bar{g}\bar{t}} E(Y_{i\bar{t}}|g_i = \bar{g}) \tag{6}$$

This representation simply follows by not exploiting the orthogonality property of $\tilde{D}_{g_i t}$ with respect to the group and time indicators. The weights are defined in the same way as in (5). Of course, $\sum_{(g,t)} w_{gt} = 0$, since OLS residuals sum up to zero by construction.

Note that the ATT is itself a group size weighted average of GATEs of all treated (g, t):

$$ATT = \sum_{(\bar{g},\bar{t}):D_{\bar{g}\bar{t}}=1} \omega_{\bar{g}\bar{t}} GATE_{\bar{g},\bar{t}} \tag{7}$$

, where the weights are given by: $\omega_{\bar{g}\bar{t}} := \frac{|\bar{g}|}{\sum_{(\bar{g},\bar{t})} |\bar{g}| D_{\bar{g}\bar{t}}}$. The TWFE coefficient would generally identify the ATT if the weights in the decompositions (5) and (7) would be identical, i.e. $w_{gt} = \omega_{gt}$ for all treated (g, t). To see why this is not the case, it's necessary to further characterize the weights in the decomposition result (5), which are a function of $\tilde{D}_{g_i t}$ and $|g|$. Note that by the so called 'within' transformation (which is an application of the FWL theorem) it holds for balanced panels that:

$$\tilde{D}_{\bar{g}\bar{t}} = D_{\bar{g}\bar{t}} - E(D_{\bar{g}t}) - E(D_{g_i\bar{t}}) + E(D_{g_i t})$$

$E(D_{\bar{g}t})$ is the share of periods under treatment of a specific group $\bar{g}$, $E(D_{g_i\bar{t}})$ is the share of treated units in a given period $\bar{t}$, and $E(D_{g_i t})$ is the share of treated observations in the whole panel. For all $(\bar{g}, \bar{t}) : E(D_{\bar{g}t}), E(D_{g_i\bar{t}}) \in [0, 1]$ and $E(D_{g_i t}) \in (0, 1)$, since OLS mechanics (and also causal identification) require that at least one group is treated at some point in time and simultaneously not all groups are always treated. Even though $D_{g_i t} \in \{0, 1\}$, only the treated state is relevant here, since average treatment effects of untreated (g, t) are not weighted into the TWFE regression coefficient.

Using these findings and the fact that the denominator of the weights is a positive constant, it's possible to further characterize the TWFE coefficient based on the decomposition (5): The weight of a treated (g, t) is small (large), if a large (small) share of units is treated in that period and the group is treated in many (few) periods. More intuitively, the weights also depend positively on the group size. In a staggered adoption setting, this is equivalent to saying that GATEs from the end of the panel and from groups which receive their initial treatment early have smaller weights. GATEs can even receive negative weights, if $E(D_{\bar{g}t}) + E(D_{g_i\bar{t}}) > 1 + E(D_{g_i t})$. One can construct examples where this is indeed the case: Assume that there are only two equally sized groups, observed in three periods. Group $g = 1$ is untreated in the first and treated in the second and third period, whereas group $g = 2$ is only treated in the last period. In this setting:

$$\mu = GATE_{1,2} + (-0.5)GATE_{1,3} + (0.5)GATE_{2,3}$$

It follows from the above stated arguments that the TWFE regression coefficient is in fact not generally identical to the ATT, because the weights do not only depend on the group size but also on the groups' treatment histories. If however $GATE_{\bar{g},\bar{t}} = ATT$ for all treated $(\bar{g}, \bar{t})$ the TWFE regression coefficient still identifies the ATT, because it's then possible to draw the $GATE_{\bar{g},\bar{t}}$ terms in (5) out of the sum. This is intuitive, since the population regression function fully resembles the CEF in this case as already shown earlier (see equation (4)). Theoretically it would be however still possible that the TWFE regression coefficient coincides with the ATT even under treatment effect heterogeneity if, for some reason, the biases offset each other.

Building an Intuition for GATE Weights

Now I want to describe two ways to intuitively think about the weights of GATEs in the decomposition result (5):

The first way is to have a closer look at the auxiliary regression of the treatment indicator on group and time indicators. Practically, this regression is a linear probability model. The linear probability model is well known for its tendency to predict probabilities below zero and above one, what is the reason why researchers often times prefer to use logit or probit regressions instead. Given that the sign of a weight is determined by the associated residuals from this regression, it is clear that a negative weight arises if the predicted treatment probability is larger than one. From this perspective, the negative weighting problem arises because OLS is designed to provide the best linear prediction (in a mean squared error sense) of the outcome variable based on the regressors. This also explains why the GATE from a group, which is treated in many periods and from a period with many treated units is more likely to receive a negative weight, since OLS exploits the information to predict the treatment status. However, this perspective is yet not too intuitive, since the auxiliary regression is only a regression implied by the original TWFE specification.

Alternatively, there is a very intuitive way to link the regression weights back to 2x2 DiD estimands. Consider the two groups and three periods example from above. The decomposition of μ in terms of average observed outcomes (6) then yields:

$$\begin{aligned}\mu = (-0.5)E(Y_{i1}|g_i = 1) + E(Y_{i2}|g_i = 1) + (-0.5)E(Y_{i3}|g_i = 1)\\ +(0.5)E(Y_{i1}|g_i = 2) + -E(Y_{i2}|g_i = 2) + (0.5)E(Y_{i3}|g_i = 2)\end{aligned}$$

and after some rearrangement:

$$\begin{aligned}\mu = 0.5\{[E(Y_{i2}|g_i = 1) - E(Y_{i1}|g_i = 1)] - [E(Y_{i2}|g_i = 2) - E(Y_{i1}|g_i = 2)]\}+\\ 0.5\{[E(Y_{i3}|g_i = 2) - E(Y_{i2}|g_i = 2)] - [E(Y_{i3}|g_i = 1) - E(Y_{i2}|g_i = 1)]\}\end{aligned}$$

Subsequently it is feasible to interpret the TWFE regression coefficient as a weighted average of two 2x2 DiD estimands: Under the CTA, the first estimand identifies $GATE_{1,2}$ and, without making further assumptions, the second estimand identifies $GATE_{2,3} - [GATE_{1,3} - GATE_{1,2}]$. Given that, in contrast to the first estimand, the control group is treated in both periods, the second difference does not solely reflect the time trend but also the difference in the group's GATEs between both periods.

Only under the condition that the first group's GATEs are identical in both periods, the estimand in fact identifies $GATE_{2,3}$. Figure 1 depicts the example. The red line segment marks the case where GATEs of group 1 are heterogeneous. Since $GATE_{1,3} > GATE_{1,2}$ in the figure, the 2x2 DiD estimand that uses group 1 as a control group underestimates $GATE_{2,3}$. Since both decomposition results are equivalent, it's easy to verify that the above decomposition implies the previously derived weights for GATEs. Now it becomes pretty clear, where the negative weight for $GATE_{1,3}$ and the large weight for $GATE_{1,2}$ come from. It's the result of an incorrectly identified time trend, that leads to a subtraction of $GATE_{1,3}$ and an addition of $GATE_{1,2}$ in the second estimand. Another important observation is, that even in the case where GATEs are homogeneous over time, the TWFE regression coefficient does not even explicitly 'try' to weight in $GATE_{1,3}$. Intuitively, this is not the case since it's impossible to build a suitable 2x2 DiD estimand: For the first difference one could only exploit the different treatment states of group 1 between periods 1 and 3. However, group 2 also has diverging treatment states in the same periods and therefore is not a valid control group.

To further enhance the intuition for the weights of GATEs, consider an extension of the previous example with an always treated, also equally sized, third group. The implied weights on average observed outcomes and GATEs are given by:

t	$g_i = 1$	$g_i = 2$	$g_i = 3$	w_{1t}	w_{2t}	w_{3t}
1	0	0	1	(-0.5)	(0)	0.5
2	1	0	1	0.5	(-0.5)	0
3	1	1	1	0	0.5	-0.5

After some rearrangement, it's again possible to write the TWFE regression coefficient as an average of now six 2x2 DiD estimands:

$$\begin{aligned}
\mu = 1/6\{&[E(Y_{i2}|g_i = 1) - E(Y_{i1}|g_i = 1)] - [E(Y_{i2}|g_i = 2) - E(Y_{i1}|g_i = 2)]\} \\
+1/6\{&[E(Y_{i2}|g_i = 1) - E(Y_{i1}|g_i = 1)] - [E(Y_{i2}|g_i = 3) - E(Y_{i1}|g_i = 3)]\} \\
+1/6\{&[E(Y_{i3}|g_i = 1) - E(Y_{i1}|g_i = 1)] - [E(Y_{i3}|g_i = 3) - E(Y_{i1}|g_i = 3)]\} \\
+1/6\{&[E(Y_{i3}|g_i = 2) - E(Y_{i2}|g_i = 2)] - [E(Y_{i3}|g_i = 1) - E(Y_{i2}|g_i = 1)]\} \\
+1/6\{&[E(Y_{i3}|g_i = 2) - E(Y_{i2}|g_i = 2)] - [E(Y_{i3}|g_i = 3) - E(Y_{i2}|g_i = 3)]\} \\
+1/6\{&[E(Y_{i3}|g_i = 2) - E(Y_{i1}|g_i = 2)] - [E(Y_{i3}|g_i = 3) - E(Y_{i1}|g_i = 3)]\}
\end{aligned}$$

The TWFE coefficient consists of all 'admissible' 2x2 DiD estimands, that would be suitable to identify GATEs in case of (group specific) time constant GATEs: Since group 3 is always treated, it's impossible to build an 2x2 DiD estimand to identify one of its GATEs, because this requires different treatment states in two points in time. For the same reason the group is, in principle, well suited as a control group. Average outcomes of group 3 therefore never appear in the first difference, but four times in the second difference of a 2x2 DiD estimand. The first result implies that GATEs of group 3 generally receive smaller weights. The second result can push its weights in both directions, depending on whether the associated average observed outcome appears as the first or the second element of the second difference. If it appears as the first element, the weight becomes smaller and if it appears as the second element, the weight becomes larger. Since no other group receives treatment in period 1, this period is only

used as a contrast to identify GATEs of other groups in periods 2 and 3 with group 3 as a control group (three times). Therefore, $GATE_{3,1}$ gets a relatively large weight due to the fact that the average observed outcome only appears as the second element of the second difference. In period 3, all groups are treated and there are three 2x2 DiD estimands that use group 3 as a control group. Subsequently, the average observed outcome of group 3 in period 3 appears as the first element of the second difference, what results in a smaller weight for $GATE_{3,3}$. The same line of argumentation explains why the weight of $GATE_{1,2}$ is larger than the weight of $GATE_{1,3}$: Group 1 is used as a control group to build a 2x2 DiD estimand for the identification of $GATE_{2,3}$ (in the fourth line). The larger weight of $GATE_{2,3}$ compared to $GATE_{1,3}$ also stems from the fact, that group 2 is treated less often and therefore there are more contrasts to build 2x2 DiD estimands (three vs. one 2x2 DiD estimands).

The described mechanisms rationalize, why weights are smaller if a group is treated in many periods throughout the panel and if at a point in time a lot of units receive treatment. Strezhnev (2018, Proposition 1) shows that the decomposition in terms of 2x2 DiD estimands can be extended to all settings with staggered adoption. Goodman-Bacon (2021, Theorem 1) proposes an alternative decomposition result for staggered adoption settings, where DiD estimands are build from the average outcomes over each group's complete pre- and post treatment periods. Unfortunately, up to date, there is no decomposition result in terms of DiD estimands that also applies to non-staggered adoption settings. However, the basic intuition for the weights is still plausible in this case.

3.2 Dynamic TWFE Regression

In binary treatment settings with staggered adoption researchers usually not only want to identify an ATT-type parameter, but they may also be interested in treatment effect dynamics with respect to the initial treatment period. To do so, it's common to replace the single treatment dummy of the static TWFE specification by a set of indicator variables, which indicate the time difference to the initial treatment period. Those specifications are usually referred to as 'dynamic' TWFE specifications.

The classical application domain of dynamic TWFE specifications are evaluations of legislation amendments where jurisdictions adopt the new law at different points in time. Stevenson and Wolfers (2006) exploit the different timing of divorce law reforms across the US to study how the establishment of a unilateral divorce rule affects domestic violence, spousal homicides, and suicides. Here, it seems a priori reasonable to assume that the effects of the policy are not instantaneous but slowly build up over time (and this is what the authors find). In many cases, dynamic specifications are even sensible if the treatment is transient by nature, but the treatment effect evolves dynamically over time. Gallagher (2014) studies for instance the effect of regional flood events on flood insurance take-up in the US. While the actual event is limited to a single time period, he finds that insurance take-up spikes in the year after a flood and then steadily declines to a baseline level over a time horizon of about nine years. Hence, it makes sense to code all periods after the flood event as treated periods.

Potential Outcomes and Terminology

In order to account for dynamic treatment effects in the staggered adoption setting, it makes sense to further specify the treatment groups and adjust the definition of potential outcomes slightly.

Since all units that receive their initial treatment in the same period experience the same treatment history, they belong to the same treatment group. For that reason it's also valid to use the initial treatment period as a group label: $g_i = \min\{t : D_{it} = 1\}$. Units that don't receive treatment until period T ('never-treated' units) are labelled by $g_i = \infty$. I make the assumption that never-treated units don't receive treatment in periods $t > T$ and that 'always-treated' units ($g_i = 1$) don't already receive treatment in periods $t < 1$. Based on this definition for g_i, I define the 'event-time' as the time difference to a unit's (/ group's) initial treatment at a given point in time $s := (t - g_i)$. Additionally, I define for each (i, t) the sequence of dummy variables $D_{g_i t,r} := I(t - g_i = r)$, where r takes on values between $-(T-1)$ and $(T-1)$. $D_{g_i t,r} = 1$ if unit i at time t experienced its initial treatment r periods ago. Of course, for a given (i, t) only $D_{g_i t,s} = 1$ and for never-treated units $D_{g_i t,r} = 0$ for all event-times.

For each (i, t), I do no longer define only two potential outcomes based on the potential treatment status at a given point in time (treated or untreated): Instead of the treated potential outcome, I define T potential outcomes based on the potential initial treatment period $\tau \in \{1, ..., T\}$, $Y_{it}(\tau)$. It's the outcome that would be realized in period t if unit i receives its initial treatment in period τ. I also define a 'never-treated' potential outcome $Y_{it}(\infty)$, i.e. the outcome that would be realized if the unit would never be treated. $Y_{it}(\infty)$ is analogous to the untreated potential outcome in the classical potential outcomes framework. The CTA can be imposed for $Y_{it}(\infty)$ in the same way as for $Y_{it}(0)$. Treatment effects are then defined as the difference between the period τ potential outcome and the never-treated potential outcome:

$$\rho_{it}(\tau) := Y_{it}(\tau) - Y_{it}(\infty), \tau \in \{1, ..., T\}$$

In principle it would be also possible to define treatment effects with respect to any other potential outcome, but the never-treated potential outcome seems to be the most natural choice here. Sun and Abraham (2020) and Callaway and Sant'Anna (2020) also use the never-treated potential outcome as a contrast to define treatment effects. Note that treatment effects are also defined for $t < \tau$ what allows to accommodate anticipatory treatment effects. Herein lies the main advantage of this modified potential outcomes framework in comparison to the classical framework. Anticipatory behaviour is plausible, if the treatment is announced or foreseeable. Büttner and Madzharova (2021) study for instance the effect of pre-announced VAT changes on unit sales and find sizeable anticipatory effects.

Of course, for each (i, t) only the potential outcome where τ is the true initial treatment period g_i is observed. Potential and observed outcomes are therefore related through:

$$Y_{it} := Y_{it}(\infty) + \sum_{r=-(T-1)}^{T-1} D_{g_i t,r}[Y_{it}(t - r) - Y_{it}(\infty)]$$

The treatment effect associated with the observed outcome, $\rho_{it}(g_i)$, is the 'realized' treatment effect and is denoted by $\rho_{it} := \rho_{it}(g_i)$. $\rho_{i(g_i+r)}$ then refers to the realized treatment effect of unit i, r periods after its initial treatment. The average of all realized treatment effects of a given timing group $\bar{g}$ at event-time r is given by $GATE_{\bar{g},\bar{g}+r} = E(\rho_{i(g_i+r)}|g_i = \bar{g})$.

Dynamic TWFE Specification

The natural identification target in settings where event-time dependent treatment effects are plausible is the average of all realized treatment effects in event-time r, $ATE_r := E(\rho_{it}|D_{g_it,r} = 1)$, with $r \in \{-(T-1), ..., (T-1)\}$. To achieve this for (a subset of) all event-times simultaneously, it's common to regress the outcome variable on group and time indicators and (a subset of) the above defined event-time indicators.

The dynamic TWFE population regression could then read:

$$Y_{it} = \hat{\alpha}_{g_i} + \hat{\beta}_t + \sum_{r=-K}^{-2} \mu_r D_{g_it,r} + \sum_{r=0}^{R} \mu_r D_{g_it,r} + \epsilon_{it} \tag{8}$$

$R, K \in \{1, ..., T-1\}$ are respectively the largest and smallest event-times which are represented by indicator variables. μ_r denotes the population regression coefficient associated with event-time r.

It's necessary to exclude at least one event-time indicator in the dynamic TWFE specification. Otherwise the event-time indicators would sum up to one for all (i,t) where $g_i \neq \infty$ and to zero for the never-treated group's (i,t). This would make the event-time indicators perfectly multi-collinear to a linear combination of the $g_i \neq \infty$ group indicators. Here the one-period-ahead indicator is omitted, what is common in the applied literature. A given event-time coefficient μ_r is then usually interpreted as the difference between ATE_r and ATE_{-1}. Borusyak, Jaravel & Spiess (2021) additionally note, that another event-time indicator needs to be dropped to avoid perfect multicollinearity if there is no never-treated group. This additional multicollinearity arises because, without a never-treated group, the event-time for each observation is always a linear function of the initial treatment period g_i (i.e. the group) and the calendar time t ($s = t - g_i$ by definition). Besides this, the choices of K and R are flexible depending on assumptions about treatment anticipation and treatment effect dynamics. Without anticipatory behaviour one can, in principle, drop the third term of (8) by setting $K = 1$. What is often done in practice is to interpret negative event-time coefficients as a 'built-in' (joint) test for treatment anticipation and/ or pre-treatment trends. If there is no anticipation and the CTA holds, these coefficients should accordingly be equal (or close) to zero.

It's also possible, and often done in practice, to 'bin' several event-time periods to a single indicator. Binning is most commonly implemented to pool event-times which are smaller or larger than a certain threshold to a single indicator. E.g., one could define an indicator that bins all event-times above the threshold $\bar{r} \geq 0$: $D_{g_it,\geq\bar{r}} := D_{g_it,\bar{r}} + ... + D_{g_it,(T-1)}$. Note that for $\bar{r} = 0$ and $K = 1$, the dynamic TWFE specification coincides with the static TWFE specification (3) and can therefore be rightfully interpreted as a generalization of the latter. The justification put forward for this practice is usually to

increase statistical power, because the number of units that experience an event-time becomes smaller in the distance to the initial treatment period by construction.

Sometimes researchers also use 'trimming' to balance their panel in event-times and then run the dynamic TWFE specification on this subset of the panel. In particular, (i) only treatment groups are included that experience all event-times which are covered in (8) and (ii) only (i,t) of included treatment groups that also belong to one of the covered event-times are included. This modification avoids the issue of compositional differences across event-times in the period-balanced panel, but comes at the cost of a smaller and period-unbalanced panel.

Decomposing Coefficients from the Dynamic TWFE Specification

From the same line of argumentation as in the static TWFE specification, it becomes clear that it's impossible to identify the CEF without further assumptions by the population regression (8). Under the CTA only, the CEF takes the following form (cf. Gardner (2021)):

$$\begin{aligned} E(Y_{i(g_i+r)}|g_i = \bar{g}) &= \alpha_{\bar{g}} + \beta_{\bar{g}+r} + GATE_{\bar{g},\bar{g}+r} \\ &= \alpha_{\bar{g}} + \beta_{\bar{g}+r} + ATE_r + (GATE_{\bar{g},\bar{g}+r} - ATE_r) \end{aligned} \tag{9}$$

Only under the assumption that the GATEs are homogeneous within each event-time, i.e. for each $r : GATE_{\bar{g},\bar{g}+r} = ATE_r$ for all $\bar{g}$, the regression coefficients certainly identify all ATE_r terms. If this is not the case the population regression function does not fully resemble the CEF and hence does not generally identify the event-time specific average treatment effects. OLS will yield the best linear approximation to the CEF in a mean squared error sense, but it's unknown whether the event-time coefficients of the dynamic TWFE specification (8) are still a reasonable summary measure of the GATEs within each event-time. However, given that the dynamic TWFE specification can be seen as a generalization of the static TWFE specification, a similar conjecture seems to be plausible. Note that the dynamic specification is able to accommodate event-time dependent treatment effects, while the static specification would already be misspecified in that case (even without anticipatory behaviour). One interesting question which is ex ante not completely clear from the above reasoning is, whether heterogeneity of GATEs within a specific event-time leads to identification issues of other event-times' coefficients, which actually would satisfy the identification conditions.

Following Sun & Abraham (2021, Proposition 2), the population regression coefficient of the (not-excluded) event-time indicator $D_{g_i t,r}$ in (8) can be expressed as a weighted sum of GATEs of all groups and all event-times:

$$\mu_r = \sum_{\bar{g}} w^r_{\bar{g},r} GATE_{\bar{g},\bar{g}+r} + \sum_{r' \neq r, r' \notin \Re} \sum_{\bar{g}} w^r_{\bar{g},r'} GATE_{\bar{g},\bar{g}+r'} + \sum_{r' \in \Re} \sum_{\bar{g}} w^r_{\bar{g},r'} GATE_{\bar{g},\bar{g}+r'} \tag{10}$$

To account for the flexible definition of the dynamic TWFE specification (8) regarding incorporated event-times, I define a set that consists of all excluded event-times: $\Re := \{-(T-1), ..., -(K+1), -1, R+1, ..., T-1\}$. The $w^r_{\bar{g},r'}$-terms are weights for GATE-terms, where the superscript generally denotes the event-time of the decomposed coefficient, the first subscript denotes the treatment group of the GATE the

weight is affiliated to, and the second subscript denotes the event-time of the GATE the weight belongs to. Group summation generally ranges over all groups that experience the event-time in the second subscript of the respective weight. I don't provide a derivation of the decomposition result here, since it's somewhat cumbersome (cf. Sun and Abraham (2020, Appendix B)). The steps are however very similar to the decomposition of the coefficient in the static TWFE specification presented in the previous section.

The first term in (10) is a weighted average of all GATEs in event-time r, since $\sum_{\bar{g}} w^r_{\bar{g},r} = 1$. As in the static specification, weights can potentially be negative what is a problem as long as $GATE_{\bar{g},\bar{g}+r}$ is not identical in all groups that experience event-time r. The second term is a sum, where each element is itself a weighted sum of GATEs of an included event-time other than r. For a given $r' \notin \Re$, $\sum_{\bar{g}} w^r_{\bar{g},r'} = 0$. Therefore, if $GATE_{\bar{g},\bar{g}+r'}$ is constant across all groups that experience event-time r', the sum cancels out. To make sure that the coefficient for event-time r is not 'contaminated' by GATEs from other included event-times, they need to be homogeneous within each included event-time for all groups. From this it's clear that an event-time coefficient can be contaminated, even when all GATEs of this event-time are in fact homogeneous. Finally, the third term is a weighted sum of all excluded event-times' GATEs, where $\sum_{r' \in \Re} \sum_{\bar{g}} = -1$. This term cancels out when all GATEs of excluded event-times are simultaneously zero. If only negative event-time indicators are excluded this would be fulfilled in case of no anticipatory behaviour. If these GATEs are non-zero, they need to be identical within *and* across all excluded event-times simultaneously to allow for a desirable interpretation of the coefficient. This is of particular relevance, if anticipatory behaviour is likely and in the same time more than one event-time indicator needs to be excluded to avoid the above mentioned multicollinearity.

The weights in (10) for all event-time coefficients in (8) can be obtained by running the following auxiliary regression for each event-time r' (comprising excluded event-times) that a treatment group $g = \bar{g}$ experiences separately:

$$D_{g_i t,r'} I(g_i = \bar{g}) = \breve{\alpha}_{g_i} + \breve{\beta}_t + \sum_{r \notin \Re} w^r_{\bar{g},r'} D_{g_i t,r} + \breve{\epsilon}_{it}$$

The outcome variable in this auxiliary regression is an indicator that takes the value 1 for observations of group $\bar{g}$ in period $\bar{g} + r'$. $\breve{\alpha}_{g_i}$ and $\breve{\beta}_t$ are the unit and time indicator coefficients respectively. Note that this regression yields the weights for all included event-time coefficients, where the GATE of group $\bar{g}$ in period $\bar{g} + r'$ appears.

Only if all three conditions are met, the decomposition result reduces to ATE_r. If only the one period ahead indicator is excluded, the expression would then reduce to $\mu_r = ATE_r - ATE_{-1}$. Another important implication of the decomposition result (10) is that under no anticipation but treatment effect heterogeneity, testing the CTA based on negative event-time coefficients is invalid. The coefficients can be in principle non-zero, even if the CTA holds (or zero if the CTA does not hold), if GATEs are heterogeneous within one or more included event times and/ or in the excluded event-times. Sun and Abraham (2020) show that additional complications arise, if several event-times are binned as single indicators. The GATEs are then required to be homogeneous within *and* across each of the binned event-times simultaneously to avoid difficulties, what is also intuitive based on the CEF argument from above.

Building an Intuition for the Contamination of Coefficients

While the negative weighting problem is already known from the static specification, the contamination of event-time coefficients by GATEs from other event-times is novel and also puzzling at first glance. To build some intuition for this result, consider again the two groups and three periods example from the previous section. Recall that the groups are now labelled according to their initial treatment period. Here, the event-times -1 to 1 of group 2 and the event-times -2 to 0 of group 3 can be observed. Since there is no never-treated group, I drop the indicators for the event-times -2 and -1 from the specification here. The population regression function (8) then takes the specific form:

$$Y_{it} = \hat{\alpha}_{g_i} + \hat{\beta}_t + \mu_0 D_{g_i t,0} + \mu_1 D_{g_i t,1} + \epsilon_{it}$$

I deploy the decomposition result (10) to write the two event-time coefficients as:

$$\begin{aligned}
\mu_0 =& [w^0_{2,0} GATE_{2,2+0} + w^0_{3,0} GATE_{3,3+0}] + \\
& [w^0_{2,1} GATE_{2,2+1}] + \\
& [(w^0_{2,-1} GATE_{2,2-1} + w^0_{3,-1} GATE_{3,3-1}) + w^0_{3,-2} GATE_{3,3-2}] \\
\mu_1 =& [w^1_{2,1} GATE_{2,2+1}] + \\
& [w^1_{2,0} GATE_{2,2+0} + w^1_{3,0} GATE_{3,3+0}] + \\
& [(w^1_{2,-1} GATE_{2,2-1} + w^1_{3,-1} GATE_{3,3-1}) + w^1_{3,-2} GATE_{3,3-2}]
\end{aligned}$$

Each of the three lines corresponds to the respective term in (10). By construction $w^0_{2,1} = 0$ and $w^1_{2,1} = 1$, since only group 2 experiences the event-time 1. Additionally assume that there are no anticipatory effects, such that the third line in both decomposition results disappears. In this example it turns out that $w^0_{2,0} = 1$, $w^0_{3,0} = 0$, and hence $\mu_0 = GATE_{2,2+0}$. On the other hand, $w^1_{2,0} = 1$ and $w^1_{3,0} = -1$, what leads to $\mu_1 = GATE_{2,2+1} + [GATE_{2,2+0} - GATE_{3,3+0}]$. So while μ_0 only contains GATEs from event-time 0 (as one would expect), μ_1 is contaminated by GATEs from event-time 0, as long as they differ between the groups.

It's easy to verify that in this specific case, μ_0 can be expressed as a 2x2 DiD estimand that uses group 3 as a control group:

$$\mu_0 = [E(Y_{i2}|g_i = 2) - E(Y_{i1}|g_i = 2)] - [E(Y_{i2}|g_i = 3) - E(Y_{i1}|g_i = 3)]$$

One may wonder how it's even possible for OLS to determine μ_1 from the data: There are no units which remain untreated until period 3 and subsequently it should be impossible to build a 2x2 DiD estimand (or any other contrast) that disentangles the coefficient from the time trend. To achieve this, it's necessary to make the additional assumption that the GATEs for event-time 0 are homogeneous, i.e. $GATE_{3,3+0} = GATE_{2,2+0}$. This assumption allows to identify the time trend by imputing $GATE_{3,3+0}$ by $GATE_{2,2+0}$ (which is identified by μ_0), what in turn allows to identify $GATE_{2,2+1}$. If the condition holds $GATE_{2,2+1}$ can then be identified by the following sum of 2x2 DiD estimands:

$$\{[E(Y_{i3}|g_i = 2) - E(Y_{i1}|g_i = 2)] - [E(Y_{i3}|g_i = 3) - E(Y_{i1}|g_i = 3)]\} + \mu_0$$

μ_0 needs to be added in order to identify the time trend from the second difference of the 2x2 DiD estimand in curly brackets. On the contrary, if the condition is violated the expression yields $GATE_{2,2+1} + [GATE_{2,2+0} - GATE_{3,3+0}]$, what coincides with the decomposition of μ_1 from above. To the best of my knowledge, I'm the first one to put forward this intuition for the contamination of event-time coefficients in such a clear exposition.

Figure 2 depicts the decomposition result for the event-time coefficient μ_1. The red line segment between $t = 2$ and $t = 3$ represents the case where $GATE_{2,2+0} \neq GATE_{3,3+0}$. Since $GATE_{2,2+0} > GATE_{3,3+0}$ in this example, it follows that μ_1 is larger than $GATE_{2,2+1}$. On the contrary, the black dashed line segment represents the case where $GATE_{2,2+0} = GATE_{3,3+0}$.

Now it becomes clear why μ_1 is contaminated by GATEs from event-time 0: OLS assumes that all GATEs in the same event-time are homogeneous and therefore uses them to identify GATEs of other event-times by means of (unwarranted) extrapolation. It just uses all contrasts to identify parameters that would be admissible, if the structure imposed by the dynamic TWFE specification would resemble the CEF. If it wouldn't do so, it wouldn't leverage all available information to predict the outcome variable under the mean squared error criterion. In a sense, OLS just 'doesn't know better'. The line of argumentation, why contamination occurs in the dynamic specification is in principle the same as for the negative weights in the static specification: OLS makes unwarranted assumptions about the homogeneity of treatment effects to build control groups.

Albeit the decomposition result of event-time coefficients in terms of 2x2 DiD estimands does not easily generalize to complexer settings – it's still a very useful way to intuitively think about contamination of event-time coefficients. Borusyak et al. (2021, Proposition 4) note that, conditional on no anticipatory behaviour, for all event-times $r \geq \max(g_i) - \min(g_i)$, there is no way to identify event-time coefficients without unwarranted extrapolation from prior event-times (regardless of the identification technique). The reason for this is, that there are no untreated units left that could be used as a 'clean' control group. However, the decomposition result (10) suggests that contamination even occurs for event-times, where identification would be feasible without extrapolation. This reflects the notion of OLS to exploit all contrasts that would be admissable when treatment effects satisfy the required homogeneity assumptions.

Following this logic, contamination should be worse for large event-times, given that the number of untreated units, and therefore the number of clean contrasts, decreases over time. To illustrate this, consider again the example with two groups but with a fourth period added. In this case, one could try to identify $GATE_{2,2+2}$ in a similar fashion by:

$$\begin{aligned}
&\{[E(Y_{i4}|g_i = 2) - E(Y_{i1}|g_i = 2)] - [E(Y_{i4}|g_i = 3) - E(Y_{i1}|g_i = 3)]\} + \\
&\{[E(Y_{i3}|g_i = 2) - E(Y_{i1}|g_i = 2)] - [E(Y_{i3}|g_i = 3) - E(Y_{i1}|g_i = 3)] + \\
&[E(Y_{i2}|g_i = 2) - E(Y_{i1}|g_i = 2)] - [E(Y_{i2}|g_i = 3) - E(Y_{i1}|g_i = 3)]\}
\end{aligned}$$

To identify $GATE_{2,2+2}$ it now needs to hold that $GATE_{2,2+1} = GATE_{3,3+1}$ (second line), where $GATE_{2,2+1}$ again relies on $GATE_{3,3+0} = GATE_{2,2+0}$ (third line). Identification subsequently relies on one additional homogeneity restriction.

3.3 TWFE Regression with a Non-binary Treatment

In many settings treatment variables can take on more than two values. Consider for instance a scenario, where the outcome variable is unemployment at the county level and the treatment variable is a minimum wage, which is set at the state level. Additionally, the minimum wage can be potentially revised multiple times in the observed time periods. The DiD principle naturally extends from the binary treatment setting: Units that experience changes in treatment intensity over time can be compared to units with constant treatment intensity. Units that experience changes in treatment intensity can be also compared to units that experience a different change in treatment intensity.

To estimate treatment effect parameters in the non-binary setting, it's common to simply use the TWFE specification (3) with a non-binary treatment variable.

Potential Outcomes and Terminology

With a non-binary treatment variable it's necessary to define a potential outcome for each possible value of the treatment $Y_{it}(D), D \in \mathbb{D} \subseteq \mathbb{R}$. $Y_{it}(D)$ is the potential outcome of unit i at time t under treatment 'dose' D. One way to define a treatment effect is to choose a benchmark dose for comparison; $D = 0$ lends itself as a natural point of comparison: $\rho_{it}(D) = Y_{it}(D) - Y_{it}(0)$. This definition can be seen as a generalization of the classical potential outcomes framework. When $\mathbb{D} = \{0, 1\}$, $\rho_{it}(0) = 0$ and $\rho_{it}(1)$ is identical to the treatment effect defined in section 2.2. Alternatively, one could be interested in the effect induced by a one unit increase in the treatment $Y_{it}(D) - Y_{it}(D-1)$, or, if the treatment is continuous, a marginal increase in the treatment $d(Y_{it}(D))/d(D)$ (Callaway et al., 2021). A unit's treatment history is still assumed to be bound to its group affiliation. The observed outcome is then given by $Y_{it} := Y_{it}(D_{g_i t})$. In settings where treatment is assigned at a group level (e.g. in the above mentioned minimum wage example) this assumption is fulfilled. However, in many practical settings it may be impossible to form groups with uniform treatment histories, especially if the treatment is continuous. The CTA is imposed for the zero-dose potential outcome, $Y_{it}(0)$, in the same way as for the untreated/ never-treated potential outcome in the previous sections.

Identification in the 2x2 Setting

Consider the setting with two groups and two periods introduced in section 2.2, but with a non-binary treatment. A natural generalization of the 2x2 DiD estimand (1) for non-binary treatments is then given by:

$$\frac{E(Y_{i2} - Y_{i1}|g_i = 2) - E(Y_{i2} - Y_{i1}|g_i = 1)}{(D_{22} - D_{21}) - (D_{12} - D_{11})} \tag{11}$$

The 2x2 DiD estimand (1) is normalized by the difference of the changes in the groups' treatment dose over time, to account for the fact that both groups' treatment state can differ over time. Note that with a binary treatment, $D_{22} = 1, D_{21} = 0$ and $D_{12} = D_{11} = 0$ (i.e. the treatment histories of the generic 2x2 setting introduced in

section 2.2), the expression collapses to the 2x2 DiD estimand (1). Also note that the expression is only defined for $(D_{22} - D_{21}) \neq (D_{12} - D_{11})$. This reflects the DiD principle to exploit diverging treatment histories for identification: If both groups would experience the same change in treatment dose, the time variation would be indistinguishable from the treatment variation. Feldstein (1995) studies for instance the effect of a change in the marginal income tax rate on taxable income based on this approach. He exploits a tax reform, where different income groups are subject to different changes in the marginal income tax rate.

Until now I didn't make any claims about the treatment effect parameter the estimand is meant to identify. Without further assumptions than the the CTA, (11) identifies:

$$\frac{E[\rho_{i2}(D_{22}) - \rho_{i1}(D_{21})|g_i = 2] - E[\rho_{i2}(D_{12}) - \rho_{i1}(D_{11})|g_i = 1]}{(D_{22} - D_{21}) - (D_{12} - D_{11})}$$

This expression has yet not really any manageable interpretation, mainly due to the general definition of potential outcomes. Assume that the treatment effect function takes a linear form $\rho_{it}(D) = \ddot{\rho}_{it}D$, where $\ddot{\rho}_{it}$ is the observation specific slope parameter. Unfortunately, the set of assumptions is still to weak to give the resulting expression a sensible causal interpretation, because $\ddot{\rho}_{it}$ can vary systematically between groups, periods, or both. Only if one is willing to assume that $E(\ddot{\rho}_{i\bar{t}}|g_i = \bar{g}) = E(\ddot{\rho}_{it})$ for all $(\bar{g}, \bar{t}) \in \{1, 2\}^2$, the estimand identifies an easy to interpret causal coefficient. These assumptions are by construction not necessary in the binary treatment case, because only one group has a non-zero treatment dose at one point in time and linearity of the treatment effect directly follows from the binary treatment.

The extended 2x2 DiD estimand for non-binary treatments (11) can therefore yield very misleading parameters, if one or more of the parametric assumptions is violated. Assume for instance that the linear model for the potential outcomes holds and that the average slope parameter only varies by group but not by time, i.e. $E(\ddot{\rho}_{i\bar{t}}|g_i = \bar{g}) = E(\ddot{\rho}_{it}|g_i = \bar{g})$. In particular, assume that $D_{21} = D_{11} = 0$, $D_{22} = 1$, $D_{12} = 2$, and both groups are equally sized. Let $E(\ddot{\rho}_{it}|g_i = 2) = 3$ and $E(\ddot{\rho}_{it}|g_i = 1) = 1$. Hence, $E(\ddot{\rho}_{it}) = 2$ but the DiD estimand (11) yields -1. So even when the average treatment slope parameters are positive, it's feasible that the estimand indicates a negative parameter, if treatment effects are heterogeneous across groups.

Extension of the Decomposition Result (5)

The decomposition of the treatment coefficient in the static TWFE regression (5) can be extended to non-binary treatments in a straightforward way, because regression mechanics do not depend on the scale of the outcome variable. Under the CTA, the CEF takes the following form in the non-binary treatment setting:

$$E(Y_{it}|g_i = \bar{g}) = \alpha_{\bar{g}} + \beta_{\bar{t}} + E[\rho_{i\bar{t}}(D_{g_i\bar{t}})|g_i = \bar{g}]$$

The treatment coefficient μ from the TWFE regression (3) with non-binary treatment can then be expressed as:

$$\begin{aligned}\mu &= \frac{\sum_{(\bar{g},\bar{t})} |\bar{g}| \tilde{D}_{\bar{g}\bar{t}} E[\rho_{i\bar{t}}(D_{g_i\bar{t}})|g_i = \bar{g}]}{\sum_{(\bar{g},\bar{t})} |\bar{g}| \tilde{D}^2_{\bar{g}\bar{t}}} \\ &= \sum_{(\bar{g},\bar{t}):D_{\bar{g}\bar{t}}\neq 0} w_{\bar{g}\bar{t}} \frac{E[\rho_{i\bar{t}}(D_{g_i\bar{t}})|g_i = \bar{g}]}{D_{\bar{g}\bar{t}}} \end{aligned} \tag{12}$$

, where $w_{\bar{g}\bar{t}} = \frac{|\bar{g}|\tilde{D}_{\bar{g}\bar{t}} D_{\bar{g}\bar{t}}}{\sum_{(\bar{g},\bar{t})} |\bar{g}| \tilde{D}^2_{\bar{g}\bar{t}}}$. Note that the weights coincide with the definition of the weights in (5), if $D_{\bar{g}\bar{t}}$ is binary. Furthermore, the weights also sum to one as in the binary treatment case. $\frac{E(\rho_{i\bar{t}}(D_{g_i\bar{t}})|g_i=\bar{g})}{D_{\bar{g}\bar{t}}}$ can be interpreted as the average treatment effect at dose $D_{\bar{g}\bar{t}}$ of group $\bar{g}$ in period $\bar{t}$ per unit of dose. If the treatment is binary, the expression again collapses back to the decomposition result in (5) (after taking into account different potential outcomes notations). When all three homogeneity conditions, which were previously introduced for the 2x2 setting with a non-binary treatment, are fulfilled, the decomposition result collapses to $E(\ddot{\rho}_{it})$.

While the interpretation of weights is somewhat more complex than in the binary treatment case, it still holds that weights depend on group size, but also on treatment dose, and the treatment history of the respective group itself and all other groups. In particular, sign and magnitude of a weight is determined by the weight's numerator, which can be expanded to:

$$|\bar{g}|\tilde{D}_{\bar{g}\bar{t}} D_{\bar{g}\bar{t}} = |\bar{g}|[D_{\bar{g}\bar{t}} - E(D_{\bar{g}t}) - E(D_{g_i\bar{t}}) + E(D_{g_it})] D_{\bar{g}\bar{t}}$$

It's also easy to verify, that the decomposition result (12) yields the same result as the extended 2x2 DiD estimand (11) for the previously established example. Here, the weights are given by $w_{11} = w_{21} = 0$, $w_{22} = -1$, and $w_{12} = 2$.

3.4 Simulations

While the established decomposition results are mathematically exact, it still may be hard to evaluate under which conditions the differences between the TWFE coefficients and the target estimands are noticeable in more practical settings. Therefore it could be insightful to run simulations for different patterns of treatment effect heterogeneity and different heterogeneity degrees. Intuitively, one would expect that deviations become larger as heterogeneity increases. One interesting question could also be, how the presence and the size of a never-treated control group influences the coefficients. One may expect that deviations between the TWFE coefficients and the target estimands become smaller, as the control group's size increases relative to the ever-treated units, because the number of 'clean' comparisons for identification becomes larger.

The subsequent simulation results may be sensitive to the underlying assumptions and therefore should mainly serve for illustrative purposes. Because of this, my interpretations of the results remain conservative and shouldn't be generalized to other settings.

Setting

I consider a staggered adoption setting with ten periods ($T = 10$). There are nine equally sized treatment groups as well as a never-treated group with flexible size, $g_i \in \{2, 3, ..., 10, \infty\}$. Group labels are assigned according to the initial treatment period (as in section 3.2). The observed outcomes are generated by the following process:

$$Y_{it} = \alpha_i + \beta_t + \rho_{it} D_{g_i t}$$

I do not add an error term here to fully focus on the OLS mechanics. Therefore, it's enough to generate a single treatment path for each ever-treated group. In principle, this is equivalent to directly observing the CEF. Alternatively one could add a mean-zero exogenous error term and simulate a large number of units per group to get the equivalent result. For the simulations the relative size of the never-treated group will vary between zero- and three-times the number of all ever-treated units.

I consider four different patterns of treatment effect heterogeneity, where $d \geq 0$ is the heterogeneity parameter:

a) Calendar time dependent treatment effect: $\rho_{it} = 1 + (t - 1) \cdot d$

   The treatment effect is a linear function of the calendar time with positive slope.

b) Group dependent treatment effect : $\rho_{it} = 1 + (g_i - 1) \cdot d$

   Units experience a treatment effect linear in their first treatment date, where later treated units have a higher treatment effect.

c) Event-time dependent treatment effect: $\rho_{it} = 1 + (t - g_i) \cdot d$

   The treatment effect is a linear function of the event-time with positive slope.

d) Event time and group dependent treatment effect: $\rho_{it} = 1 + (t - g_i) + (t - g_i) \cdot g_i \cdot d$

   The treatment effect depends both on the event-time and the first treatment date, in particular, later treated units' treatment effects increase faster in event-time.

In each case, a larger value of d corresponds to a higher level of treatment effect heterogeneity, where $d = 0$ implies in the first three cases homogeneous treatment effects, i.e. $\rho_{it} = 1$. In the last case, $d = 0$ corresponds to homogeneous treatment effects within each event time (equivalent to the third case with $d = 1$). For the simulations I consider the set of values $d \in \{0, 0.05, 0.25, 0.5, 1\}$.

In the simulations, I systematically vary the level of treatment effect heterogeneity and the relative size of the never-treated group for different forms of treatment effect heterogeneity. In particular, I consider the first three patterns for the static TWFE specification (3) and the last pattern for the dynamic TWFE specification (8). Recall that the latter specification is always correctly specified for the purely event-time dependent heterogeneity pattern c). I then compare the TWFE coefficients to the ATT and the event-time dependent average treatment effects for the static and the dynamic specifications respectively.

Static TWFE Specification

Figure 3 displays the simulation results of the static TWFE specification for the first three heterogeneity patterns, different levels of heterogeneity, and different relative sizes of the never-treated group. I report the deviation of the TWFE coefficient from the ATT in per cent to account for the fact that different heterogeneity levels imply different ATTs. The ATT is however always positive here, since the treatment effects are strictly positive. A deviation < -1 would therefore indicate that the TWFE coefficient has a negative sign. Note that the ATT is independent from the relative size of the never-treated group by construction.

Unsurprisingly, if treatment effects are homogeneous ($d = 0$), the TWFE coefficient hits the ATT. In all three cases, a higher level of treatment effect heterogeneity coincides with a larger deviation of the TWFE coefficient for a given size of the never-treated group. This is very intuitive, since a higher level of treatment effect heterogeneity corresponds to a stronger violation of the identification conditions. The size of the never-treated group relative to the number of ever-treated units has a less clear impact: While the coefficient converges to 'some' constant as the relative size of the never-treated group increases in all three cases, this constant is not necessarily the ATT. Interestingly, for the group dependent pattern (panel b)) the deviation seems to be independent from the never-treated group's size. For the calendar time dependent heterogeneity pattern (panel a)) the TWFE coefficient approaches the ATT first, hits it, and diverges again as the never-treated group's relative size increases further.

Dynamic TWFE Specification

Figure 4 depicts the coefficient path of the dynamic TWFE specification and the true event-time specific average treatment effects for different heterogeneity levels and relative sizes of the never-treated group. In contrast to the static specification, I report level values and not deviations in per cent, since this would not allow to have meaningful interpretations for the event-time coefficients of pre-treatment periods. The never-treated group's size relative to the number of ever-treated units increases from left to right and the degree of treatment effect heterogeneity increases from top to bottom. Hence, all plots in a column correspond to the same size of the never-treated group and all plots in a row correspond to the same level of treatment effect heterogeneity.

I omit the indicator variables for event-times -1 and -9 from the dynamic TWFE specification, to avoid perfect multicollinearity in the case without a never-treated group. While it's in principle not necessary to omit more than one indicator if there are never-treated units, I stick to this specification for the sake of coherence. Importantly, the patterns remain essentially identical if the indicator for event-time -9 is also included.

Like in the static specification, coincides a higher degree of treatment effect heterogeneity generally with larger deviations of the TWFE coefficients from the true parameters. Also, deviations seem to be systematically larger for later (positive) event-times. Interestingly, without a never-treated group, the TWFE coefficients lie systematically below the true parameters, while they systematically overstate the true parameters with a control group. One can also clearly see that the dynamic TWFE specification

can yield non-zero coefficients for pre-periods even if they are all zero, what is an implication of the decomposition result by Sun and Abraham (2020). This problem amplifies in the degree of heterogeneity and, a bit surprisingly, is also more pronounced if a never-treated group exists.

4 Alternative Estimators

In this section I introduce, evaluate, and compare alternative estimators, which have been developed in response to the theoretical shortcomings of TWFE regressions when treatment effects are heterogeneous. Each subsection corresponds to one estimator/ identification approach. The structure of the subsections is standardized to maximize the readability and comparability: I start by explaining the setting, the identification target, and state the underlying identifying assumptions. Then, I introduce and explain the identification approach in detail. Lastly, I consider statistical inference, but without going into any details.

To account for all peculiarities of the approaches, it's necessary to switch between the two different potential outcome notations from the previous section. For each estimator, I use the potential outcome notation that's also used by the respective authors.

4.1 de Chaisemartin and D'Haultfœuille (2020)

Setting

de Chaisemartin and D'Haultfœuille (2020) focus on a setting with binary treatment and arbitrary treatment trajectories, i.e. units that receive treatment once are allowed to become untreated in later periods. I will therefore make use of the classical potential outcomes framework introduced in section 2.2. Their goal is to improve upon conventional TWFE regressions by developing an estimator that identifies a 'well-defined' causal effect under some forms of treatment effect heterogeneity.

Assumptions

A priori, they impose three assumptions:

A.1 CTA. For all $(\bar{g}, \bar{t}) \in \{1, ..., G\} \times \{2, ..., T\}$:

$$E(Y_{i\bar{t}}(0) - Y_{i\bar{t}-1}(0)|g_i = \bar{g}) = E(Y_{i\bar{t}}(0) - Y_{i\bar{t}-1}(0))$$

A.2 Common trends of GATEs. For all $(\bar{g}, \bar{t}) \in \{1, ..., G\} \times \{2, ..., T\}$:

$$E(\rho_{i\bar{t}} - \rho_{i\bar{t}-1}|g_i = \bar{g}) = E(\rho_{i\bar{t}} - \rho_{i\bar{t}-1})$$

Assumption 2 requires that GATEs evolve over time in the same way for all treatment groups and thereby parallels the CTA: Level differences between groups in GATEs are still allowed, as long as the change between consecutive periods is the same for

all groups. Put differently, GATEs are allowed to be equal to the sum of a time independent but group specific component and a time specific but group independent component. Equivalently, one could summarize assumptions 1 and 2 as a 'common trends assumption under treatment', since treatment effects are defined as the difference between the treated and untreated potential outcomes in this framework. Figure 5 depicts the assumption in a setting with two groups and two periods. Assumption 2 would be for instance plausible in a scenario, where treatment effects depend on the business cycle and therefore only on calendar time. However, dynamic treatment effects are effectively ruled out, since this would imply a dependence of treatment effects on group and time simultaneously.

A.3 'Stable' groups. For all $t \in \{2, ..., T\}$:

a) If there is at least one group $\bar{g}$ with $D_{\bar{g}t-1} = 0$ and $D_{\bar{g}t} = 1$, there is at least one group $g' \neq \bar{g}$ with $D_{g't-1} = 0$ and $D_{g't} = 0$.

b) If there is at least one group $\bar{g}$ with $D_{\bar{g}t-1} = 1$ and $D_{\bar{g}t} = 0$, there is at least one group $g' \neq \bar{g}$ with $D_{g't-1} = 1$ and $D_{g't} = 1$.

A.4 No treatment anticipation and no treatment spillovers from treated to untreated periods.

By the no spillover assumption, if a unit receives treatment in a given period, the treatment effect is not allowed to persist in subsequent formally untreated periods.[3] Note that this assumption is already implicitly imposed in the treated/ untreated potential outcomes framework.

I will further explain the purpose of the assumptions after introducing the identification approach.

Identification

1. For $\bar{t} \in \{2, ..., T\}$, use all $i : D_{g_i\bar{t}-1} = 0, D_{g_i\bar{t}} = 1$ and all $i : D_{g_i\bar{t}-1} = 0, D_{g_i\bar{t}} = 0$ to build the following DiD estimand:

   $$DID_{+,\bar{t}} := E(Y_{i\bar{t}} - Y_{i\bar{t}-1} | D_{g_it} = 1, D_{g_it-1} = 0) - E(Y_{i\bar{t}} - Y_{i\bar{t}-1} | D_{g_it} = 0, D_{g_it-1} = 0)$$

   If for a given period there are no units with $D_{g_i\bar{t}-1} = 0, D_{g_i\bar{t}} = 1$, set $DID_{+,\bar{t}} = 0$ (since the estimand is not defined properly).

2. For $\bar{t} \in \{2, ..., T\}$, use all $i : D_{g_i\bar{t}-1} = 1, D_{g_i\bar{t}} = 0$ and all $i : D_{g_i\bar{t}-1} = 1, D_{g_i\bar{t}} = 1$ to build the following DiD estimand:

   $$DID_{-,\bar{t}} := E(Y_{i\bar{t}} - Y_{i\bar{t}-1} | D_{g_it} = 1, D_{g_it-1} = 1) - E(Y_{i\bar{t}} - Y_{i\bar{t}-1}) | D_{g_it} = 0, D_{g_it-1} = 1)$$

   If for a given period there are no units with $D_{g_i\bar{t}-1} = 1, D_{g_i\bar{t}} = 0$, set $DID_{-,\bar{t}} = 0$.

[3] de Chaisemartin and D'Haultfœuille (2021) consider a binary treatment setting, where treatment spillovers from treated to untreated periods are not ruled out. However, except in the special case of staggered adoption, their approach is not able to identify an easy to interpret causal parameter. This is mainly due to the fact, that there are 2^T potential treatment paths in non-staggered settings, where treatment effects cannot be easily aggregated with treatment spillovers. Instead they focus on economically interpretable cost-benefit ratios. In the special case of staggered adoption, their approach is numerical identical to the approach of Callaway and Sant'Anna (2020), which I review in section 4.4.

3. Build the weighted average of all DiD estimands from step 1 and step 2:

$$\sum_{\bar{t}=2}^{T}(w_{+,\bar{t}}DID_{+,\bar{t}} + w_{-,\bar{t}}DID_{-,\bar{t}})$$

, where $w_{+,\bar{t}} := \frac{\sum_i I(D_{g_i\bar{t}} = 1, D_{g_i\bar{t}-1} = 0)}{\sum_{(i,t):t\geq 2} I(D_{g_it} \neq D_{g_it-1})}$ and $w_{-,\bar{t}} := \frac{\sum_i I(D_{g_i\bar{t}} = 0, D_{g_i\bar{t}-1} = 1)}{\sum_{(i,t):t\geq 2} I(D_{g_it} \neq D_{g_it-1})}$.

It's straightforward to see that, under assumptions 1, 3, and 4, $DID_{+,\bar{t}}$ identifies the average treatment effect of all units that switch from untreated to treated ('joiners') between periods $\bar{t}-1$ and $\bar{t}$ in the latter period, $E(\rho_{i\bar{t}}|D_{g_it} = 1, D_{g_it-1} = 0)$. Note that one could equivalently express $DID_{+,\bar{t}}$ as a group size weighted average of all joiner groups' GATEs in period $\bar{t}$. If there is only one group $\bar{g}$ that joins the treatment in period $\bar{t}$, $E(\rho_{i\bar{t}}|D_{g_it} = 1, D_{g_it-1} = 0) = GATE_{\bar{g},\bar{t}}$ holds. Assumption 3 assures here, that there is always an untreated control group to build the DiD estimand.

The rationale behind $DID_{-,\bar{t}}$ may be however less clear. If assumptions 1, 3, and 4 hold, $DID_{-,\bar{t}}$ identifies:

$$DID_{-,\bar{t}} = E(\rho_{i\bar{t}-1}|D_{g_it} = 0, D_{g_it-1} = 1) + E(\rho_{i\bar{t}} - \rho_{i\bar{t}-1}|D_{g_it} = 1, D_{g_it-1} = 1)$$

Now, if additionally assumption 2 holds, $DID_{-,\bar{t}}$ identifies $E(\rho_{i\bar{t}}|D_{g_it} = 0, D_{g_it-1} = 1)$, i.e. the average treatment effect of all units that switch from treated to untreated ('leavers') between periods $\bar{t}-1$ and $\bar{t}$ in the latter period, $E(\rho_{i\bar{t}}|D_{g_it} = 0, D_{g_it-1} = 1)$. Assumption 2 allows to reconstruct the average treatment effect of the leavers in their untreated period $\bar{t}$, by imputing the treatment effect evolution from those units remaining treated.

The third step builds a weighted average of all DiD estimands from step 1 and step 2. $w_{+,\bar{t}}$ is the number of joiners in period $\bar{t}$ relative to the overall number of treatment status changes in the panel and $w_{-,\bar{t}}$ has the analogous interpretation. Subsequently, $\sum_{\bar{t}=2}^{T}(w_{+,\bar{t}}+w_{-,\bar{t}}) = 1$. The weighted average therefore identifies the average treatment effect of all (i,t), where the treatment status switches in comparison to the previous period: $E(\rho_{it}|D_{g_it} \neq D_{g_it-1})$. Note that if a unit joins the treatment and stays treated for some periods thereafter before leaving the treatment, only the treatment effects of the first period under treatment and of the first untreated period receive consideration.

If there is no stable group to build a well-defined DiD estimand (a violation of assumption 3), one can simply omit the respective DiD estimand and adjust the denominator of the weights accordingly. Of course, it's also possible to separately identify average treatment effects of the joiners and leavers, by calculating weighted averages of only $DID_{+,\bar{t}}$ and $DID_{-,\bar{t}}$ respectively. In a staggered adoption setting, all $DID_{-,\bar{t}}$ vanish from the weighted average, since groups can't leave the treatment after joining once. Assumption 2 becomes redundant for identification in this case.

The approach intuitively circumvents the (negative) weighting problem of the static TWFE regression (3), by only exploiting treatment contrasts which are valid under the stated assumptions and by separating the identification and weighting of treatment effects.

Placebo Test

de Chaisemartin and D'Haultfœuille (2020) also propose an intuitive placebo test for the CTA and the common trends of GATEs assumptions. The basic idea is to apply the above mentioned framework to time periods $\bar{t}-2$ and $\bar{t}-1$ for two sets of groups: Units with constant treatment status in $\bar{t}-2$ and $\bar{t}-1$ and a different treatment status in $\bar{t}$ and units with constant treatment status in all three periods.

In particular:

$$DID^{pl}_{+,\bar{t}} := E(Y_{i\bar{t}-1}-Y_{i\bar{t}-2}|D_{g_it}=1, D_{g_it-1}=D_{g_it-2}=0)-E(Y_{i\bar{t}-1}-Y_{i\bar{t}-2}|D_{g_it}=D_{g_it-1}=D_{g_it-2}=0)$$

$$DID^{pl}_{-,\bar{t}} := E(Y_{i\bar{t}-1}-Y_{i\bar{t}-2}|D_{g_it}=D_{g_it-1}=D_{g_it-2}=1)-E(Y_{i\bar{t}-1}-Y_{i\bar{t}-2}|D_{g_it}=0, D_{g_it-1}=D_{g_it-2}=1)$$

The remainder is then analogous to the third step from above. Of course, the placebo test requires an extension of the stable groups assumption. Under the stated assumptions, the test quantity is then zero.

Inference

One can easily transfer the identification strategy to a sample setting, by replacing the population quantities by their sample counterparts. de Chaisemartin and D'Haultfœuille (2020) show that the resulting estimator is unbiased, consistent, and asymptotically Normal under some standard conditions.

4.2 Borusyak, Jaravel, and Spiess (2021) and Gardner (2021)

Setting

Borusyak et al. (2021) mainly focus on the binary treatment setting with staggered adoption. It's however possible to apply the framework to non-staggered settings under some additional assumptions. Their goal is to develop a framework, which is able to flexibly identify weighted sums of treatment effects without imposing any homogeneity restrictions. In accordance with the authors, I make use of the treated/ untreated potential outcomes framework introduced in section 2.2.

Assumptions

A.1 CTA.

A.2 No treatment anticipation.

While the CTA is crucial, it's feasible to modify the procedure to allow for some anticipation.

Identification

In its simplest form, their 'imputation' approach takes the following form:

1. Restrict the population to untreated observations, i.e. $(i,t) : D_{g_it} = 0$. Run an OLS regression of the outcome variable on unit and time indicators within the untreated stratum:
$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \epsilon_{it} \tag{13}$$

2. For all treated observations, i.e. $(i,t) : D_{g_it} = 1$, impute the (unknown) potential outcome under no treatment, $Y_{it}(0)$, by using the unit and time indicator coefficients from the first step:
$$\tilde{Y}_{it}(0) = \tilde{\alpha}_i + \tilde{\beta}_t$$

3. Approximate the treatment effects, ρ_{it}, for all treated observations by:
$$\tilde{\rho}_{it} = Y_{it} - \tilde{Y}_{it}(0)$$

4. Approximate a weighted sum of treatment effects, $\sum_{(i,t):D_{g_it}=1} w^*_{it}\rho_{it}$, by replacing the unobserved treatment effects with their approximated counterparts from step 3:
$$\sum_{(i,t):D_{g_it}=1} w^*_{it}\tilde{\rho}_{it}$$
The weights w^*_{it} can be chosen freely to approximate the treatment effect summary measure of interest. E.g., if one is interested in the overall ATT, all w^*_{it} would be equal to one over the number of treated (i,t).

The imputation approach tackles the shortcomings of conventional TWFE regressions in three way: First, it uses only untreated (i,t) to approximate counterfactual outcomes for treated (i,t) and thereby does not rely on treatment effect homogeneity assumptions for causal identification. Second, it also avoids any concerns regarding spurious identification of long run treatment effects in dynamic TWFE specifications: If at a point in time all units are treated, it becomes impossible to identify the time fixed effect, since the indicator would be zero for all (i,t) in the untreated stratum. Therefore, it's also impossible to obtain treatment effect approximations for this point in time, because it's not feasible to approximate the untreated potential outcome. The same line of argumentation applies to always-treated units. Finally, it separates the identification and weighting of treatment effects.

While the imputation approach intuitively seems to be valid to identify (weighted sums of) treatment effects, there is no straightforward proof for this due to the unconventional multi-step procedure. Luckily, it's feasible to represent it in a way for which a well-established theory exists. One can show that the first three steps of the imputation approach are equivalent to a regression of the outcome variable on unit and time indicators and a unique indicator for each treated (i,t):
$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \tilde{\rho}_{it}D_{g_it} + \epsilon_{it} \tag{14}$$

By including a separate indicator for each treated (i,t) the regression is, by construction, correctly specified if the identifying assumptions hold. Therefore, $\tilde{\rho}_{it}$ is always a valid approximation of ρ_{it}. Note that the regression is only applicable, if never-treated units exist and there are no always-treated units. Otherwise, a linear combination

of the treated (i,t)s' indicators would be multicollinear to the time or unit indicators. This parallels the identification requirements in the imputation representation. Of course, it's always possible to fulfil this requirement by restricting the population appropriately.

Nevertheless, it's still necessary to proof that the correspondence between the imputation approach and the regression approach holds to be confident about the implications. To see this, consider the OLS minimization problem underlying the population regression (14):

$$[(\tilde{\alpha}_i)_{i=1}^N, (\tilde{\beta}_t)_{t=1}^T, (\tilde{\rho}_{it})_{(i,t):D_{it}=1}]' = \arg\min \sum_{(i,t)} (Y_{it} - a_i - b_t - p_{it} D_{g_i t})^2$$

$$= \arg\min \{ \sum_{(i,t):D_{g_i t}=0} (Y_{it} - a_i - b_t)^2 + \sum_{(i,t):D_{g_i t}=1} (Y_{it} - a_i - b_t - p_{it})^2 \}$$

For all $(i,t) : D_{g_i t} = 1$, it holds that $Y_{it} - a_i - b_t - p_{it} = 0$, by setting $p_{it} = Y_{it} - a_i - b_t$. Hence, treated (i,t) don't need to be taken into account in the minimization problem. A regression of Y_{it} on unit and time indicators in the untreated stratum (i.e. regression (13)) therefore yields the same unit and time indicator coefficients as regression (14) on the whole population. Now, since $p_{it} = Y_{it} - \tilde{\alpha}_i - \tilde{\beta}_t$ minimizes the residuals for all treated (i,t) the regression coefficients are identical to $\tilde{\rho}_{it}$ from the third step of the imputation representation. Note that the residuals from regression (14) are zero for all treated (i,t). Given that $\tilde{\rho}_{it}$ relies on a single (i,t), it will be a (very) noisy approximation of ρ_{it} even in the population. This is however not a problem as long as the approximations do not systematically differ from the true value, since deviations will average out in the aggregation step.

The imputation representation is not only intuitive, but also has a computational advantage: Including a single indicator for each treated (i,t) may become computationally challenging if the number of observations is large, whereas the imputation approach only requires a regression with unit and time indicators, for which fast algorithms are available.

The imputation approach can be applied to staggered adoption settings with some anticipatory behaviour. In this case, one can simply restrict the population in the first step appropriately to pre-periods outside of the effect window to avoid any sort of contamination. Treatment effects can then also be approximated for pre-periods within the effect window. It's also feasible to apply the procedure to non-staggered binary treatment settings, if spillovers of treatment effects can be ruled out. This means that if a unit receives treatment in a given period the effect is not allowed to persist in subsequent (formally) untreated periods. Otherwise, the identification of unit and time effects in the first step would be contaminated by treatment effects. One can additionally flexibly account for unit and time dependent covariates, by including them in the regression specification in the first step to predict the counterfactual outcomes of the treated observations. In the same way it's also feasible to include more complex sets of fixed effects and for instance unit specific linear time trends.

Testing for Pre-Treatment Trends and Anticipatory Behaviour

Note that the imputation approach does not come with a 'built-in' test for pre-treatment trends and/ or anticipatory behaviour as it's the case in dynamic TWFE specifications, where lead coefficients are usually used as a visual test. However, as explained in section 3.2, this practice may be unwarranted if treatment effects are heterogeneous, because pre-treatment coefficients can be contaminated by post-treatment treatment effects. Borusyak et al. (2021) propose an alternative testing approach that closely resembles the logic of the imputation estimator:

1. Restrict the population to untreated observations, i.e. $(i,t) : D_{g_it} = 0$. Run an OLS regression of the outcome variable on unit and time indicators and additionally a set of lead indicators:

$$Y_{it} = \hat{\alpha}_i + \hat{\beta}_t + \sum_{r=-K}^{-1} \mu_r D_{g_it,r} + \epsilon_{it} \tag{15}$$

   , with $K > 0$.
2. Use the F-test to test for the hypothesis $\mu_{-K} = ... = \mu_{-1} = 0$.

This test prevents contamination by only using the untreated stratum. Coefficients from the first step can be alternatively used for a visual test in the spirit of the dynamic TWFE specification, but without the associated pitfall. Whether this is a test for parallel pre-trends, treatment anticipation, or both simultaneously depends on the assumptions one is willing to impose.

Inference

The imputation approach can be directly transferred to a sample setting. Borusyak et al. (2021) show, that the resulting estimator is then unbiased, consistent, and asymptotically Normal for weighted sums of treatment effects. They also analytically derive the asymptotically valid standard error estimator.

Gardner (2021)

Gardner (2021) proposes a very similar 'two stage' DiD approach. He mainly focusses on the identification of the overall ATT, but also proposes an extension for event-time specific average treatment effects. His approach takes the following form:

1. Regress the outcome variable on unit and time indicators in the subpopulation with $(i,t) : D_{g_it} = 0$:
$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \epsilon_{it}$$
2. For *all* (i,t), subtract the unit and time indicator coefficients from step 1 from the observed outcome:
$$Y_{it} - \tilde{\alpha}_i - \tilde{\beta}_t$$

3. Regress the adjusted observed outcome from step 2 on an intercept and the treatment indicator in the *complete* population:

$$(Y_{it} - \tilde{\alpha}_i - \tilde{\beta}_t) = \lambda_0 + \lambda_1 D_{g_i t} + \tilde{\epsilon}_{it}$$

While the first step is identical to the first step of the imputation approach, the second and third steps differ.[4] However, it can be shown that the two stage approach is just a special case of the imputation approach:

Since the regression in the third step is saturated, it follows that $\lambda_0 = E(Y_{it} - \tilde{\alpha}_i - \tilde{\beta}_t | D_{g_i t} = 0)$. Given that for all untreated (i, t), the adjusted outcome from the second step is identical to the residual from the first step and the fact that residuals always sum to zero, it follows that $\lambda_0 = 0$ by construction (so the intercept can be dropped in principle). Subsequently,

$$\lambda_1 = E(Y_{it} - \tilde{\alpha}_i - \tilde{\beta}_t) | D_{g_i t} = 1) = E(Y_{it} - \tilde{Y}_{it}(0) | D_{g_i t} = 1) = E(\tilde{\rho}_{it} | D_{g_i t} = 1)$$

, i.e. λ_1 is numerically identical to the weighted average of treatment effect approximations $\tilde{\rho}_{it}$ from the fourth step of the imputation approach, after setting all w^*_{it} to one over the number of treated (i, t).

In a case where event-time specific average treatment effects are of interest, Gardner (2021) proposes to replace the binary treatment dummy in the third step by the event-time indicators $D_{g_i t,0}, ..., D_{g_i t,R}$. By the same line of argumentation, it's straightforward to show that the event-time coefficients from this regression are identical to (appropriately) weighted averages of treatment effect approximations obtained from the fourth step of the imputation approach. As a result, the two stage DiD estimator is in fact just a special case of the imputation estimator.

To draw inference, Gardner (2021) interprets his two stage DiD procedure, in contrast to Borusyak et al. (2021), as a generalized method of moments estimator.

4.3 Sun and Abraham (2020)

Setting

Sun and Abraham (2020) consider the binary treatment setting with staggered adoption. They propose an estimator that fixes the problems of dynamic TWFE specifications when treatment effects are heterogeneous. In particular, their goal is to identify event-time specific average treatment effects (ATE_r).

[4] Gardner (2021) actually uses group indicators in the regression on the untreated stratum (step 1), whereas Borusyak et al. (2021) use unit indicators. It can be shown that both approaches are respectively numerical identical with unit and group indicators. The treatment effect approximations from step 3 of the imputation approach obviously differ between the unit and group indicator versions. However, these differences completely vanish in the aggregation step, as long as all observations within a (g, t) receive the same weight. This result follows from two findings: First, by the FWL theorem, the time indicator coefficients from regression (13) are identical to the time indicator coefficients in the same regression with group indicators. Second, the average of all unit indicator coefficients from regression (13) of the same group is numerically identical to the respective group indicator. The second observation can be validated by taking into account that both regression versions are saturated in unit/ group indicators, if the time indicator coefficients are fixed (what follows from the first finding). I use unit indicators here for both approaches, because the proof is somewhat more compelling.

In compliance with the authors, I use the dynamic potential outcomes notation introduced in section 3.2.

Assumptions

A.1 CTA.

A.2 No treatment anticipation.

A.3 Existence of a never-treated group.

I will demonstrate later that assumptions 2 and 3 can be easily relaxed.

Identification

The 'interaction-weighted' estimator takes the following form:

1. Specify a TWFE regression, that includes an interaction term between each treatment group indicator (except for the never-treated group) and each applicable event-time indicator (except for event-time -1):

$$Y_{it} = \hat{\alpha}_{g_i} + \hat{\beta}_t + \sum_{\bar{g} \neq \infty} \sum_{r \neq -1} \mu_{\bar{g},r} I(g_i = \bar{g}) D_{g_i t,r} + \epsilon_{it} \tag{16}$$

2. For each event-time r of interest, build the following weighted average of interaction term coefficients from step 1:

$$\sum_{\bar{g}} \mu_{\bar{g},r} P(g_i = \bar{g} | g_i \in \{1 - r, ..., T - r\})$$

   $P(g_i = \bar{g} | g_i \in \{1 - r, ..., T - r\})$ is the share of units in group $\bar{g}$ among all units that experience event-time r, i.e. $i : 1 \leq g_i + r \leq T$.

Note that regression (16) could be equivalently written by including interactions between group indicators and period indicators, where the period indicator for $g_i - 1$ is excluded for each group. This follows from the definition of event-times. From a purely mechanical perspective, it's necessary to exclude one event-time from the regression specification (16), to avoid perfect multicollinearity. This follows from the same line of argumentation as in the 'classical' dynamic TWFE specification (8), since for each (i, t): $\sum_{\bar{g}} I(g_i = \bar{g}) D_{g_i t,r} = D_{g_i t,r}$ holds. Also, the never-treated group needs to be excluded, since $I(g_i = \infty) D_{g_i t,r} = 0$ for any event-time what again causes a multicollinearity problem. The never-treated group then takes the role of a reference group in the specification and event-time -1 takes the role of a reference event-time.

It's easy to verify that the TWFE regression (16) is saturated – it includes a separate parameter for each (g, t). The population regression function identifies the CEF completely non-parametrically and the interaction coefficients identify the associated GATEs under the identifying assumptions:

$$\begin{aligned} \mu_{\bar{g},r} &= [E(Y_{i(g+r)} | g_i = \bar{g}) - E(Y_{i(g-1)} | g_i = \bar{g})] - [E(Y_{i(g+r)} | g_i = \infty) - E(Y_{i(g-1)} | g_i = \infty)] \\ &= GATE_{\bar{g},\bar{g}+r} \end{aligned}$$

In a sense, this approach can be seen as a direct consequence of the point I made about the relation between population regression function and CEF in section 3.2 (equation (9)): The CEF is generally not linear in the event-time indicators (and the group/ time indicators) themselves, but it's always linear by defining a single variable for each (g, t). Due to the no anticipation assumption, all interactions involving negative event-times could be in principle dropped, because the regression function still fully resembles the CEF. Note that regression (16) also coincides with the 2x2 DiD regression (2) in the 2x2 setting.

Step 2 then builds a weighted sum of all interaction term coefficients (i.e. GATEs) for event-time r. The weighted sum then identifies ATE_r, since:

$$\sum_{\bar{g}} GATE_{\bar{g},\bar{g}+r} P(g_i = \bar{g} | g_i \in \{1-r, ..., T-r\}) = E(\rho_{it} | D_{g_i t,r} = 1) = ATE_r$$

If there is some anticipatory behaviour one can simply exclude a pre-treatment event-time outside of the effect window (k) instead of event-time -1 in regression (16) to avoid complications. The mechanics remain identical, since only the reference period in the 2x2 DiD estimand from above changes. Intuitively, the population then needs to be restricted to units with $g_i + k \geq 1$, since the 2x2 DiD estimand from above wouldn't be defined for all groups with $g_i + k < 1$. If there are (additionally) no never-treated units, one can exclude all interaction terms involving $g_i = \max(g_i)$ from the specification and then run the regression only on (i, t) with $t \in \{1, ..., \max(g_i) + k\}$. The restriction of the panel's time dimension makes sure that there is always a clean control group.

Inference

It's straightforward to show the validity of this approach in a random sample of the population. The specification of population regression (16) can be directly transferred to a random sample. Unbiasedness and consistency of the estimators for $\mu_{\bar{g},r}$ follows from the fact that the regression is correctly specified. The conditional probabilities can be also estimated consistently be sample shares. It then follows from the known calculus rules for probability limits that the sample analogue of the term in step 2 is a consistent estimator for ATE_r. Sun and Abraham (2020) also establish unbiasedness and asymptotic normality under standard assumptions.

4.4 Callaway and Sant'Anna (2020)

Setting

Callaway and Sant'Anna (2020) also consider the setting with a binary treatment and staggered adoption. Their framework allows to summarize GATEs in a very flexible way and is able to accommodate scenarios, where common trends may only hold after conditioning on time-constant covariates. To achieve the latter goal they build on prior work, among others the 'semiparametric' DiD estimator proposed by Abadie (2005). After introducing the identifying assumptions, I will shortly characterize this approach to facilitate the understanding of what follows. They also show that the results can be derived by relying on other estimators, namely the 'outcome regression' approach

by Heckman et al. (1997) and the 'doubly robust' estimator of Sant'Anna and Zhao (2020). From an identification standpoint all three approaches are equivalent, but they differ in terms of inference. I therefore constrain myself to the approach by Abadie (2005). Notation-wise, I use the dynamic potential outcomes framework introduced in section 3.2.

Assumptions

A.1 Conditional common trends assumption under no treatment (CCTA). For all $(\bar{g}, \bar{t}) \in \{2, ..., T\}^2$:

$$E(Y_{i\bar{t}}(\infty) - Y_{i\bar{t}-1}(\infty)|g_i = \bar{g}, x_i = \bar{x}) = E(Y_{i\bar{t}}(\infty) - Y_{i\bar{t}-1}(\infty)|x_i = \bar{x})$$

, where x_i is a vector of observed time-constant covariates.

For the same value of x_i, all groups are assumed to have the same time trend. This is a modification of the conventional CTA, where groups are assumed to experience the same time trend without conditioning on covariates. In some cases, the CCTA may be more plausible than the CTA. Consider for instance the setting studied by Ashenfelter (1978), who is interested in the effect of a job market training program on earnings. The treatment group is given by all individuals who eventually receive the training and the control group by all other individuals. The conventional CTA would fail in this setting, if individuals who experience a negative labour market shock in pre-periods are more likely to receive (or choose) the treatment: In absence of the treatment the evolution of the outcome in the treatment group would be expected to be 'more positive' than in the control group, because the negative shock vanishes. In fact, Ashenfelter (1978) finds that one period prior to the treatment earnings of trainees significantly drop relative to the control group, what is commonly referred to as 'Ashenfelter's dip' in the literature. More generally, Ashenfelter's dip describes the failure of the CTA if selection to treatment depends on transitory shocks on the outcome variable in pre-treatment periods. The non-parallel (unconditional) trends then occur, because the shocks differ systematically between treatment and control group. If one would just compare the outcome evolution of units with the same magnitude of pre-treatment shock in both groups, trends would be however parallel. So x_i could be the difference in individual earnings of individual i between periods $t-2$ and $t-1$, if t is the first period under treatment.

Generally, the CCTA does not imply the CTA, as long as the distribution of x_i differs in both groups. To see this, consider two treatment groups $\bar{g} \neq g'$, where the distribution of the scalar-valued (discrete) x_i is not identical in both groups. Under

the CCTA and by using the law of iterated expectation it holds that:

$$
\begin{aligned}
&E(Y_{i\bar{t}}(\infty) - Y_{i\bar{t}-1}(\infty)|g_i = \bar{g}) \\
&= E[E(Y_{i\bar{t}}(\infty) - Y_{i\bar{t}-1}(\infty)|g_i = \bar{g}, x_i)] \\
&= \sum_{\bar{x}} E(Y_{i\bar{t}}(\infty) - Y_{i\bar{t}-1}(\infty)|g_i = \bar{g}, x_i = \bar{x}) P(x_i = \bar{x}|g_i = \bar{g}) \\
&= \sum_{\bar{x}} E(Y_{i\bar{t}}(\infty) - Y_{i\bar{t}-1}(\infty)|g_i = g', x_i = \bar{x}) P(x_i = \bar{x}|g_i = \bar{g}) \\
&\neq \sum_{\bar{x}} E(Y_{i\bar{t}}(\infty) - Y_{i\bar{t}-1}(\infty)|g_i = g', x_i = \bar{x}) P(x_i = \bar{x}|g_i = g') \\
&= E(Y_{i\bar{t}}(\infty) - Y_{i\bar{t}-1}(\infty)|g_i = g')
\end{aligned}
$$

A.2 Limited treatment anticipation. There is a known $\delta \geq 0$ such that for all $t < g_i - \delta$:

$$Y_{it} = Y_{it}(\infty)$$

Assumption 2 simply requires that for all event-times smaller than $-\delta$, there is no anticipatory behaviour.

Semiparametric DiD estimator by Abadie (2005)[5]

Consider the generic 2x2 setting and the notation introduced in section 2.2, but assume that for $Y_{it}(0)$ the CCTA holds instead of the CTA. The simple 2x2 DiD estimand (1) then no longer generally identifies $E(Y_{i2}(1) - Y_{i2}(0)|g_i = 2)$, because the unconditional time trend of group 1 is in general not identical to the unconditional time trend of group 2. This directly follows from the above reasoning regarding the CCTA. It's therefore necessary to find another way to identify the treatment group's time trend based on observed outcomes. The basic idea is to manipulate the control group's observed time trend, $E(Y_{i2}(0) - Y_{i1}(0)|g_i = 1)$, to resemble the treatment group's time trend by making use of the CCTA.

In particular, the unobserved time trend of group 2 can be expressed in the following way:

$$
\begin{aligned}
&E(Y_{i2}(0) - Y_{i1}(0)|g_i = 2) \\
&= \sum_{\bar{x}} E(Y_{i2}(0) - Y_{i1}(0)|g_i = 1, x_i = \bar{x}) \frac{P(x_i = \bar{x}|g_i = 2)}{P(x_i = \bar{x}|g_i = 1)} P(x_i = \bar{x}|g_i = 1) \\
&= \sum_{\bar{x}} E\left((Y_{i2}(0) - Y_{i1}(0)) \frac{P(x_i = \bar{x}|g_i = 2)}{P(x_i = \bar{x}|g_i = 1)} |g_i = 1, x_i = \bar{x} \right) P(x_i = \bar{x}|g_i = 1) \\
&= E\left[E\left((Y_{i2}(0) - Y_{i1}(0)) \frac{P(x_i|g_i = 2)}{P(x_i|g_i = 1)} |g_i = 1, x_i \right) \right] \\
&= E\left((Y_{i2}(0) - Y_{i1}(0)) \frac{P(x_i|g_i = 2)}{P(x_i|g_i = 1)} |g_i = 1 \right) \\
&= E\left((Y_{i2}(0) - Y_{i1}(0)) \frac{P(g_i = 2|x_i)/P(g_i = 2)}{(1 - P(g_i = 2|x_i))/(1 - P(g_i = 2))} |g_i = 1 \right)
\end{aligned}
$$

[5]This paragraph partly follows lecture notes of Clément de Chaisemartin (de Chaisemartin, 2021).

The third and fourth equalities hold due to the law of iterated expectation and the last equality follows from Bayes' theorem.

Hence, $E(Y_{i2}(1) - Y_{i2}(0)|g_i = 2)$ is identified by the following estimand:

$$E(Y_{i2} - Y_{i1}|g_i = 2) - E\left((Y_{i2} - Y_{i1})\frac{P(x_i|g_i = 2)}{P(x_i|g_i = 1)}|g_i = 1\right) \tag{17}$$

Put differently, the treatment group's time trend can be identified by the time trend of the control group, after weighting each unit's difference in observed outcomes appropriately. $P(x_i|g_i = 2)$ should be interpreted as the share of units in group 2, that has the same value of the covariate as unit i from group 1.[6] If a unit's x_i is more prevalent in the treatment group than in the control group, it receives a larger weight and vice versa. By weighting the control group units, the distribution of x_i becomes the same in the treatment and the (weighted) control group. Since the time trend only depends on x_i, treatment and control groups then have the same overall time trend.[7] Note that the estimand is only defined, if $0 \leq P(g_i = 2|x_i = \bar{x}) < 1$ for all possible values of $\bar{x}$: If there are no units in the control group with $x_i = \bar{x}$, it becomes impossible to resemble the distribution of x_i in the treatment group by appropriately weighting the control group. Also note that if x_i has the same distribution in the treatment and in the control group, (17) collapses to the generic 2x2 DiD estimand (1). This follows from the previously established fact, that the CCTA implies the CTA in this case.

To illustrate the mechanics, consider a simple example: Assume that x_i is binary, $x_i \in \{0, 1\}$, with $P(x_i = 1|g_i = 2) = 1$ and $P(x_i = 1|g_i = 1) = 0.5$. Hence, for all units in the $g_i = 1$ group with $x_i = 0$, $(Y_{i2} - Y_{i1})$ is set to zero, and for all units with $x_i = 1$, $(Y_{i2} - Y_{i1})$ is weighted up by factor two. Put differently, each unit with $x_i = 0$ is imputed by a duplicate of one of the units with $x_i = 1$. This is effectively equivalent to taking the average only on the subset of units in the control group with $x_i = 1$. Since the units of the reweighted control group all have the same value of x_i as the units in the treatment group, it follows from the CCTA that their time trends are identical.

The weighted 2x2 DiD estimand (17) can be estimated with a two step procedure: In the first step, $P(g_i = 2|x_i)$ can be either estimated non-parametrically or by a logit, probit, or linear probability regression for each unit in the sample. The unconditional probability $P(g_i = 2)$ can be estimated by the sample share. The estimator for the weighted 2x2 DiD estimand is then simply the usual 2x2 DiD estimator, with the control group units weighted by the weights from the first step.

Identification

1. Identify all possible GATEs based on a never-treated group or all not-yet-treated groups.

    a) Identification based on a never-treated group:

[6] A more unambiguous notation would be probably given by $P(x_j = x_i|g_j = 2)$.

[7] In practice researchers often include unit and time dependent covariates to their TWFE specification, to relax the CTA. In contrast to the approach by Abadie (2005), this practice however clearly requires strong parametric assumptions.

For each treatment group $\bar{g} \in \{2+\delta, 3+\delta, ..., T\}$ with $\bar{t} \in \{\bar{g}-\delta, \bar{g}-\delta+1, ..., T\}$ identify all $GATE_{\bar{g},\bar{t}}$ by the weighted 2x2 DiD estimands:

$$E(Y_{i\bar{t}} - Y_{i(\bar{g}-\delta-1)}|g_i = \bar{g}) - E\left((Y_{i\bar{t}} - Y_{i(\bar{g}-\delta-1)})\frac{P(x_i|g_i = \bar{g})}{P(x_i|g_i = \infty)}|g_i = \infty\right)$$

b) Identification based on all not-yet-treated groups:

Define the sets for $\bar{g}$ and $\bar{t}$ as in (a). Identify all $GATE_{\bar{g},\bar{t}}$ by the weighted 2x2 DiD estimands:

$$E(Y_{i\bar{t}} - Y_{i(\bar{g}-\delta-1)}|g_i = \bar{g}) - E\left((Y_{i\bar{t}} - Y_{i(\bar{g}-\delta-1)})\frac{P(x_i|g_i = \bar{g})}{P(x_i|g_i > \bar{t}+\delta)}|g_i > \bar{t}+\delta\right)$$

If anticipatory effects are of no interest, adjust the set for $\bar{t}$ to $\{\bar{g}, \bar{g}+1, ..., T\}$.

2. Build a weighted sum of all $GATE_{\bar{g},\bar{t}}$, identified by the weighted 2x2 DiD estimands from step 1:

$$\sum_{(\bar{g},\bar{t})} w^{CS}_{\bar{g}\bar{t}} GATE_{\bar{g},\bar{t}}$$

, where $w^{CS}_{\bar{g}\bar{t}}$ are weights specified by the researcher.

The identification strategy of Callaway and Sant'Anna (2020) relies on a separate (weighted) 2x2 DiD estimands for each $GATE_{\bar{g},\bar{t}}$. Only two minor adjustments are necessary to transfer the semiparametric DiD estimand (17) from the 2x2 setting to the multi-period staggered adoption setting: First, one needs to take into account anticipatory behaviour when choosing a reference period, to avoid contamination. Under the assumption that $-\delta$ is the smallest (non-positive) event-time where anticipatory behaviour occurs, it's natural to choose $\bar{g}-\delta-1$ as a reference period to identify all GATEs of group $\bar{g}$. Of course, without anticipatory behaviour, i.e. $\delta = 0$, this implies $\bar{g}-1$ as a reference period. Second, it's necessary to pick a 'clean' control group, which has not experienced any kind of treatment effect at time $\bar{t}$, i.e. $g_i > \bar{t}+\delta$. This could be either the never-treated group or all not-yet-treated groups at a given point in time. If a never-treated group is available, it's possible to identify GATEs of all groups over the whole time span where clean contrasts are available, i.e. $(\bar{g},\bar{t}) \in \{2+\delta, ..., T\} \times \{\bar{g}-\delta, ..., T\}$. In case of no never-treated group, one can instead rely on the last-treated group as a control group, what then allows to identify only GATEs for all groups and time periods before the last group's anticipation window starts, i.e. $(\bar{g},\bar{t}) \in \{2+\delta, ..., \max(g_i)-\delta-1\} \times \{\bar{g}-\delta, ..., \max(g_i)-\delta-1\}$.[8] Researchers sometimes may want to choose the last-treated group (or another group) as a control group, even if a never-treated group is available, if the never-treated group differs strongly from the other groups. Lastly, one can also pick all groups not-yet-treated by time $\bar{t}+\delta$ simultaneously as a control group. This has the advantage that all available information is exploited to identify the time trend, what is particularly relevant for statistical inference. On the downside, this practice induces the control group to change its composition over time as more groups become treated.

[8] Callaway and Sant'Anna (2020) note, that there is a trade-off between treatment anticipation and the CCTA: Under no anticipation, the CCTA only needs to hold for $\bar{t}-(\bar{g}-1)$ periods. If one wants to allow for treatment anticipation, the time span increases in the effect window, i.e. $\bar{t}-(\bar{g}-\delta-1)$. This trade-off needs to be handled in a case-specific way.

After all GATEs of interest are identified, they can be aggregated in different ways by choosing appropriate weights in step 2. Next to the overall ATT and event-time specific average treatment effects (which resemble the identification targets of the static and dynamic TWFE specifications respectively), they also propose weights for calendar-time specific and group-specific aggregation schemes. In each case, the weights are simply given by population shares of groups with respect to the relevant reference group. In principle, there are however almost no limits in this framework to construct weights in order to identify policy relevant treatment effect summary measures.

Note that the approach is numerically identical to the approach of Sun and Abraham (2020) if the CTA instead of the CCTA is imposed, by setting the weights appropriately and choosing the never-treated group as a control group.

Inference

Estimation of single GATEs can be conducted analogously to the standard 2x2 setting, where the propensity score needs to be estimated separately for each eventually-treated group (and time, if all not-yet-treated groups form the control group). To estimate the treatment effect summary measures from step 2 they propose a plug-in type estimator, that replaces each theoretical GATE and the associated weight by the respective sample counterparts. They also establish asymptotic normality of GATEs and aggregated treatment effect measures and propose a bootstrap procedure for asymptotically valid simultaneous inference.

4.5 Comparison

Given that the two-stage approach of Gardner (2021) is a special case of the imputation approach by Borusyak et al. (2021) (hereafter 'B,J&S'), there is no need to treat them separately. The same holds for the interaction-weighted estimator of Sun and Abraham (2020) with respect to the approach by Callaway and Sant'Anna (2020) (hereafter 'C&S').

All three approaches have in common, that they manually separate the identification and aggregation of treatment effects to mitigate the improper weighting under treatment effect heterogeneity in TWFE regressions: The approaches of C&S and de Chaisemartin and D'Haultfœuille (2020) (hereafter 'dC&DH') rely on the well understood 2x2 DiD estimand to nonparametrically identify aggregated group and time specific average treatment effects as building blocks. B,J&S in contrast stick to the TWFE principle, but avoid complications by identifying unit and time effects only based on untreated units and subsequently directly estimate treatment effects.

The approaches of C&S and B,J&S are robust to arbitrary forms of treatment effect heterogeneity and are suitable to flexibly identify (almost) any estimand of treatment effects. In contrast, the approach of dC&DH is only robust to some forms of treatment effect heterogeneity and is limited to a specific aggregation scheme. This can be rated as a clear advantage of the former two.

A unique feature of the C&S approach is, that it allows for non- or semiparametric identification in cases where common trends only hold after conditioning on (time-

constant) observed covariates. On the other hand their estimator is limited to staggered adoption settings, whereas the other two are also applicable in non-staggered settings.

The imputation estimator might be favourable when it comes to inference, since it exploits all observations for identification, whereas the other two approaches necessarily always only use, potentially very small, subsets of the data. Borusyak et al. (2021) demonstrate the resulting efficiency gain of their imputation estimator with Monte-Carlo simulations. However, this comes at the cost that the CTA needs to hold over the whole time span, whereas for the C&S'A estimator the (C)CTA only needs to hold for the 'relevant' periods. In practice, it might be admittedly difficult to argue that the weaker version holds while the stronger version does not.

Finally, all estimators are implemented in the broadly used statistical packages Stata and *R*. Table 1 provides a systematic overview of all estimators.

5 Empirical Application: Flood Insurance Take-up in the US

In this section, I apply the different estimators introduced in the previous section to real world data and compare the results to the TWFE estimator. The purpose of this exercise is to get a feeling, whether the theoretical shortcomings of TWFE specifications translate into sizeable differences in applications.

In particular, I use data from Gallagher (2014), who studies the effect of regional flood events on flood insurance take-up in the US using the dynamic TWFE specification. He finds that flood insurance take-up spikes in communities affected by a flood in the subsequent years, but later returns to the baseline level. One may expect treatment effect heterogeneity with respect to the business cycle in this context. It also seems possible that the severity of flood damages varies across time (e.g. due to climate change) and thereby the propensity to participate in an insurance scheme in the aftermath of a flood.

After a short characterization of the institutional background and data sources, I replicate one of the main regression specifications and compare the results to the alternative estimators from the previous section. Given that the estimator by de Chaisemartin and D'Haultfœuille (2020) is not able to capture dynamic effects, I limit my comparison to the approaches by Borusyak et al. (2021) and Callaway and Sant'Anna (2020).

5.1 Institutional Background and Data

The 'National Flood Insurance Program' (NFIP) was created by the federal government of the US in 1968. It allows homeowners to purchase a flood insurance at premiums aligned to the local flooding risk. Homeowners can decide each calendar year whether they want to participate in the insurance scheme. Importantly, homeowners can simply drop out by not paying the premium for the subsequent year. Gallagher (2014) relies on NFIP data aggregated at the community-year level for his analysis. He additionally uses requests from the so called 'Presidential Disaster Declaration' (PDD) system as a

data source for regional floods. The PDD system is a mechanism in the US that allows counties to apply for federal assistance in the aftermath of natural disasters. He is further able to identify the communities within each county, which were hit by a flood from the location of damaged public infrastructure.

Using these data sources, Gallagher (2014) constructs a balanced panel of 10,841 communities (cities, towns, villages etc.) distributed across 2,725 counties for the time period 1990-2007. One observation (i, t) corresponds to a community i in year t in this context. He additionally constructs a longer panel spanning the time period 1980-2007. Due to limitations of the PDD data, it's however not feasible to verify which communities within a PDD county were actually affected by the flood event.

One important peculiarity of the setting at hand is, that communities can be hit by a regional flood in multiple years: In the 1990-2007 panel, 3,092 communities were hit by one and 3,822 by two or more flood events. In the dynamic TWFE specification this circumstance can be easily accommodated, by allowing multiple event-time indicators to be active for a given observation. E.g., if a community is hit by a flood in t and in $t + 2$, the indicators for event time 0 and -2 are equal to 1 in year t. However, this practice clearly additionally requires additivity of treatment effects. Since the alternative estimators do not allow to accommodate multiple treatments, I restrict the sample to all communities hit by a flood event at most once in the time period 1990-2007 (7,019 communities) for the comparison.

In this context it's also worth mentioning that communities could have already been subject to a flood event prior to 1990. Gallagher (2014) does not control for this circumstance when working with the 1990-2007 panel. If this is the case, lagged treatment effects of pre-1990 floods are incorrectly attributed. As long as flood trajectories of communities do not follow a systematic pattern (e.g. floods occur every four years), this circumstance is however uncritical for identification. Given that the restricted 1990-2007 panel contains only communities which were hit at most once in 18 years, this is not too implausible. The 1980-2007 panel allows to account for this, since PDD requests at the county level are available from 1958 onward. Unfortunately, using the longer panel would reduce the sample size dramatically when removing all communities with multiple treatments due to the coarser definition of treated communities: As soon as one community within a county is affected by a flood, all communities within that county are coded as treated. I therefore stick to the 1990-2007 panel for the comparison, but acknowledge the imperfection.

Finally, the spatial structure of the data also causes some difficulties. It's plausible to assume that flood insurance take-up does not only respond in counties directly hit by a flood, but also in other communities, especially those located close to the affected community. In a more technical sense, these spillover effects correspond to a violation of the well known 'stable unit treatment value assumption' (SUTVA). Gallagher (2014) presents evidence that flood insurance take-up also increases, but to a much smaller extent, in neighbouring communities by including additionally a set of event-time indicators for neighbouring communities/ counties to the dynamic TWFE specification. Practically this means, that estimates are generally likely downwards biased in tendency, given that some actually treated communities are used as control units.

Since regional floods are random natural events it's reasonable to assume no anticipatory behaviour in this context.

5.2 Replication

Gallagher (2014) uses the log of flood policies per person in community i in year t ($ln(takeup_{it})$) as the outcome variable for his dynamic TWFE specification. In his main specification the event-times -17 to 11 and 11 to 17 are binned to a single indicator respectively. He justifies this practice with the gain in statistical power, given that these event-time coefficients are naturally identified by relatively few observations. Importantly, his specification also slightly deviates from the specification introduced in section 3.2, by including state by year indicators instead of simple year indicators, i.e. each year indicator is partitioned into 48 indicators - one for each state. Thereby he relaxes the CTA by only requiring communities within a state to follow the same time trend. The practice of including group specific time indicators is actually very common in the applied literature. In principle it would be even feasible to include county by year indicators here; given the large number of counties this would however induce almost 50,000 indicators, thereby wasting too many degrees of freedom.

His TWFE specification therefore takes the following explicit form:

$$ln(takeup_{it}) = \hat{\alpha}_i + \hat{\beta}_{st} + \mu_{\leq -11} D_{it,\leq -11} + \sum_{r=-10}^{10} \mu_r D_{it,r} + \mu_{\geq 11} D_{it,\geq 11} + \epsilon_{it} \qquad (18)$$

Figure 6 a) replicates Figure 2 in Gallagher (2014). Depicted are coefficient estimates from specification (18) together with 95% confidence intervals, based on the full 1990-2007 sample. He interprets the statistically insignificant pre-treatment point estimates as evidence in favour of the CTA. Insurance take-up increases by 8% in the year of a flood and peaks at 9% in the following year. Take-up then declines steadily and returns to its baseline level approximately ten years after the flood event. Figure 6 c) corresponds to the same specification estimated on the subsample of all counties treated at most once. The coefficient pattern is fairly similar in both samples, but the confidence intervals are much wider for the subsample, what reflects the smaller sample size. Figures b) and d) correspond to the same samples as a) and c) respectively, but the specifications do not include binned event-time indicators. Ultimately, figures e) and f) are estimated by specifications with simple year indicators on the subsample. The main pattern remains stable in the binned specification; the point estimates for the post-treatment periods are only slightly higher. In the specification without binning, the point estimates for event times 15-17 do however not really fit to the established pattern here. One potential explanation for this odd pattern, which can be also observed in figure 6 d), could be that the event-time coefficients are identified by different sets of communities. In particular, the coefficients for event-times 15-17 are only identified by very few communities treated at the very beginning of the panel, which might experience a vastly different treatment effect pattern. Figure 7 depicts the distribution of event-times across communities in the panel. Even more importantly, there are 102 communities in the subsample, which were hit by the flood event in 1990. These units are always-treated and are in the same time the only units, which can identify the

coefficient for event-time 17. From a DiD perspective always-treated units are however not eligible to identify treatment effects, given that that there is no pre-treatment period. I discussed this issue in more detail in section 3.2. The coefficient for event-time 17 in the dynamic TWFE specification therefore necessarily needs to be a result of unwarranted extrapolation. The coefficients for event-times 15 and 16 are also likely strongly affected by the 1990-treatment cohort, given that the 1991 and 1992 cohorts are very small.

5.3 Comparison to Alternative Estimators

In contrast to the (dynamic) TWFE specification, the alternative approaches foreclose the use of always-treated units ex ante: The imputation approach prevents this pitfall simply due to the fact that it's impossible to estimate unit fixed effects for always-treated units from the untreated stratum (always-treated units do not appear in the untreated stratum). With the C&S approach, estimation of always-treated units' treatment effects is also impossible, because a pre-treatment period is required. To make sure that results are not solely driven by compositional differences between samples, I therefore additionally drop the 102 always-treated communities.

While the imputation approach can accommodate state-specific time trends naturally and is therefore able to fully reproduce the original TWFE specification in Gallagher (2014), this is unfortunately not the case for the C&S estimator. In principle, it would be possible to subset the dataset by states and apply the estimator for each state separately. In a second step, state specific GATEs could be aggregated to overall GATEs again. However, there is no such (implemented) extension so far and inference is also not straightforward in this case. Given that more complex sets of fixed effects are very common in the applied literature, this can be evaluated as a clear disadvantage of the C&S estimator. In order to allow for the estimation of a particular state by year fixed effect in the first step of the imputation estimator, it's required to have a non-empty set of untreated communities in the same state and year. This follows by the same argumentation as before. Therefore, it becomes impossible in a staggered adoption setting to estimate state by year fixed effects for a state, as soon as all communities within that state are treated. For this reason I drop one more community from the sample. My final sample then consists of 6,916 communities.

Figure 8 a) plots coefficient estimates together with 95% confidence intervals from the dynamic TWFE specification (18) and the imputation estimator, where distant event-times are binned as before. Estimates from the TWFE estimator are displayed in blue and estimates from the imputation estimator are displayed in green here. Recall that the imputation estimator does not produce coefficients for pre-treatment periods by default. The displayed pre-treatment coefficients stem from the pre-trends test described in section 4.2, that only uses the untreated observations. I use $K = 16$ for the regression on the untreated stratum (15) in all depicted settings. To check whether potential differences for post-treatment event-times only stem from the inclusion of pre-treatment event-times, I also depict the coefficient estimates from TWFE specifications without lead indicators (non-solid blue circles). The post-treatment estimates from the imputation estimator are in general larger than those from the TWFE estimator (except for the binned event-time coefficient), but differences are modest. Differences

to the 'lags only' TWFE specification are even smaller. Figure 8 b) displays the same comparison but without binning. Note that the imputation estimator's coefficients for event-times 0-10 are numerically identical in both figures, because they are averages of the same sets of treatment effect estimates respectively. This is in principle not the case for the TWFE coefficients, even though they remain stable here. Deviations become more sizeable for the previously binned event-times, but are generally very noisy and in both cases statistically insignificant. In pre-treatment event-times, the imputation estimates are however systematically and sizeably larger than the TWFE estimates, while the distance between the estimates stays approximately constant across the event-times. The reason for this observation is probably, that regression (15) uses event-time -17 as a reference, whereas the TWFE specification uses event-time -1 as a reference. After normalizing for level differences, both sets of coefficients are again similarly close to one another as in post-treatment event-times.

Figure 8 c) is identical to figure 8 a), except that I include year indicators instead of state by year indicators. Interestingly, the estimates from the TWFE specification and the imputation estimator are almost indistinguishable up to event time 11. Unluckily, there exists so far no off-the-shelf implementation to create binned coefficients for the C&S estimator. Figure 8 d) presents the same setting but without binning. Estimates from the C&S estimator are additionally displayed in red here. They generally also differ only slightly and not systematically from the TWFE and imputation estimates for post-treatment event-times. For pre-treatment event-times, the C&S estimates are in tendency closer to zero than their TWFE counterparts. Somewhat surprisingly, the software implementations in both *R* and Stata deviate slightly from the proposed estimator in Callaway and Sant'Anna (2020): Note that there is a (non-zero) point estimate for event time -1, but no estimate for event time -17. In other words, the software implementation always uses 1990 as pre-period and not the year prior to the initial treatment.

6 Conclusion

TWFE regressions are a very common way to implement DiD research designs. In this thesis, I discussed potential shortcomings of TWFE regressions for causal inference when treatment effects are heterogeneous. The shortcomings are founded in the fact that TWFE is, beyond simple settings, a parametric approach that requires homogeneous treatment effects to identify 'well defined' causal parameters by regression coefficients. I first examined TWFE specifications with a single binary treatment variable: The treatment coefficient is a weighted sum of treatment effects, where weights can be far off their target values and can even become negative. Intuitively, the TWFE regression exploits all contrasts to identify the coefficient that would be admissible when treatment effects are homogeneous and thereby misallocates variation of the outcome variable. I then turned towards TWFE specification, where the single treatment indicator is replaced by a set of dummies indicating the relative time to the initial treatment period. While the weighting problem also arises in this case, event-time coefficients can additionally be contaminated by treatment effects from other event-times. The intuition for this result is similar to the intuition for the negative weighting problem:

TWFE implicitly assumes homogeneous treatment effects within each event-time and thereby performs unwarranted extrapolations across event-times. I also highlighted, that the restrictions on treatment effects become even more stringent with non-binary treatments.

Afterwards, I reviewed and compared alternative robust estimators proposed by de Chaisemartin and D'Haultfœuille (2020), Borusyak et al. (2021), Gardner (2021), Sun and Abraham (2020), and Callaway and Sant'Anna (2020). It turned out that the estimator by Sun and Abraham (2020) can be thought of as a special case of the estimator by Callaway and Sant'Anna (2020) and that the estimator by Gardner (2021) is a special case of the estimator by Borusyak et al. (2021). My comparison revealed that all three approaches have strengths and weaknesses that depend on the setting at hand. The estimator by Callaway and Sant'Anna (2020) can be favourable in cases where common trends might only hold after conditioning on time constant covariates, e.g. in the Ashenfelter's dip scenario. The approach by Borusyak et al. (2021) may be however more implementable in many applied settings, since it preserves much of the flexibility of TWFE specifications.

Finally, I applied the TWFE estimator and the estimators of Borusyak et al. (2021) and Callaway and Sant'Anna (2020) to real world data from Gallagher (2014), who studies the effect of regional flood events on flood insurance take-up in the US by using TWFE regressions. It turned out that differences between the three estimators are modest to negligible here. However, the results should be taken with a grain of salt due to potential violations of identifying assumptions. The comparison also reinforced the previous impression, that the application range of the imputation estimator by Borusyak et al. (2021) is generally broader.

In conclusion, it's fair to say that the literature on treatment effect heterogeneity in TWFE regressions greatly contributes to the understanding of regression anatomy. In practice, the alternative robust estimators can be a useful robustness check for the TWFE estimator. However, the TWFE estimator itself will likely remain a solid baseline approach in the future, mainly due to its higher flexibility.

References

Abadie, A. (2005): "Semiparametric Difference-in-Differences Estimators," *The Review of Economic Studies*, 72, 1–19.

Angrist, J. D. and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Ashenfelter, O. (1978): "Estimating the Effect of Training Programs on Earnings," *The Review of Economics and Statistics*, 60, 47–57.

Borusyak, K., X. Jaravel, and J. Spiess (2021): "Revisiting Event Study Designs: Robust and Efficient Estimation," *unpublished*.

Büttner, T. and B. Madzharova (2021): "Unit Sales and Price Effects of Preannounced Consumption Tax Reforms: Micro-level Evidence from European VAT," *American Economic Journal: Economic Policy*, 13, 103–134.

Callaway, B., A. Goodman-Bacon, and P. H. Sant'Anna (2021): "Difference-in-Differences with a Continuous Treatment," *unpublished, arXiv:2107.02637v2*.

Callaway, B. and P. H. Sant'Anna (2020): "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, forthcoming.

Davidson, R. and J. G. MacKinnon (2003): *Econometric Theory and Methods*, Oxford University Press.

de Chaisemartin, C. (2021): "Econometrics for 1st and 2nd year PhDs - UCSB," *Lecture Notes*.

de Chaisemartin, C. and X. D'Haultfœuille (2020): "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 110, 2964–2996.

——— (2021): "Difference-in-Differences Estimators of Intertemporal Treatment Effects," *unpublished*.

Feldstein, M. (1995): "The Effect of Marginal Tax Rates on Taxable Income: A Panel Study of the 1986 Tax Reform Act," *Journal of Political Economy*, 103, 551–572.

Gallagher, J. (2014): "Learning about an Infrequent Event: Evidence from Flood Insurance Take-Up in the United States," *American Economic Journal: Applied Economics*, 6, 206–233.

Gardner, J. (2021): "Two-stage Differences in Differences," *unpublished*.

Goodman-Bacon, A. (2021): "Difference-in-Differences with Variation in Treatment Timing," *Journal of Econometrics*, forthcoming.

Heckman, J. J., H. Ichimura, and P. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, 64, 605–654.

Imai, K. and I. S. Kim (2021): "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data," *Political Analysis*, 29, 405–415.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Sant'Anna, P. H. and J. Zhao (2020): "Doubly Robust Difference-in-Differences Estimators," *Journal of Econometrics*, 219, 101–122.

Stevenson, B. and J. Wolfers (2006): "Bargaining in the Shadow of the Law: Divorce Laws and Family Distress," *The Quarterly Journal of Economics*, 121, 267–288.

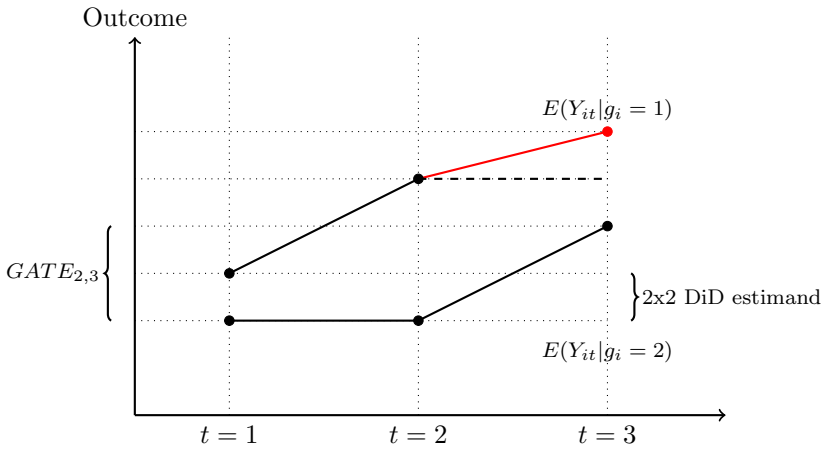
Strezhnev, A. (2018): "Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs," *unpublished*.

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

Wooldridge, J. M. (2021): "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators," *unpublished*.

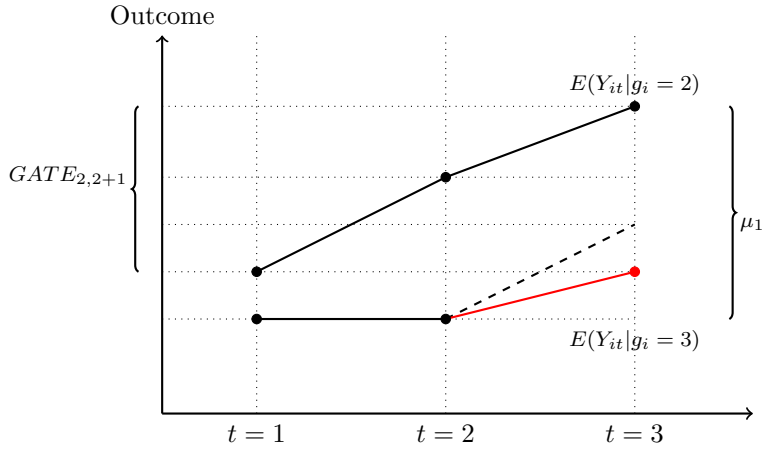
Appendix: Figures and Tables

Figure 1: Graphical Illustration of the Decomposition Result for μ in the Example with Two Groups and Three Periods



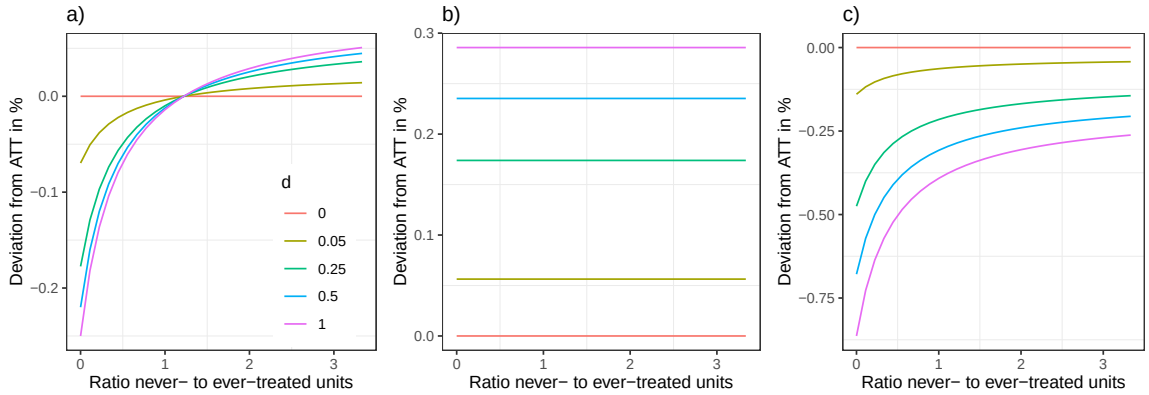
Notes: For simplicity I assume that there is no time trend here. Hence, changes in a group's average observed outcome over time are only due to treatment effects. The red solid line segment between $t = 2$ and $t = 3$ marks the case, where GATEs of group 1 are heterogeneous. Since $GATE_{1,3} > GATE_{1,2}$ in this example, it follows that the second 2x2 DiD estimand underestimates $GATE_{2,3}$. The black dashed line segment between $t = 2$ and $t = 3$ corresponds the case, where GATEs of group 1 are homogeneous.

Figure 2: Graphical Illustration of the Decomposition Result for μ_1 in the Example with Two Groups and Three Periods



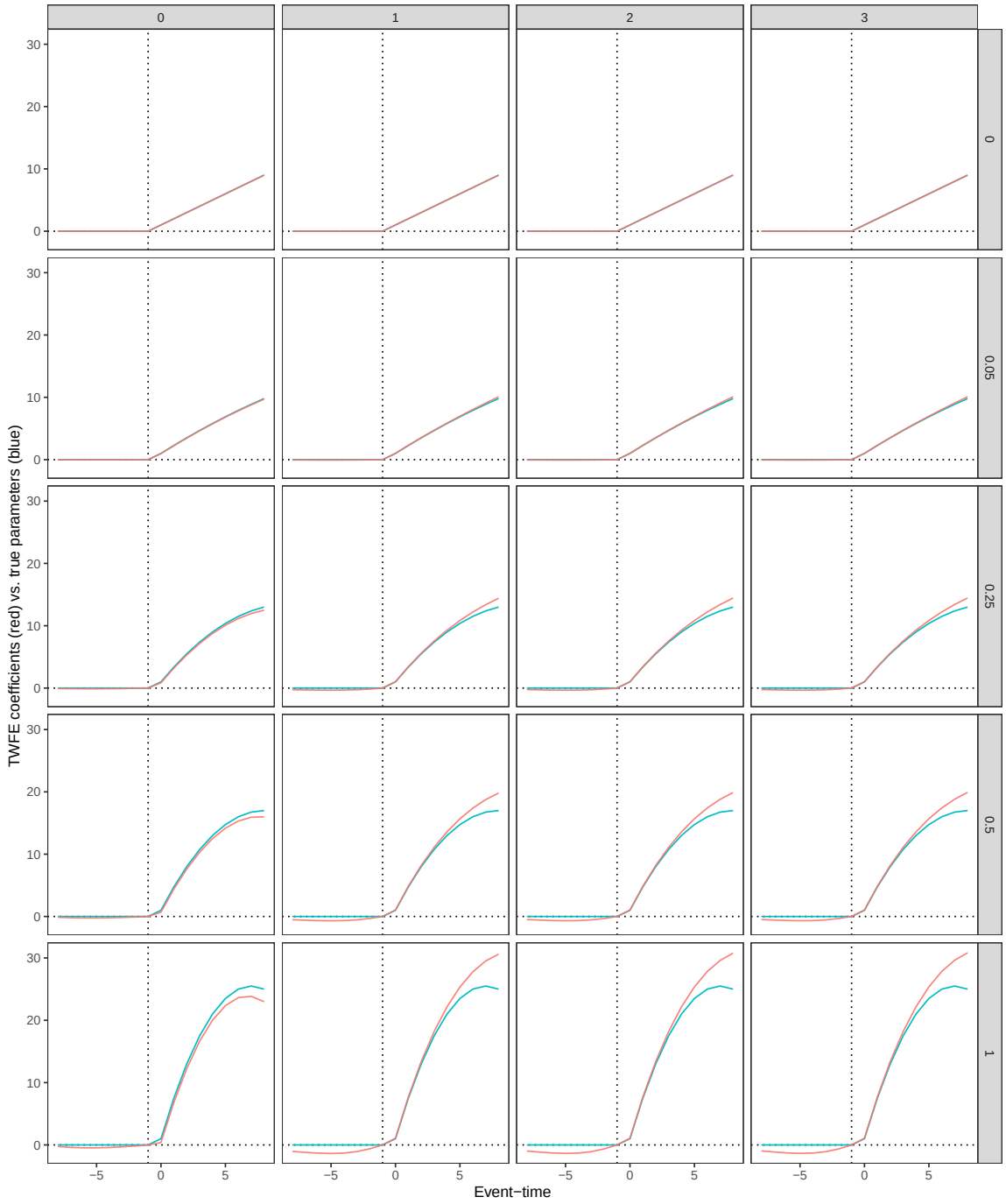
Notes: For simplicity I assume that there is no time trend here. Hence, changes in a group's average observed outcome over time are only due to treatment effects. The red line segment between $t = 2$ and $t = 3$ represents the case where $GATE_{2,2+0} \neq GATE_{3,3+0}$. Since $GATE_{2,2+0} > GATE_{3,3+0}$ in this example, it follows that μ_1 is larger than $GATE_{2,2+1}$. The black dashed line segment between $t = 2$ and $t = 3$ represents the case where $GATE_{2,2+0} = GATE_{3,3+0}$.

Figure 3: Effect of Heterogeneity and the Never-treated Group's Size on the Treatment Coefficient in the Static TWFE Specification



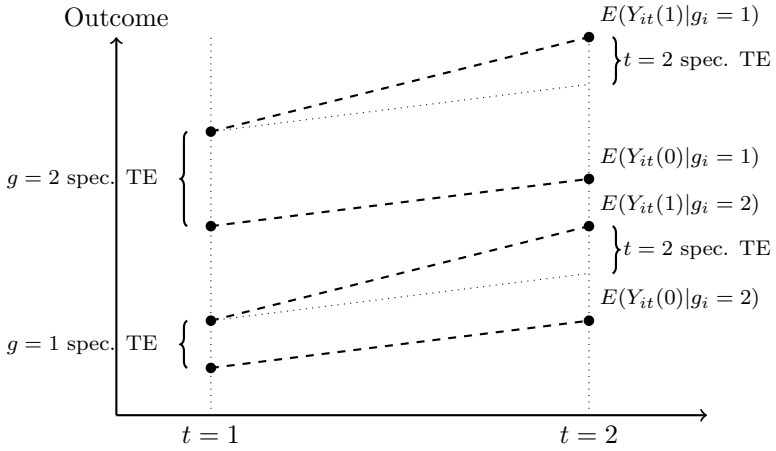
Notes: Panel a): Calendar time dependent treatment effect; Panel b): Group dependent treatment effect; Panel c): Event-time dependent treatment effect. The y-axis displays the deviation of the coefficient of the static TWFE specification from the ATT in %. The x-axis displays the ratio of the number of never-treated units to the sum of all ever-treated units. Note that the y-axis is not uniformly scaled across the panels.

Figure 4: Effect of Heterogeneity and the Never-treated Group's Size on Coefficients in the Dynamic TWFE Specification



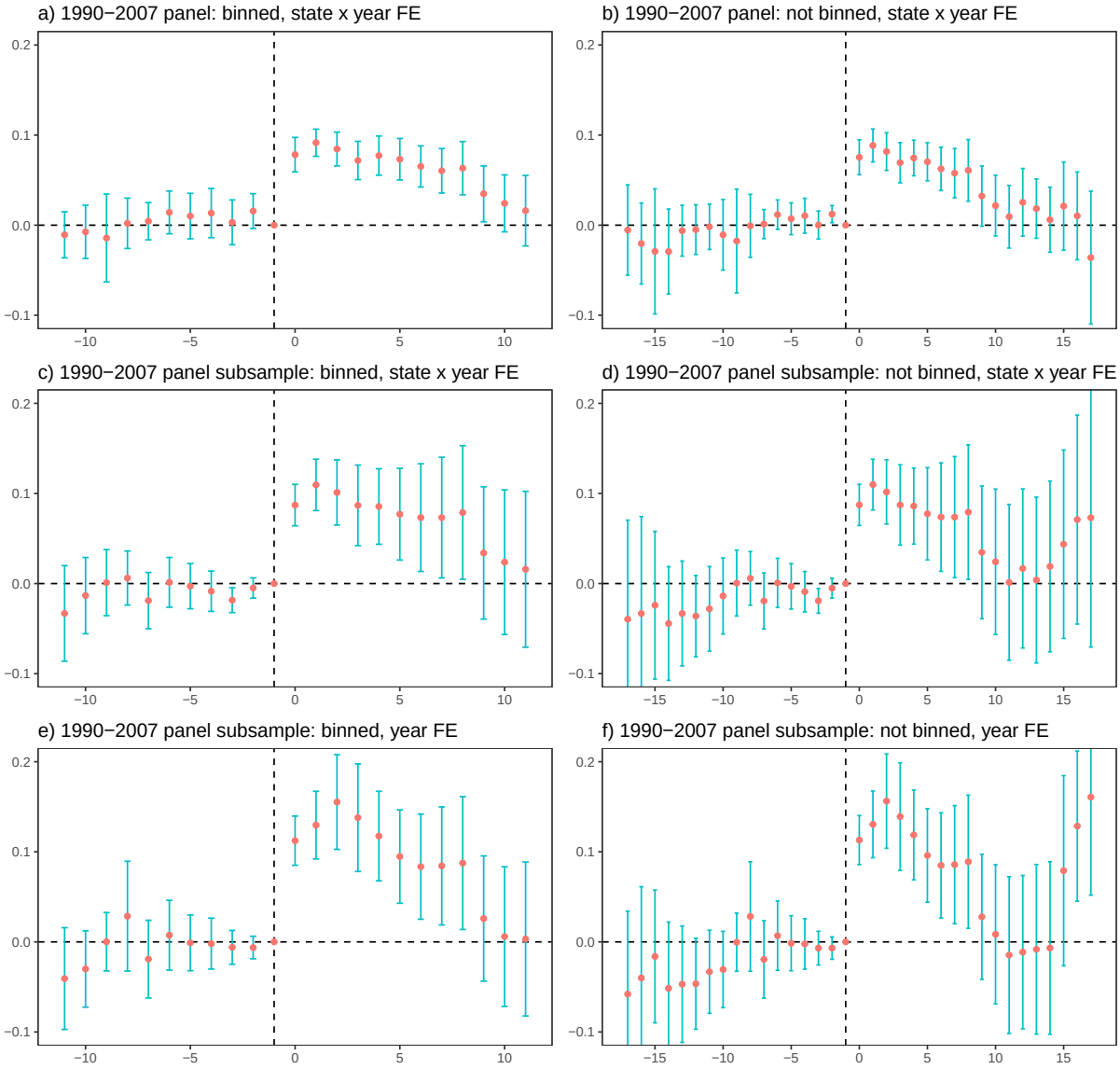
Notes: All subfigures depict the event-time and group dependent treatment effect pattern d). In each case, the indicators for event time -1 and -9 are omitted from the dynamic TWFE specification. Treatment effect heterogeneity increases from top to bottom and relative size of the never-treated group increases from left to right. The dashed vertical line marks event-time -1.

Figure 5: Common Trends of GATEs in de Chaisemartin and D'Haultfœuille (2020)



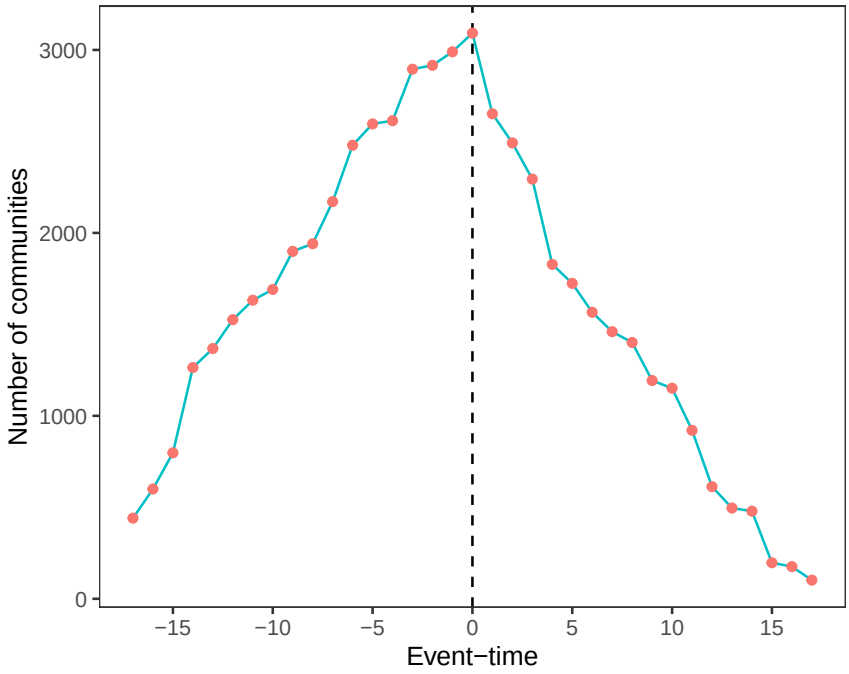
Notes: The figure depicts the common trends of GATEs assumption, underlying the approach of de Chaisemartin and D'Haultfœuille (2020) in a setting with two groups and two periods. The time specific GATE component in period $t = 1$ is 0 in this example.

Figure 6: Replication of Gallagher (2014)



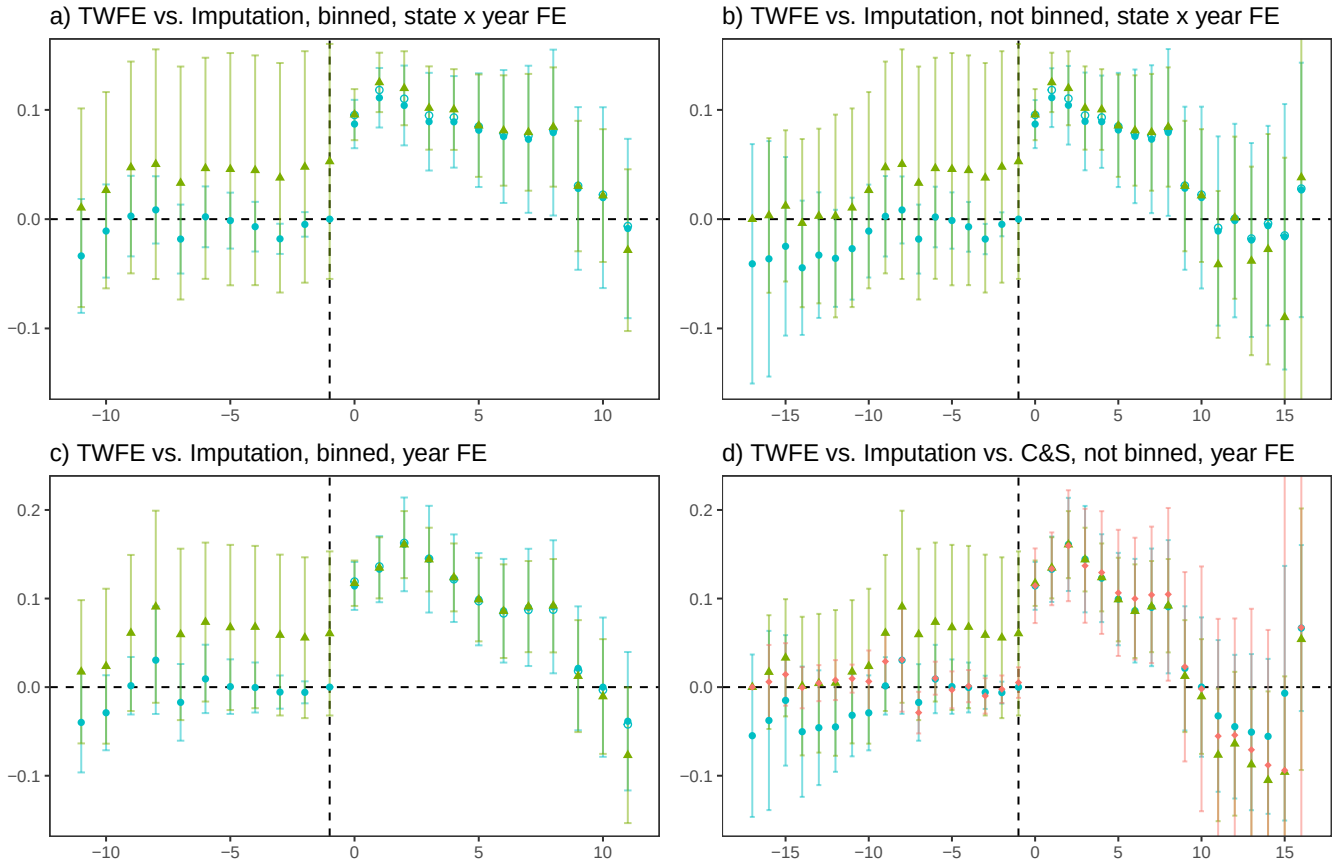
Notes: All figures display coefficient estimates of dynamic TWFE specifications applied to the data from Gallagher (2014), with different sample sizes, sets of fixed effects, and with or without binning of distant event-times. The vertical axis represents log flood policies per person and the horizontal axis represents the event-time. Each red point depicts an event-time point estimate from a dynamic TWFE specification. 95% confidence intervals are outlined in blue. Standard errors are clustered by state. Each specification includes community (/ unit) indicators. All figures in the first column display coefficient estimates obtained from equation (18). The second column contains coefficient estimates from the same specifications but without binning of event-times. In the first row, both specifications are estimated on the original dataset from Gallagher (2014), where communities can be hit by floods multiple times. The second row contains results for the sample restricted to communities treated at most once, with state by year indicators. The last row corresponds to the same subsample but with year indicators instead.

Figure 7: Number of Units in each Event-time



Notes: The figure is based on the subsample of the 1990-2007 panel restricted to all units treated at most once in this time period.

Figure 8: Comparison of Alternative Estimators to the TWFE Estimator Using Data from Gallagher (2014)



Notes: All figures display coefficient estimates together with 95% confidence intervals of different estimators applied to the data from Gallagher (2014) in different settings. Blue corresponds to TWFE estimates (solid = with leads, non-solid = without leads), green corresponds to estimates from the imputation estimator by Borusyak et al. (2021), and red corresponds to estimates from the estimator by Callaway and Sant'Anna (2020). Confidence Intervals of the 'lags only' TWFE coefficients are not depicted. For the sake of clarity, the 'lags only' specification is not depicted in figure d). Each figure only displays the estimators which are applicable in the respective setting. The vertical axis represents log flood policies per person and the horizontal axis represents the event-time. The sample is identical in all four figures and consists of 6,916 communities. Standard errors are clustered by state.

Table 1: Comparison of Alternative Estimators

Estimator	dC & DH (2020)	B, J & S (2021)	Gardner (2021)	S & A (2021)	C & SA (2020)
Setting	Binary treatment Arbitrary adoption	Binary treatment Staggered adoption Arbitrary adoption under additional assumptions	Binary treatment Staggered adoption Arbitrary adoption under additional assumptions	Binary treatment Staggered adoption	Binary treatment Staggered adoption
Target estimand	Average treatment effect of switching observations	Flexible	ATT and event-time specific average treatment effects	Event-time specific average treatment effects	Flexible
Assumptions	CTA for all periods No anticipation Stable groups Common trends of GATEs	CTA for all periods Limited anticipation	CTA for all periods Limited anticipation	CTA for relevant periods Limited anticipation	CCTA or CTA for relevant periods Limited anticipation
Robustness to	Heterogeneity w.r.t. calendar time and group but not both simultaneously	All kinds of heterogeneity	All kinds of heterogeneity	All kinds of heterogeneity	All kinds of heterogeneity
Inference	Uses only subset of observations Pointwise inference	Exploits all observations Pointwise inference	Pointwise inference	Uses only subset of observations Pointwise inference	Uses only subset of observations Simultaneous inference
Software	R: `DIDmultiplegt` Stata: `did_multiplegt`*	R: `didimputation` `did_imputation`*	R & Stata: `did2s`	Stata: `eventstudyinteract`	R: `did`* Stata: `csdid`

Notes: * Recommended implementation.