Two-way fixed effects estimators with heterogeneous treatment effects*

Clément de Chaisemartin† Xavier D'Haultfœuille‡

April 26, 2018

Abstract

Around 20% of all empirical articles published by the American Economic Review between 2010 and 2012 estimate treatment effects using linear regressions with time and group fixed effects. In a model where the effect of the treatment is constant across groups and over time, such two-way fixed effects regressions identify the treatment effect of interest under the standard "common trends" assumption. But these regressions have not been analyzed yet allowing for treatment effect heterogeneity. We study two commonly used two-way fixed effects regressions. We start by showing that without assuming constant treatment effects, those regressions identify weighted sums of the average treatment effects in each group and period, where some weights may be negative. The weights can be estimated, and can help researchers assess whether their results are robust to heterogeneous treatment effects across groups and periods. When some weights are negative, their estimates may not even have the same sign as the true average treatment effect if treatment effects are heterogenous. We then propose another estimator that does not rely on any treatment effect homogeneity assumption. We revisit two empirical articles that have estimated two-way fixed effects regressions. In both cases, around half of the weights attached to those regressions are negative. In one application, our new estimator is of a different sign than the first two-way fixed effects estimator we study, and significantly different from the second one.

*We are very grateful to Guido Imbens, Kyle Meng, Jesse Shapiro, Doug Steigerwald, and seminar participants at Bergen, Stanford, and UCSB for their helpful comments.

†University of California at Santa Barbara, clementdechaisemartin@ucsb.edu

‡CREST, xavier.dhaultfoeuille@ensae.fr

1 Introduction

A popular method to identify treatment effects in the absence of experimental data is to compare groups experiencing different evolutions of their exposure to the treatment over time. Under the assumption that their outcome would have followed the same evolution if their exposure had also followed the same evolution, any difference we observe in the evolution of the outcome between those groups should be due to the treatment effect. In practice, this idea is implemented by estimating linear regressions controlling for both group and time fixed effects, which we hereafter refer to as two-way fixed effects regressions. We conducted a literature review, and found that 19.6% of all empirical articles published by the American Economic Review between 2010 and 2012 use two-way fixed effects regressions. When the effect of the treatment is constant across groups and over time, such regressions identify the treatment effect of interest under the standard "common trends" assumption. However, assuming constant treatment effects across groups and over time is often implausible. For instance, the effect of minimum wage on employment may vary across US counties, and may change over time. Perhaps surprisingly, two-way fixed effects regressions have not been analyzed yet without assuming constant treatment effects. This is the purpose of this paper.

We study two popular two-way fixed effects regressions. In the first one, hereafter referred to as the fixed effects (FE) regression, the outcome is regressed on group and period fixed effects, and on the average value of the treatment in the group $\times$ period a unit belongs to. In the second one, hereafter referred to as the first difference (FD) regression, the first difference of the group-level mean of the outcome is regressed on period fixed effects, and on the first difference of the group-level mean of the treatment.

We start by showing that under the standard common trends assumption, the respective coefficients of the treatment variable in those two regressions identify weighted sums of the average treatment effect on treated units (ATT) in each group and at each period. The weights are not equal to the proportion that each group $\times$ period accounts for in the full population, so those coefficients are not equal to the ATT in the full population. Perhaps more worryingly, many of the weights attached to those coefficients may be negative. In two empirical applications, we find that more than 50% of the weights are negative. When many weights attached to a coefficient are negative, this coefficient is not robust to heterogeneous treatment effects across groups and periods. Even if the treatment effect is positive for all units in the population, this coefficient may be negative. We also show that a simple function of the weights identifies the minimal value of the standard deviation of the treatment effect across groups and periods under which the ATT in the full population may actually have the opposite sign than the FE or FD coefficient. This lower bound can be used as a summary measure of a coefficient's robustness to heterogeneous treatment effects: the larger this lower bound, the more robust the coefficient is.

We then consider two supplementary assumptions. The first one, hereafter referred to as the treatment monotonicity assumption, requires that within each group, the treatment of all units evolve in the same direction between each pair of consecutive periods. The second one, hereafter referred to as the stable treatment effect assumption, requires that in each group, the average treatment effect of units treated in period $t-1$ remain stable between $t-1$ and t. We show that under the common trends, treatment monotonicity, and stable treatment effect assumptions, the coefficients of the treatment variable in the FE and FD regressions identify weighted sums of the local average treatment effect (LATE) of the "switchers" in each group and at each period, where "switchers" are units that experience a change in their treatment between two consecutive periods. In most instances, some of the weights attached to the FE coefficient under those three assumptions are negative. On the other hand, in some special cases, including for instance the staggered adoption design where groups go from being fully untreated to fully treated at heterogeneous dates, the weights attached to the FD coefficient are all positive.[1]

We also propose a new estimand. This estimand identifies an easily interpretable parameter, the LATE of all switchers in the population. It does not rely on any homogeneous treatment effect assumption. Instead, it relies on a conditional common trends assumption, whose plausibility can be assessed by looking at pre-trends, as in standard differences-in-differences (DID). This new estimand can be used when for each pair of consecutive dates, there are groups whose exposure to the treatment does not change between those two dates, a condition that is often met in practice. In one of the two applications we revisit, the corresponding estimator is significantly different from the FD estimator, and of a different sign than the FE estimator.

We derive our decomposition of the FE and FD coefficients as weighted sums of ATTs or LATEs with a binary treatment variable and without covariates in the regression. Nevertheless, we show that this decomposition can still be obtained with a non-binary, ordered treatment, and with covariates in the FE and FD regressions. Interestingly, the weights remain unchanged in these two cases. Moreover, researchers have also estimated two-stage-least-squares (2SLS) versions of the FE and FD regressions we consider, and we show that our main conclusions also apply to those regressions. Therefore, our results are widely applicable in practice.

Our paper is related to two strands of the literature. First, it is related to papers studying the consequences of treatment effect heterogeneity in linear models, including for instance White (1980), Angrist (1998), or Słoczyński (2014). For instance, White (1980) shows that in a model with heterogeneous effects, a regression of an outcome on an exogenous continuous regressor identifies a weighted average of marginal effects. An important difference between those and our paper is that in our setting, treatment effect heterogeneity may entail negative weights.

[1] Stata dofiles estimating the weights attached to any FE or FD regression under the two sets of assumptions we consider can be found on the authors' webpages.

Our paper is also related to the literature studying DID. In particular, our main result generalizes Theorem 1 in de Chaisemartin and D'Haultfœuille (2018). When the data only bears two groups and two periods, the Wald estimand considered therein is equal to the FE and FD coefficients we study in this paper. Our main result may thus be seen as an extension of that theorem to multiple periods and groups.[2] In research independent from ours, Borusyak and Jaravel (2016) study the staggered adoption design. They show that if the common trends assumption holds, and if the treatment effect only depends on the number of periods elapsed since a unit has started receiving the treatment, then the coefficient of treatment in the FE regression identifies a weighted sum of treatment effects, with weights that may be negative. This result can be obtained from our result for the FE regression, by further assuming that the treatment effect only depends on the number of periods since treatment was received.

The paper is organized as follows. Section 2 presents the regressions we consider, our assumptions, and our parameters of interest. Section 3 presents our main results, considering first the special case with two groups and two periods to build intuition. Section 4 extends these results to non-binary treatments, regressions with covariates, and 2SLS regressions. Section 5 presents our alternative estimand. Section 6 discusses inference. Section 7 presents our literature review of the articles published in the AER, and the results of our two empirical applications.

2 Two-way fixed effects regressions, assumptions, and parameters of interest

2.1 Regressions and assumptions

Assume that one is interested in measuring the effect of a treatment D on some outcome Y. We first assume for simplicity that D is binary, but most of our results apply to any ordered treatment, as we show in Subsection 4.1. Then, let $Y(0)$ and $Y(1)$ denote the two potential outcomes of the same unit without and with the treatment. The observed outcome is $Y = Y(D)$. We also assume that the population can be divided into a finite number of time periods represented by a random variable $T \in \{0, 1, ..., \overline{t}\}$, and into a finite number of groups represented by a random variable $G \in \{0, 1, ..., \overline{g}\}$. Each group $\times$ period cell may bear several or a single unit. In the latter case, conditional on G and T all the random variables we consider are degenerate. In the former case, those variables may have a non-degenerate distribution. Some of the assumptions we introduce below have a different interpretation in each of these two cases, as we highlight in the end of this section by reviewing several concrete examples.

[2] In fact, a preliminary version of our main result appeared in a working paper version of de Chaisemartin and D'Haultfœuille (2018) (see Theorems S1 and S2 in de Chaisemartin and D'Haultfoeuille, 2015).

We now introduce some notation we use throughout the paper. For any random variable R and for every $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$, let $R_{g,.}$, $R_{.,t}$, and $R_{g,t}$ respectively be random variables such that $R_{g,.} \sim R|G = g$, $R_{.,t} \sim R|T = t$, and $R_{g,t} \sim R|G = g, T = t$, where $\sim$ denotes equality in distribution. Finally, let $FD_R(g,t) = E(R_{g,t}) - E(R_{g,t-1})$ denote the conditional first-difference operator. This notational shortcut is useful to avoid the notational burden of, e.g., evaluating the function $(g,t) \mapsto E(D_{g,t})$ at $(G, T-1)$.

We study the two following linear regressions coefficients.

Regression 1 *(Fixed-effects regression)*

Let β_{fe} denote the coefficient of $E(D|G,T)$ in an OLS regression of Y on a constant, $(1\{G = g\})_{1\leq g\leq \overline{g}}$, $(1\{T = t\})_{1\leq t\leq \overline{t}}$, and $E(D|G,T)$.

Regression 2 *(First-difference regression)*

Let β_{fd} denote the coefficient of $FD_D(G,T)$ in an OLS regression of $FD_Y(G,T)$ on a constant, $(1\{T = t\})_{2\leq t\leq \overline{t}}$, and $FD_D(G,T)$, conditional on $T \geq 1$.

The FE regression is the OLS regression of the outcome on group and time fixed effects, and on the mean value of the treatment in the group a unit belongs to and at the period she is observed. This regression is very pervasive: 11 articles published in the AER between 2010 and 2012 have estimated it. Other articles have estimated regressions similar to it, e.g. with two treatment variables instead of one. The FD regression is the OLS regression of the group-level first difference of the mean outcome on the group-level first difference of the mean treatment, with time fixed effects. It is also very pervasive: seven articles published in the AER between 2010 and 2012 have estimated it, and other articles have estimated regressions similar to it. When $\overline{t} = 1$ and $T \perp\!\!\!\perp G$, meaning that groups' distribution remains stable over time, one can show using the Frisch-Waugh theorem that $\beta_{fe} = \beta_{fd}$. β_{fe} usually differs from β_{fd} when $\overline{t} > 1$ or T is not independent from G.

We now introduce the main assumptions we consider.

Assumption 1 *(Common trends) For all $t \in \{1,...,\overline{t}\}$, $E(Y(0)|G,T = t) - E(Y(0)|G,T = t-1)$ does not depend on G.*

Assumption 2 *(Treatment monotonicity) There exist random variables $D(0),...,D(\overline{t})$ such that:*

1. *$D = D(T)$;*
2. *For all $t \in \{1,...,\overline{t}\}$, $D(t) \perp\!\!\!\perp T|G$;*
3. *For all $t \in \{1,...,\overline{t}\}$, $P(D(t) \geq D(t-1)|G) = 1$ or $P(D(t) \leq D(t-1)|G) = 1$.*

Assumption 3 *(Stable treatment effect) For all* $(g,t) \in \{0,...,\overline{g}\} \times \{1,...,\overline{t}\}$,
$E(Y(1)-Y(0)|G=g,T=t,D(t-1)=1) = E(Y(1)-Y(0)|G=g,T=t-1,D(t-1)=1)$.

The common trends assumption requires that the mean of $Y(0)$ follow the same evolution over time in the treatment and control groups. This assumption also underlies the standard DID estimand (see, e.g., Abadie, 2005). The treatment monotonicity assumption requires that each unit has $\overline{t}+1$ variables $D(0)$, $D(1)$, ..., $D(\overline{t})$ attached to her, which respectively denote her treatment at T =0, 1, ..., $\overline{t}$. It also requires that the distribution of those variables be stable across periods in each group. Finally, it implies that between each pair of consecutive periods, in a given group there cannot be both units whose treatment increases and units whose treatment decreases. The stable treatment effect assumption requires that in every group, the average treatment effect among units treated in period $t-1$ does not change between $t-1$ and t.

We now review several examples to clarify the restrictions implied by the treatment monotonicity assumption. Some articles have estimated the FE or the FD regression with a treatment that is constant within each group × period cell. For instance, in Gentzkow et al. (2011) the treatment is the number of newspapers in county g and election year t. Then, Point 3 automatically holds: $P(D(t) \geq D(t-1)|G) = 1$ in groups where the treatment increases between $t-1$ and t, and $P(D(t) \leq D(t-1)|G) = 1$ in groups where it decreases. Point 2 will also automatically hold if the regression is estimated at the group × period-level, unless some groups appear or disappear over time. Other articles have estimated the FE or the FD regression with a treatment that varies within each group × period cell. For instance, in Enikolopov et al. (2011), the treatment is whether someone has access to an independent TV channel, groups are Russia's subregions, and periods are election years. Then, Point 3 will fail to hold if there are groups where the treatment of some units diminishes between $t-1$ and t, while the treatment of other units increases. When the FE or the FD regression is estimated with individual-level panel data, the variables $(D(t))_{0\leq t\leq \overline{t}}$ are observed: they are just the treatments of each unit at each period. Then, one can assess from the data whether Point 3 holds or not.[3] On the other hand, when the FE or the FD regression is estimated with individual-level repeated cross-sections, only $D(T)$ is observed so Point 3 is not testable.

2.2 Parameters of interest, and a discussion of treatment effect heterogeneity

Our two parameters of interest are $\Delta^{TR} = E(Y(1)-Y(0)|D=1)$ and $\Delta^{S} = E\left(Y(1)-Y(0)|S\right)$, where $S = \{D(T-1) \neq D(T), T \geq 1\}$ denotes units whose treatment status switches between $T-1$ and T. Δ^{TR} is the Average Treatment effect on the Treated (ATT), and Δ^{S} is the Local Average Treatment Effect (LATE) of the switchers. For any $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$, let

[3]This test may reveal that Point 3 fails. However, our results still hold if the treatment variables satisfy the threshold crossing Equation (3.2) in de Chaisemartin and D'Haultfœuille (2018), which is weaker than Point 3.

$\Delta^{TR}_{g,t} = E(Y(1) - Y(0)|D = 1, G = g, T = t)$ denote the ATT in group g and at period t. For any $(g,t) \in \{0, ..., \overline{g}\} \times \{1, ..., \overline{t}\}$, let $\Delta^{S}_{g,t} = E(Y(1) - Y(0)|S, G = g, T = t)$ denote the LATE of switchers in group g and at period t. Then $\Delta^{TR} = E[\Delta^{TR}_{G,T}|D = 1]$ and $\Delta^{S} = E[\Delta^{S}_{G,T}|S]$: Δ^{TR} and Δ^{S} are respectively equal to weighted averages of the $(\Delta^{TR}_{g,t})_{g,t}$ and $(\Delta^{S}_{g,t})_{g,t}$.

Throughout the paper, we assume that groups experience common trends, but that the effect of the treatment may be heterogeneous between groups and / or over time. We now discuss two examples where this may happen.

First, assume one wants to learn the effect of the minimum wage on the employment levels of some US counties. For simplicity, let us assume that the minimum wage can only take two values, a low and a high value. Also, let us assume that there are only two periods, the 90s and the 2000s. Between these two periods, the amount of competition from China for the US industry increased substantially. Thus, for the common trends assumption to hold for counties A and B, the effect of that increase in competition should be the same in those two counties, in the counterfactual state of the world where A and B have a low minimum wage at both dates. For that to be true, the economy of those two counties should be pretty similar. For instance, if A has a very service-oriented economy, while B has a very industry-oriented economy, it is unlikely that their employment levels will react similarly to Chinese competition.

Now, if the economies of A and B are similar, they should also have similar effects of the minimum wage on employment, thus implying that the treatment effect is homogenous between groups. On the other hand, the treatment effect may vary over time. For instance, the drop in the employment levels of A and B due to Chinese competition will probably be higher if their minimum wage is high than if their minimum wage is low. This is equivalent to saying that the effect of the minimum wage on employment diminishes from the first to the second period: due to Chinese competition in the second period, the minimum wage may have a more negative effect on employment then.

Second, assume one wants to learn the effect of a job training program implemented in some US counties on participants' wages. Let us suppose that individuals self-select into the training according to a Roy model:

$$D = 1\{Y(1) - Y(0) > c_{G,T}\}, \tag{1}$$

where $c_{g,t}$ represents the cost of the training for individuals in county g and period t. Here, the common trends condition requires that average wages without the training follow the same evolution in all counties. As above, for this to hold counties used in the analysis should have similar economies, so let us assume that those counties are actually identical copies of each other: at each period, their distribution of wages without and with the training is the same. Therefore, $(g,t) \mapsto E(Y(1) - Y(0)|G = g, T = t)$ is constant. However, $c_{g,t}$ may vary across counties and over time: some counties may subsidize the training more than others, and some counties may change their subsidies over time. Then, $(g,t) \mapsto \Delta^{TR}_{g,t} = E(Y(1)-Y(0)|Y(1)-Y(0) > c_{g,t})$ will not be constant, despite the fact that all counties in the sample have similar economies and experience similar trends on their wages. Similarly, $(g,t) \mapsto \Delta^{S}_{g,t}$ will also not be constant.

Overall, when the treatment is assigned at the group $\times$ period level as in the minimum wage example, the economic restrictions underlying the common trends assumption may also imply homogeneous treatment effect between groups. However, those restrictions usually do not imply that the treatment effect is constant over time. Moreover, when the treatment is assigned at the individual level, as in the job training example, the economic restrictions underlying the common trends assumption neither imply homogeneous treatment effects between groups, nor homogeneous treatment effects over time.

3 Main results

3.1 The special case with two groups and periods

To introduce our main result, let us start by considering the special case where the population only bears two groups and two periods, which we studied in de Chaisemartin and D'Haultfœuille (2018). In that case, one has

$$\beta_{fe} = \beta_{fd} = \frac{E(Y_{1,1}) - E(Y_{1,0}) - E(Y_{0,1}) + E(Y_{0,0})}{E(D_{1,1}) - E(D_{1,0}) - E(D_{0,1}) + E(D_{0,0})}, \tag{2}$$

where the right-hand side of the previous display is the DID comparing the evolution of the mean outcome from period 0 to 1 in groups 0 and 1, divided by the DID comparing the evolution of the mean treatment in those two groups. Let DID_D denote the denominator of this ratio.

Let us first assume that $D = G \times T$: only units in group 1 and period 1 receive the treatment, a case often referred to as a "sharp" DID. Then, $\beta_{fe} = \beta_{fd} = E(Y_{1,1}) - E(Y_{1,0}) - E(Y_{0,1}) + E(Y_{0,0})$, and one can show that

$$\beta_{fe} = \beta_{fd} = \Delta^{TR}_{1,1}.$$

The common trends assumption requires that if all units had remained untreated, the mean outcome would have followed parallel trends in groups 0 and 1. In a sharp DID, the only departure from the scenario where nobody is treated is that units in group 1 and period 1 receive the treatment. Therefore, any discrepancy between the trends of the mean outcome in the two groups must come from the effect of the treatment in group 1 and period 1, so β_{fe} and β_{fd} identify the ATT in group 1 and period 1.

Now, let us assume that $D \neq G \times T$: there may be treated units in each of the four group $\times$ period cells, a case that is often referred to as a "fuzzy" DID. Then, one can show, using (2) and Point 1 of Lemma 2 in the appendix, that under the common trends assumption, β_{fe} and β_{fd} identify a weighted sum of the four ATTs in each group and time periods, where two ATTs enter with negative weights:

$$\beta_{fe} = \beta_{fd} = \frac{E(D_{1,1})}{DID_D}\Delta^{TR}_{1,1} - \frac{E(D_{1,0})}{DID_D}\Delta^{TR}_{1,0} - \frac{E(D_{0,1})}{DID_D}\Delta^{TR}_{0,1} + \frac{E(D_{0,0})}{DID_D}\Delta^{TR}_{0,0}. \tag{3}$$

Intuitively, in a fuzzy DID there are potentially four departures from the scenario where nobody is treated: some units may be treated in group 1 and period 1, in group 1 and period 0, etc. Therefore, the discrepancy between the trends of the mean outcome in the two groups can come from the treatment effect in any group and time period.

If we further impose the treatment monotonicity and stable treatment effect assumptions, one can show, using (2) and Point 2 of Lemma 2, that β_{fe} and β_{fd} identify a weighted sum of switchers' LATEs in groups 0 and 1:

$$\beta_{fe} = \beta_{fd} = \frac{E(D_{1,1}) - E(D_{1,0})}{DID_D}\Delta^{S}_{1,1} - \frac{E(D_{0,1}) - E(D_{0,0})}{DID_D}\Delta^{S}_{0,1}. \tag{4}$$

Let us give the intuition underlying Equation (4) in the case where in both groups the treatment rate increases from period 0 to 1. Then, under the treatment monotonicity assumption, for $g \in \{0, 1\}$ $\Delta^{TR}_{g,1}$ can be decomposed into the period 1 ATE of units already treated in period 0, and the period 1 ATE of units switching into treatment. Under the stable treatment effect assumption, the period 1 ATE of units already treated in period 0 is equal to their period 0 ATE, $\Delta^{TR}_{g,0}$. Thus, those terms cancel out in Equation (3), and we are left with a weighted sum of switchers' LATEs in both groups. When the share of treated units increases in one group but decreases in the other, then both LATEs in Equation (4) enter with a positive weight. In this case, one can actually show that β_{fe} and β_{fd} identify the LATE of all switchers. On the other hand, when the share of treated units increases or decreases in both groups, then one of the two LATEs in Equation (4) enters with a negative weight.

3.2 Two-way fixed effects estimators as weighted sums of average treatment effects

Our main result extends Equations (3) and (4) to the case where the population bears more than two groups and periods. Then, we show that under the common trends assumption (resp. the common trends, treatment monotonicity, and stable treatment effect assumptions), β_{fe} and β_{fd} still identify weighted sums of $(\Delta^{TR}_{g,t})_{g,t}$ (resp. $(\Delta^{S}_{g,t})_{g,t}$), with weights that can be negative, as in the two-groups and two-periods case.

We start by defining the weights attached to $(\Delta^{TR}_{g,t})_{g,t}$ in the FE and FD regressions under the common trends assumption. First, for any $(g, t) \in \{0, ..., \overline{g}\} \times \{0, ..., \overline{t}\}$, let $\varepsilon_{fe,g,t}$ denote the residual of observations in group g and at period t in the regression of $E(D|G,T)$ on a constant, $(1\{G=g\})_{1\leq g\leq \overline{g}}$, and $(1\{T=t\})_{1\leq t\leq \overline{t}}$. Then, provided that $E[\varepsilon_{fe,G,T}E(D|G,T)]\neq 0$, let

$$w_{fe,g,t}=\frac{\varepsilon_{fe,g,t}E(D)}{E[\varepsilon_{fe,G,T}E(D|G,T)]},\quad W_{fe}=w_{fe,G,T}.$$

As in this section we assume that the treatment is binary, we have $W_{fe}=\varepsilon_{fe,G,T}/E(\varepsilon_{fe,G,T}|D=1)$. However, this simplification does not carry through to the case where the treatment is not binary, so we stick to the definition of W_{fe} in the previous display to ensure it applies both to the binary- and non-binary-treatment cases.

Second, for any $(g,t)\in\{0,...,\overline{g}\}\times\{1,...,\overline{t}\}$, let $\varepsilon_{fd,g,t}$ denote the residual of observations in group g and at period t in the regression of $FD_D(G,T)$ on a constant and $(1\{T=t\})_{2\leq t\leq \overline{t}}$, conditional on $T\geq 1$. For any $g\in\{0,...,\overline{g}\}$, let also $\varepsilon_{fd,g,0}=0$, $\varepsilon_{fd,g,\overline{t}+1}=0$. Then, for any $(g,t)\in\{0,...,\overline{g}\}\times\{0,...,\overline{t}\}$, let

$$\widetilde{w}_{fd,g,t}=\varepsilon_{fd,g,t}-\varepsilon_{fd,g,t+1}\frac{P(G=g,T=t+1)}{P(G=g,T=t)}.$$

Finally, provided that $E[\widetilde{w}_{fd,G,T}E(D|G,T)]\neq 0$, let

$$w_{fd,g,t}=\frac{\widetilde{w}_{fd,g,t}E(D)}{E[\widetilde{w}_{fd,G,T}E(D|G,T)]},\quad W_{fd}=w_{fd,G,T}.$$

Note that for every $k\in\{fe,fd\}$, $E(W_k|D=1)=1$: conditional on $D=1$, W_k is a weighting variable with an expectation equal to 1.

We then define the weights attached to $(\Delta^S_{g,t})_{g,t}$ in the two regressions under the common trends, treatment monotonicity, and stable treatment effect assumptions. First, for all $(g,t)\in\{0,...,\overline{g}\}\times\{1,...,\overline{t}\}$, let

$$s_{g,t}=\text{sgn}\left(E(D_{g,t})-E(D_{g,t-1})\right),$$

where for any real number x, $\text{sgn}(x)=1\{x>0\}-1\{x<0\}$. $s_{g,t}$ is equal to 1 (resp. -1, 0) for groups where the share of treated units increases (resp. decreases, does not change) between $t-1$ and t. Let also, for all $(g,t)\in\{0,...,\overline{g}\}\times\{1,...,\overline{t}\}$,

$$\widetilde{\omega}_{fe,g,t}=\frac{s_{g,t}E[\varepsilon_{fe,G,T}1\{G=g,T\geq t\}]}{P(G=g,T=t)}.$$

Finally, provided that $E[\widetilde{\omega}_{fe,G,T}|S]\neq 0$, let

$$\omega_{fe,g,t}=\frac{\widetilde{\omega}_{fe,g,t}}{E[\widetilde{\omega}_{fe,G,T}|S]},\quad \Omega_{fe}=\omega_{fe,G,T}.$$

Second, provided that $E[s_{G,T}\varepsilon_{k,G,T}|S]\neq 0$, let us define, for all $(g,t)\in\{0,...,\overline{g}\}\times\{1,...,\overline{t}\}$,

$$\omega_{fd,g,t}=\frac{s_{g,t}\varepsilon_{fd,g,t}}{E[s_{G,T}\varepsilon_{fd,G,T}|S]},\quad \Omega_k=\omega_{fd,G,T}.$$

Note that $E(\Omega_k|S) = 1$ for $k \in \{fe, fd\}$,: in the population of switchers, Ω_k is a weighting variable with an expectation equal to 1.

Importantly, $(w_{k,g,t})_{g,t}$ and $(\omega_{k,g,t})_{g,t}$ are identified and can easily be estimated. This is obvious for $(w_{k,g,t})_{g,t}$, but less so for $(\omega_{k,g,t})_{g,t}$, as their denominators involve expectations over switchers, a population that is generally not identified. Nonetheless, we show in the appendix that under the treatment monotonicity assumption, for any function f, $E(f(G,T)|S)$ is identified by

$$E(f(G,T)|S) = \frac{\sum_{(g,t):t\geq 1} P(G=g,T=t)\,|E(D_{g,t}) - E(D_{g,t-1})|\,f(g,t)}{\sum_{(g,t):t\geq 1} P(G=g,T=t)\,|E(D_{g,t}) - E(D_{g,t-1})|}. \quad (5)$$

We can now state our main result. We say below that β_k (resp. W_k, Ω_k) is well-defined whenever there is a unique solution to the linear system corresponding to the regression attached to β_k (resp. when the denominator in their definition is not zero).

Theorem 1 *Assume that D is binary, $k \in \{fe, fd\}$ and β_k is well defined.*

1. *If Assumption 1 holds, then W_k is well defined and*

$$\beta_k = E\left[W_k \Delta^{TR}_{G,T} | D = 1\right].$$

2. *If Assumptions 1-3 hold, then Ω_k is well defined and*

$$\beta_k = E\left[\Omega_k \Delta^{S}_{G,T} | S\right].$$

In Theorem 1, we show that under Assumption 1 (resp. Assumptions 1-3), β_{fe} and β_{fd} identify weighted sums of the ATTs (resp. of switchers' LATEs) in each group and at each period. Indeed, using the law of iterated expectations, the displayed equations in the first and second statements of the theorem are respectively equivalent to

$$\beta_k = \sum_{g=0}^{\bar{g}} \sum_{t=0}^{\bar{t}} P(G=g,T=t|D=1) w_{k,g,t} \Delta^{TR}_{g,t} \quad (6)$$

$$\beta_k = \sum_{g=0}^{\bar{g}} \sum_{t=1}^{\bar{t}} P(G=g,T=t|S) \omega_{k,g,t} \Delta^{S}_{g,t}. \quad (7)$$

The weights $(P(G=g,T=t|D=1)w_{k,g,t})_{g,t}$ (resp. $(P(G=g,T=t|S)\omega_{k,g,t})_{g,t}$) differ from $(P(G=g,T=t|D=1))_{g,t}$ (resp. $(P(G=g,T=t|S))_{g,t}$), so under Assumption 1 (resp. Assumptions 1-3), β_k does not identify the ATT in the full population (resp. the LATE of all switchers).

Perhaps more worryingly, in most instances some of the weights $(P(G=g,T=t|D=1)w_{k,g,t})_{g,t}$ will be negative. Then, β_k does not satisfy the no-sign-reversal property under the common trends assumption: β_k may for instance be negative while the treatment effect is positive for everybody in the population. First, note that $w_{fe,g,t} = \varepsilon_{fe,g,t} E(D)/V(\varepsilon_{fe,G,T})$. As $E(\varepsilon_{fe,G,T}) = 0$, there must be some values of (g,t) for which $\varepsilon_{fe,g,t} < 0$, otherwise we would have $\varepsilon_{fe,g,t} = 0$ for all (g,t) and β_{fe} would not be well defined. Therefore, some of the $(w_{fe,g,t})_{g,t}$ must be negative. Accordingly, the only instance where all the weights $(P(G = g, T = t|D = 1)w_{fe,g,t})_{g,t}$ are positive is when all the (g,t) such that $\varepsilon_{fe,g,t} < 0$, meaning that $E(D_{g,t})$ is lower than its predicted value in a linear regression with group and time fixed effects, are also such that $E(D_{g,t}) = 0$, thus implying that $P(G = g, T = t|D = 1) = 0$. This condition is for instance satisfied in sharp DIDs with two groups and two periods and where $E(D_{g,t}) = 1\{g = 1, t = 1\}$. But it is unlikely to hold more generally. For instance, when $E(D_{g,t}) > 0$ for all (g,t), some of the weights $(P(G = g, T = t|D = 1)w_{fe,g,t})_{g,t}$ must be negative. Similarly, one can show, using $E(\tilde{w}_{2,G,T}) = 0$, that some of the $(w_{fd,g,t})_{g,t}$ must be negative. Here as well, the only instance where all the weights $(P(G = g, T = t|D = 1)w_{fd,g,t})_{g,t}$ are positive is when all the (g,t) such that $w_{fd,g,t} < 0$ are also such that $E(D_{g,t}) = 0$.

In some instances, the weights $(P(G = g, T = t|S)\omega_{k,g,t})_{g,t}$ will all be positive, but in other instances some of those weights will be negative. First, one can show that for any $(g,t) \in \{0, ..., \overline{g}\} \times \{1, ..., \overline{t}\}$, $\varepsilon_{fd,g,t} = E(D_{g,t}) - E(D_{g,t-1}) - (E(D_{.,t}) - E(D_{.,t-1})$. Moreover, under Assumptions 1-3, β_{fd} well defined implies $E(s_{G,T}\varepsilon_{fd,G,T}|S) > 0$. Therefore, the $(P(G = g, T = t|S)\omega_{fd,g,t})_{g,t}$ are all positive if and only if for all $t \geq 1$, in all the groups where the share of treated units strictly increases (resp. strictly decreases) between $t-1$ and t, that increase (resp. decrease) is higher (resp. lower) than $E(D_{.,t}) - E(D_{.,t-1})$, the average evolution of the treatment rate across all groups between $t-1$ and t. For instance, if for some t $E(D_{g,t}) - E(D_{g,t-1})$ is strictly positive (resp. strictly negative) for all g, but not constant across g, then some of the $(P(G = g, T = t|S)\omega_{fd,g,t})_{g,t}$ are negative. Finally, it is more difficult to derive a simple necessary and sufficient condition under which the $(P(G = g, T = t|S)\omega_{fe,g,t})_{g,t}$ are all positive. However, in Subsection 3.3, we derive such a condition in two special cases.

Theorem 1 shows that under Assumption 1 (resp. Assumptions 1-3) alone, β_{fe} and β_{fd} do not identify Δ^{TR} (resp. Δ^S), and may not even satisfy the no-sign reversal property. We now provide conditions under which they do identify Δ^{TR} (resp. Δ^S).

Assumption 4_k *(Random weights or homogeneous ATTs)* $cov\left(W_k, \Delta^{TR}_{G,T}|D = 1\right) = 0$.

Assumption 5_k *(Random weights or homogeneous LATEs)* $cov\left(\Omega_k, \Delta^{S}_{G,T}|S\right) = 0$.

Assumptions 4_k and 5_k are indexed by $k \in \{fe, fd\}$, because the assumption one needs to invoke for identification depends on whether one considers β_{fe} or β_{fd}.

Corollary 1 *Assume that D is binary, $k \in \{fe, fd\}$ and β_k is well defined.*

1. *If Assumptions 1 and 4_k hold,* $\beta_k = \Delta^{TR}$.

2. *If Assumptions 1-3 and 5_k hold,* $\beta_k = \Delta^S$.

This result follows directly from Theorem 1 and $E(W_k|D=1) = E(\Omega_k|S) = 1$. Specifically, under Assumption 1 (resp. under Assumptions 1-3), we have $\beta_k - \Delta^{TR} = \text{cov}(W_k, \Delta^{TR}_{G,T}|D=1)$ (resp. $\beta_k - \Delta^S = \text{cov}(\Omega_k, \Delta^S_{G,T}|D=1)$). Therefore, if one is ready to further impose Assumption 4_k (resp. Assumption 5_k), β_k identifies Δ^{TR} (resp. Δ^S).

Assumptions 4_{fe} and 4_{fd} (resp. 5_{fe} and 5_{fd}) hold if $\Delta^{TR}_{G,T}$ (resp. $\Delta^S_{G,T}$) is constant, or if W_k (resp. Ω_k) is not systematically correlated to $\Delta^{TR}_{G,T}$ (resp. $\Delta^S_{G,T}$). We now discuss the plausibility of those two conditions.

In empirical applications, it is often implausible that the treatment effect is constant across groups and time periods. In Section 7, we review two applications where this assumption may fail. Moreover, this assumption has strong testable implications. If Assumption 1 holds (resp. Assumptions 1-3 hold) and if there is a real number Δ such that $\Delta^{TR}_{G,T} = \Delta$ (resp. $\Delta^S_{G,T} = \Delta$), then for all g and for all $t \geq 1$,

$$E(Y_{g,t}) - E(Y_{g,t-1}) = \alpha_t + \left(E(D_{g,t}) - E(D_{g,t-1})\right)\Delta, \tag{8}$$

where α_t is the evolution of the mean of $Y(0)$ from $t-1$ to t, which is assumed to be the same in all groups. Thus, the $\bar{t}+1$ parameters $((\alpha_t)_{1\leq t\leq \bar{t}}, \Delta)$ satisfy the following overidentified system of $(\bar{g}+1)\times\bar{t}$ moment conditions: for all g and all $t \geq 1$,

$$\begin{aligned} E\Bigg[&\frac{Y1\{G=g,T=t\}}{P(G=g,T=t)} - \frac{Y1\{G=g,T=t-1\}}{P(G=g,T=t-1)} - \alpha_t \\ &-\Delta\left(\frac{D1\{G=g,T=t\}}{P(G=g,T=t)} - \frac{D1\{G=g,T=t-1\}}{P(G=g,T=t-1)}\right)\Bigg] = 0. \end{aligned}$$

To implement the corresponding overidentification test, one needs to estimate the variance of the moment conditions. When only data aggregated at the group $\times$ period level is available, this may not be feasible. In such instances, one can still test Assumption 1 (resp. Assumptions 1-3) and the constant ATTs (resp. LATEs) assumption, for instance by comparing the estimate of β_{fd} or β_{fe} in a regression weighted by the population of each group $\times$ period, and in an unweighted regression. It follows from Theorem 1 that under Assumption 1 (resp. Assumptions 1-3), the weighted and unweighted estimands identify weighted sums of ATTs (resp. LATEs), with different weights. But if $\Delta^{TR}_{G,T} = \Delta$ (resp. $\Delta^S_{G,T} = \Delta$), the weights do not matter so those two estimands identify Δ. Therefore, the corresponding estimates should not significantly differ.

When the treatment effect is not constant, Assumptions 4_{fe} and 4_{fd} (resp. Assumptions 5_{fe} and 5_{fd}) can still hold if W_k (resp. Ω_k) is not systematically correlated to $\Delta^{TR}_{G,T}$ (resp. $\Delta^S_{G,T}$). To simplify the discussion of that condition, let us momentarily assume that $\bar{t} = 1$ and $G \perp\!\!\!\perp T$. Then $\beta_{fe} = \beta_{fd}$, so Assumptions 4_{fe} and 4_{fd} are equivalent, and Assumptions 5_{fe} and 5_{fd} are also equivalent. First, Assumptions 4_{fe} and 4_{fd} hold if $\text{cov}\left(\varepsilon_{fe,G,T}, \Delta^{TR}_{G,T}|D=1\right) = 0$. This is restrictive. Positive (resp. negative) values of $\varepsilon_{fe,g,t}$ correspond to values of (g,t) for which the share of treated units is larger (resp. lower) than predicted by the regression of $E(D_{g,t})$ on group and period fixed effects. Those may also be the values of (g,t) with the largest treatment effect. For instance, if selection into treatment is determined by the Roy model in Equation (1) above, and if $Y(d)|G=g, T=t \sim \mathcal{N}(m_d + a_t, 1)$, then one can show that $\varepsilon_{fe,g,t}$ and $\Delta^{TR}_{g,t}$ are positively related when $G \perp\!\!\!\perp T$.[4] Second, Assumptions 5_{fe} and 5_{fd} hold if $\text{cov}\left(s_{G,1}\varepsilon_{fe,G,1}, \Delta^{S}_{G,1}|S\right) = 0$. This will hold if $P(s_{G,1} = 1|S) = P(s_{G,1} = -1|S)$ and $s_{G,1} \perp\!\!\!\perp (\varepsilon_{fe,G,1}, \Delta^S_{G,1})|S$.[5] However, $P(s_{G,1} = 1|S) = P(s_{G,1} = -1|S)$ implies $P(D=1|T=0) = P(D=1|T=1)$. This latter equality is testable, and it will often be rejected in the data. When it is not, one may still worry that $s_{G,T}$ is not independent of $\Delta^S_{G,T}$: groups where the share of treated units decreases may also be those where switchers experience a decrease of their treatment effect. This may result in a positive correlation between $s_{G,1}$ and $\Delta^S_{G,1}$.

Finally, note that Assumptions 4_k and 5_k are partly testable. First, if β_{fe} and β_{fd} are statistically different, under Assumption 1 (resp. Assumptions 1-3), one can reject Assumption 4_k (resp. Assumption 5_k) for at least one k in $\{fe, fd\}$. Second, assume that researchers observe a variable $P_{g,t}$ that is likely to be correlated with the intensity of the treatment effect across groups and time periods. Then, they can run a suggestive test of Assumption 4_k (resp. Assumption 5_k), by testing whether $\text{cov}(W_k, P_{G,T}|D=1) = 0$ (resp. $\text{cov}(\Omega_k, P_{G,T}|S) = 0$). In Section 7, we conduct this suggestive test in an application, and find that it is rejected.

When the treatment effect is heterogeneous across groups and periods, β_k may be a misleading measure of Δ^{TR} (resp. Δ^S) if Assumption 4_k (resp. 5_k) fails. In the corollary below, we derive the minimum amount of heterogeneity of the $(\Delta^{TR}_{g,t})_{g,t}$ (resp. $(\Delta^S_{g,t})_{g,t}$) that could actually lead β_k to be of a different sign than Δ^{TR} (resp. Δ^S). Let $\sigma^{TR} = V\left(\Delta^{TR}_{G,T}|D=1\right)^{1/2}$ (resp. $\sigma^S = V\left(\Delta^S_{G,T}|S\right)^{1/2}$) denote the standard deviation of the ATTs (resp. of switchers' LATEs) across groups and periods.

Corollary 2 *Assume that D is binary, $k \in \{fe, fd\}$ and β_k is well defined.*

1. *Suppose that Assumption 1 holds and $V(W^k|D=1) > 0$. Then, the minimal value of σ^{TR} compatible with β_k and $\Delta^{TR} = 0$ is*

$$\underline{\sigma}^{TR}_k = \frac{|\beta_k|}{V\left(W^k|D=1\right)^{1/2}}.$$

[4]Specifically, one can show that $\frac{\partial \varepsilon_{fe,g,t}}{\partial c_{g,t}} < 0$ and $\frac{\partial \Delta^{TR}_{g,t}}{\partial c_{g,t}} < 0$. Therefore, $\varepsilon_{fe,G,T}$ and $\Delta^{TR}_{G,T}$ are positively related.

[5]Assumptions 5_{fe} and 5_{fd} can also hold when both of these conditions are violated, but then the violations of those conditions must exactly compensate each other, which is unlikely to be the case.

2. *Suppose that Assumptions 1-3 hold and* $V(\Omega^k|S) > 0$. *Then, the minimal value of* σ^S *compatible with* β_k *and* $\Delta^S = 0$ *is*

$$\underline{\sigma}_k^S = \frac{|\beta_k|}{V\left(\Omega^k|S\right)^{1/2}}.$$

Estimators of $\underline{\sigma}_k^{TR}$ (resp. $\underline{\sigma}_k^S$) can be used to assess the sensitivity of β_k to treatment effect heterogeneity across groups and periods. Assume for instance that β_{fe} is large and positive, while $\underline{\sigma}_{fe}^{TR}$ is close to 0. Then, even under a minor amount of treatment effect heterogeneity, β_{fe} could be of a different sign than Δ^{TR}. Indeed, the data is compatible with β_{fe} large and positive, $\Delta^{TR} = 0$, and a small dispersion of the $(\Delta_{g,t}^{TR})_{g,t}$. To assess whether $\underline{\sigma}_k^{TR}$ and $\underline{\sigma}_k^S$ are small or large, we recommend to divide them by β_k. For instance, a standard deviation of the $(\Delta_{g,t}^{TR})_{g,t}$ equal to, say, 20% of β_k, can arguably be considered as small, as under Assumption 4_k, β_k is equal to the average of $(\Delta_{g,t}^{TR})_{g,t}$ conditional on $D = 1$.

The first (resp. second) point of Corollary 2 does not apply when $V(W^k|D = 1) = 0$ (resp. $V(\Omega^k|S) = 0$). In such a case, $\beta_k = \Delta^{TR}$ (resp. $\beta_k = \Delta^S$) under Assumption 1 (resp. Assumptions 1-3) alone, even if the $(\Delta_{g,t}^{TR})_{g,t}$ (resp. $(\Delta_{g,t}^S)_{g,t}$) are heterogeneous. However, the next proposition shows that it is rarely the case that $V(W^k|D = 1) = 0$ (resp. $V(\Omega^k|S) = 0$).

Proposition 1 *Suppose that* D *is binary,* $k \in \{fe, fd\}$ *and* β_k *is well-defined. Then:*

1. $V(W_k|D = 1) = 0$ *implies that (i) for all* g*, there exists* t *such that* $E(D_{g,t}) = 0$ *and (ii) for all* t*, there exists* g *such that* $E(D_{g,t}) = 0$.

2. $V(\Omega_k|S) = 0$ *implies that for all* $t \in \{1, ..., \overline{t}\}$, $g \mapsto s_{g,t}$ *is not constant and equal to either -1 or 1.*

The conditions above are necessary but not sufficient for the weights in Theorem 1 to be constant. The first condition is violated whenever there is a group that is at least partly treated at every period, or a period at which all groups are at least partly treated. The second condition is violated whenever there are two consecutive periods between which the mean treatment strictly increases in every group, or strictly decreases in every group.

3.3 The weights in two particular designs

We now consider two pervasive research designs in which one can derive easily interpretable necessary and sufficient conditions to have that all the weights in Theorem 1 are positive. First, we consider the staggered adoption design.

Assumption 6 *(Staggered adoption)* $D = 1\{T \geq a_G\}$ *for some* $a_G \in \{0, ..., \overline{t}, \overline{t} + 1\}$.

Assumption 7 *(Group × period regression with a balanced panel of groups) For every* $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$, $P(G=g,T=t) = \frac{1}{\overline{g}+1}\frac{1}{\overline{t}+1}$.

Assumption 6 is satisfied when each group is either fully untreated at each period, fully treated at each period, or fully untreated until a period $a_g - 1$ and fully treated from period a_g onwards. Groups g with $a_g = \overline{t}+1$ never adopt the treatment. This type of staggered adoption design is often met in practice, see e.g. Athey and Stern (2002). Hereafter, we refer to $g(e) = \text{argmin}_{g\in\{0,...,\overline{g}\}} a(g)$ as the earliest adopting group. Assumption 7 is for instance satisfied when the FE or FD regression is estimated with only one unit per group and period, and when no group appears or disappears over time. It is necessary to obtain some, but not all the results in Proposition 2 below.

Proposition 2 *Suppose that Assumption 6 holds, and that* β_{fe} *and* β_{fd} *are well defined. Then,*

1. *If Assumption 7 also holds,* $P(W_{fe} \geq 0|D=1) = 1$ *and* $P(\Omega_{fe} \geq 0|S) = 1$ *if and only if*
$$\frac{a_{g(e)}}{\overline{t}+1} \geq E(D_{.,\overline{t}}) - E(D).$$
2. *If Assumption 7 also holds,* $P(W_{fd} \geq 0|D=1) = 1$ *if and only if* $a_g \geq \overline{t}$ *for every* $g \in \{0,...,\overline{g}\}$.
3. $P(\Omega_{fd} \geq 0|S) = 1$.

The first point of Proposition 2 shows that in the staggered-adoption research design, the weights $(P(G=g,T=t|D=1)w_{fe,g,t})_{g,t}$ and $(P(G=g,T=t|S)\omega_{fe,g,t})_{g,t}$ are all positive if and only if the proportion of periods during which the earliest adopting group remains untreated is larger than the difference between the proportion of groups treated at $\overline{t}$ and the average of that proportion across all periods. This condition is rarely met. For instance, it fails when at least one group is treated at $T=0$.[6] The second point of Proposition 2 then shows that in this research design, the weights $(P(G=g,T=t|D=1)w_{fd,g,t})_{g,t}$ are all positive if and only if every group either remains untreated or becomes treated at the last period. This condition is also rarely met. Finally, the third point of Proposition 2 shows that in this research design, the weights $(P(G=g,T=t|S)\omega_{fd,g,t})_{g,t}$ are always all positive. Contrary to the first and second point, this result does not rely on Assumption 7.

Overall, in staggered adoption designs, the common trends assumption is not sufficient for β_{fe} and β_{fd} to identify convex combinations of the treatment effect in each group and period. On the other hand, if the common trends and stable treatment effect assumptions hold, β_{fd} identifies a convex combination of the LATEs of switchers in each group and period,[7] contrary to β_{fe}. Still, that convex combination is not equal to the LATE of all switchers.

[6] We then have $a_{g(e)}/(\overline{t}+1) = 0$, while $E(D_{.,\overline{t}}) - E(D) > 0$, otherwise β_{fe} would not be well defined.

[7] Under Assumption 6, Assumption 2 automatically holds.

The second design we consider is the heterogeneous adoption design.

Assumption 8 *(Heterogenous adoption)* $\overline{t} = 1$, *D is binary, and for every* $g \in \{0, ..., \overline{g}\}$, $E(D_{g,0}) = 0$ *and* $E(D_{g,1}) > 0$.

Assumption 8 is satisfied in applications with two time periods, where all groups are fully untreated at $T = 0$, and where all groups become at least partly treated at $T = 1$. This type of heterogeneous adoption design is often met in practice, see, e.g., Enikolopov et al. (2011).

Proposition 3 *Suppose that Assumption 8 holds, and that for every* $k \in \{fe, fd\}$, β_k *is well defined. Then,* $P(W_k < 0|D = 1) > 0$ *and* $P(\Omega_k < 0|S) > 0$.

Proposition 3 shows that in the heterogeneous adoption design, for every $k \in \{fe, fd\}$ some of the weights $(P(G = g, T = t|D = 1)w_{k,g,t})_{g,t}$ and $(P(G = g, T = t|S)\omega_{k,g,t})_{g,t}$ are strictly negative. Then, the causal interpretation of β_k relies on Assumption 4_k or 5_k.

4 Extensions

4.1 Non-binary, ordered treatment

We now consider the case where the treatment takes a finite number of ordered values, $D \in \{0, 1, ..., \overline{d}\}$. To accommodate for this extension, we have to modify the stable treatment effect assumption as follows.

Assumption 3O *(Stable treatment effect for ordered D) For all* $(d, g, t) \in \{1, ..., \overline{d}\} \times \{0, ..., \overline{g}\} \times \{1, ..., \overline{t}\}$, $E(Y(d)-Y(0)|G = g, T = t, D(t-1) = d) = E(Y(d)-Y(0)|G = g, T = t-1, D(t-1) = d)$.

We also need to modify the treatment effect parameters we consider. In lieu of Δ^{TR}, we consider $\text{ACR}^{TR} = E(Y(D) - Y(0))/E(D)$, the average causal response (ACR) on the treated. For any $(g, t) \in \{0, ..., \overline{g}\} \times \{0, ..., \overline{t}\}$, let $\text{ACR}^{TR}_{g,t} = E(Y_{g,t}(D) - Y_{g,t}(0))/E(D_{g,t})$ denote the ACR in group g and at period t. Let us also define the probability measure P^{TR} by $P^{TR}(A) = E(D1_A)/E(D)$ for any measurable set A. This probability measure generalizes the conditional probability $P(.|D = 1)$ for a binary treatment to non-binary treatments. We let E^{TR} and cov^{TR} denote the expectation and covariance operators associated to P^{TR}. Then note that

$$\text{ACR}^{TR} = E^{TR}\left[\text{ACR}^{TR}_{G,T}\right].$$

In lieu of Δ^S, we consider $\text{ACR}^S = E[\text{ACR}^S_{G,T}|S]$, where for all $(g, t) \in \{0, ..., \overline{g}\} \times \{1, ..., \overline{t}\}$,

$$\begin{aligned}\text{ACR}^S_{g,t} =& \sum_{d=1}^{\overline{d}} \frac{P(D_{g,t} \geq d) - P(D_{g,t-1} \geq d)}{E(D_{g,t}) - E(D_{g,t-1})} \Delta^S_{dgt}, \\ \Delta^S_{dgt} =& E\left[Y_{g,t}(d) - Y_{g,t}(d-1)|\max(D(t), D(t-1)) \geq d > \min(D(t), D(t-1))\right].\end{aligned}$$

Δ^S_{dgt} is the average effect of going from $d-1$ to d units of treatment, in period t, and among the units in group g whose treatment switches from strictly below to above d or from above to strictly below d between $T = t-1$ and $T = t$. $\text{ACR}^S_{g,t}$ is a weighted average over d of the Δ^S_{dgt}, where each of the Δ^S_{dgt} is weighted proportionally to the size of the corresponding population of switchers. $\text{ACR}^S_{g,t}$ is similar to the ACR parameter considered in Angrist and Imbens (1995). ACR^S is a weighted average of the $\text{ACR}^S_{g,t}$ across groups and periods.

Finally, we generalize Assumptions 4 and 5 as follows.

Assumption 4O$_k$ *(Random weights or homogeneous ACRs)* $cov^{TR}\left(W_k, ACR^{TR}_{G,T}\right) = 0.$

Assumption 5O$_k$ *(Random weights or homogeneous ACRs)* $cov\left(\Omega_k, ACR^S_{G,T}|S\right) = 0.$

Theorem 2 below generalizes Theorem 1 and Corollary 1 to the case where the treatment takes a finite number of ordered values.

Theorem 2 *Assume that* $D \in \{0, ..., \overline{d}\}$, $k \in \{fe, fd\}$ *and* β_k *is well-defined.*

1. *If Assumption 1 holds, then*
$$\beta_k = E^{TR}\left[W_k ACR^{TR}_{G,T}\right]. \tag{9}$$
If Assumption 4O$_k$ further holds, then $\beta_k = ACR^{TR}$.

2. *If Assumptions 1-2 and 3O hold, then*
$$\beta_k = E\left[\Omega_k ACR^S_{G,T}|S\right]. \tag{10}$$
If Assumption 5O$_k$ further holds, then $\beta_k = ACR^S$.

The first (resp. second) point of Theorem 2 shows that under Assumption 1 (resp. Assumption 1-2 and 3O), β_{fe} and β_{fd} identify weighted sums of the $(\text{ACR}^{TR}_{g,t})_{g,t}$ (resp. $(\text{ACR}^S_{g,t})_{g,t}$), where the weights are the same as those in Theorem 1. For every $k \in \{fe, fd\}$, under Assumption 4O$_k$ (resp. 5O$_k$), β_k identifies ACR^{TR} (resp. ACR^S). Yet, as already explained in the binary treatment case, these assumptions may not be plausible.

Theorem 2 extends to a continuous treatment. In such instances, one can for instance show that under Assumptions 1-2 and an appropriate generalization of Assumption 3, β_{fe} and β_{fd} identify weighted sums of the same weighted averages of the derivative of potential outcomes with respect to treatment in group g and at time t as in Angrist et al. (2000).

Finally, given that the weights in Theorem 2 are the same as in Theorem 1, and because the proof of Corollary 2 does not rely on the nature of the treatment, Corollary 2 directly applies to ordered treatments.

4.2 Including covariates

Often times, researchers also include a vector of covariates X as control variables in their regression. We show in this section that our results can be extended to this case. We start by redefining the two regressions we consider in this context.

Regression 1X *(Fixed-effects regression with covariates)*

Let β_{fe}^X *and* γ_{fe} *denote the coefficients of* $E(D|G,T)$ *and* X *in an OLS regression of* Y *on a constant,* $(1\{G=g\})_{1\leq g\leq \overline{g}}$, $(1\{T=t\})_{1\leq t\leq \overline{t}}$, $E(D|G,T)$, *and* X.

Regression 2X *(First-difference regression with covariates)*

Let β_{fd}^X *and* γ_{fd} *denote the coefficients of* $FD_D(G,T)$ *and* $FD_X(G,T)$ *in an OLS regression of* $FD_Y(G,T)$ *on a constant,* $(1\{T=t\})_{2\leq t\leq \overline{t}}$, $FD_D(G,T)$, *and* $FD_X(G,T)$, *conditional on* $T\geq 1$.

Then, we need to modify the common trends assumption as follows.

Assumption 1X$_k$ *(Common trends for* β_k^X*)* $E(Y(0)-X'\gamma_k|G,T=t)-E(Y(0)-X'\gamma_k|G,T=t-1)$ *does not depend on* G *for all* $t\in\{1,...,\overline{t}\}$.

Assumptions 1X$_{fe}$ and 1X$_{fd}$ are implied by the linear and constant treatment effect models that are often invoked to justify the use of FE and FD regressions with covariates. For instance, the use of Regression 1X is often justified by the following model:

$$Y=\gamma_G+\lambda_T+\theta D+X'\gamma_{fe}+\varepsilon, \quad E(\varepsilon|G,T,D,X)=0. \tag{11}$$

Equation (11) implies Assumption 1X$_{fe}$, but it does not imply Assumption 1. This is why we consider the former common trends assumption instead of the latter in this subsection. Similarly, the linear and constant treatment effect model rationalizing the use of Regression 2X imply Assumption 1X$_{fd}$.

Assumption 1X$_{fe}$ requires that once netted out from the partial linear correlation between the outcome and X, $Y(0)$ satisfies the common trends assumption. It may be more plausible than Assumption 1, for instance if there are group-specific trends affecting the outcome but if those group-specific trends can be captured by a linear model in X. A similar interpretation applies to Assumption 1X$_{fd}$.

Finally, we need to assume that the treatment has no effect on the covariates. Let $X(0)$ and $X(1)$ respectively denote the potential covariates of the same unit without and with the treatment.

Assumption 9 *(No treatment effect on the covariates)* $X(0)=X(1)=X$.

Assumption 9 will for instance hold if X is determined prior to the treatment.

Theorem 3 below generalizes Theorem 1 to the case where there are covariates in the regression.

Theorem 3 *Assume that D is binary, $k \in \{fe, fd\}$ and β_k is well-defined.*

1. *If Assumptions $1X_k$ and 9 hold, then*

$$\beta_k^X = E\left[W_k \Delta_{G,T}^{TR} | D = 1\right].$$

If Assumption 4_k further holds, then $\beta_k^X = \Delta^{TR}$.

2. *If Assumptions $1X_k$, 2, 3, and 9 hold, then*

$$\beta_k^X = E\left[\Omega_k \Delta_{G,T}^{S} | S\right].$$

If Assumption 5_k further holds, then $\beta_k^X = \Delta^S$.

Theorem 3 shows that under a modified version of the common trends assumption accounting for the covariates, β_k^X identifies the same weighted sum of the $(\Delta_{g,t}^{TR})_{g,t}$ as in Theorem 1. If one further imposes Assumptions 2 and 3, β_k^X identifies the same weighted sum of the $(\Delta_{g,t}^{S})_{g,t}$ as in Theorem 1. Therefore, if the regression without covariates has some negative weights and one worries that treatment effect heterogeneity may be systematically related to the weights, the addition of covariates will not alleviate that concern.

4.3 2SLS regressions

Researchers have sometimes estimated 2SLS versions of regressions 1 and 2. Our main conclusions also apply to those regressions. For instance, let β_{fe}^{2SLS} denote the coefficient of D in a 2SLS regression of Y on a constant, $(1\{G = g\})_{1 \leq g \leq \overline{g}}$, $(1\{T = t\})_{1 \leq t \leq \overline{t}}$, and D, using a variable Z constant within each group $\times$ period as the instrument for D. Z typically represents an incentive to get treated administered at the group $\times$ period level. For instance, Baum-Snow and Lutz (2011) study the effect of schools racial dissimilarity on enrollment in the US. Therein, D is a school's dissimilarity index at time t, and Z is an indicator for whether the district where the school is located has been desegrated at time t. In such 2SLS regressions, β_{fe}^{2SLS} is the ratio of the coefficients of Z in the reduced-form regression of Y on group and period fixed effects and Z, and in the first-stage regression of D on group and period fixed effects and Z. Because Z is constant within each group $\times$ period, those regressions are regressions of Y and D on group and period fixed effects and $E(Z|G,T)$. Therefore, it follows from our Theorem 1 that under a common trends assumption with respect to Z instead of D, the reduced-form coefficient of Z identifies a weighted sum of the average effect of Z on Y in each group and period, with potentially many negative weights. Similarly, the first-stage coefficient of Z identifies a weighted sum of the average effect of Z on D in each group and period, with potentially many negative weights. It is only if the average effects of Z on Y and D are constant across groups and periods, or if a version of our random weights assumption adapted to those regressions holds, that the reduced-form and first-stage coefficients of Z respectively identify the average effect of Z on Y and D, thus implying that β_{fe}^{2SLS} identifies the LATE of D on Y among units that comply with the instrument.

5 Alternative estimands

In this section, we show that Δ^S is actually identified without imposing the random weights conditions, under the following testable condition.

Assumption 10 *(Existence of "stable" groups) For all $t \in \{1, ..., \overline{t}\}$, there exists $g \in \{0, ..., \overline{g}\}$ such that $E(D_{g,t}) = E(D_{g,t-1})$.*

The stable groups assumption is often satisfied. For instance, in the staggered adoption design, for all t there must be a group whose exposure to the treatment does not change between $t-1$ and t. Indeed, the converse condition would imply that all the groups adopt the treatment at the same period, but then β_{fe} and β_{fd} would not be well defined. It is also satisfied in Gentzkow et al. (2011): between each pair of successive elections, there are counties where the number of newspapers available does not change.

For every $t \in \{1, ..., \overline{t}\}$, let us introduce the "supergroup" variable $G_t^* = s_{G,t}$. For instance, groups with $G_t^* = 1$ are those where the mean treatment strictly increases between $T = t-1$ and $T = t$. For all $(d, g, g^*, t, t') \in \{0, 1\} \times \{0, ..., \overline{g}\} \times \{-1, 0, 1\} \times \{1, ..., \overline{t}\}^2$, let $r(g|g^*, t', t) = P(G = g|G_{t'}^* = g^*, T = t)$, $r_d(g|t', t) = P(G = g|G_{t'}^* = 0, T = t, D = d)$, and let

$$Q = \frac{r(G|G_{T+1}^*, T+1, T+1)}{r(G|G_{T+1}^*, T+1, T)}, \; Q_d = \frac{r_d(G|T+1, T+1)}{r_d(G|T+1, T)}.$$

The ratio $r(g|g^*, t+1, t+1)/r(g|g^*, t+1, t)$ is the share of group g in the supergroup $G_t^* = g^*$ at $T = t+1$, divided by the share of group g in the supergroup $G_t^* = g^*$ at $T = t$. Therefore, Q is larger (resp. lower) than 1 for units whose group's size has increased faster (resp. more slowly) than their period-$T+1$ super-group's size between $T = t$ and $T = t+1$. Q_d can be interpreted similarly, but conditional on $D = d$. Note that if $G \perp\!\!\!\perp T$, $Q = Q_0 = Q_1 = 1$.

Before introducing our two estimands, let us define, for any random variable R and for all $(d, g, t) \in \{0, 1\} \times \{-1, 1\} \times \{1, ..., \bar{t}\}$, the following quantities:

$$\begin{aligned} DID^*_R(g,t) =& E(R|G^*_t = g, T = t) - E(QR|G^*_t = g, T = t-1) \\ & - (E(R|G^*_t = 0, T = t) - E(QR|G^*_t = 0, T = t-1)), \\ \delta_d =& E(Y|D = d, G^*_t = 0, T = t) - E(Q_d Y|D = d, G^*_t = 0, T = t-1) \\ TCD^*(g,t) =& E(Y|G^*_t = g, T = t) - E(Q(Y + \delta_D)|G^*_t = g, T = t-1). \end{aligned}$$

$DID^*_R(g,t)$ compares the evolution of the mean of R in the supergroups $G^*_t = g$ and $G^*_t = 0$ between $T = t-1$ and $T = t$, after reweighting units in period $t-1$ by Q to ensure that groups' distribution is the same in periods $t-1$ and t. $TCD^*(g,t)$ stands for "time-corrected difference". It compares the mean outcome in the supergroup $G^*_t = g$ at $T = t$ to the mean outcome in the same group at $T = t-1$, after imputing the trends on $Y(0)$ and $Y(1)$ observed in the supergroup $G^*_t = 0$ to the period $T = t-1$ mean outcome in the supergroup $G^*_t = g$. Here as well, the reweighting of units by Q_d and Q in period $t-1$ ensures that groups' distribution is the same in periods $t-1$ and t. Let also

$$w_{g,t} = \frac{g DID^*_D(g,t) P(G^*_t = g, T = t)}{\sum_{t'=1}^{\bar{t}} \sum_{g' \in \{-1,1\}} g' DID^*_D(g',t') P(G^*_t = g', T = t')}.$$

Note that by construction, $w_{g,t} \geq 0$. Our estimands are defined by

$$\begin{aligned} W_{DID} &= \sum_{t=1}^{\bar{t}} \sum_{g \in \{-1,1\}} w_{g,t} \frac{DID^*_Y(g,t)}{DID^*_D(g,t)} \\ W_{TC} &= \sum_{t=1}^{\bar{t}} \sum_{g \in \{-1,1\}} w_{g,t} \frac{TCD^*(g,t)}{DID^*_D(g,t)}, \end{aligned}$$

with the convention that $w_{g,t}/DID^*_D(g,t) = 0$ when $P(G^*_t = g, T = t) = 0$.

Theorem 4 below shows that W_{DID} and W_{TC} identify Δ^S without relying on the random weights conditions. The first estimand W_{DID} still relies on the stable treatment effect assumption, which restricts treatment effect heterogeneity. The second estimand W_{TC} does not rely on any restriction on treatment effect heterogeneity. Instead, it relies on the following conditional common trends condition.

Assumption 1' *(Common trends conditional on $D(t-1)$) For all $(d,t) \in \{0,1\} \times \{1, ..., \bar{t}\}$, $E(Y(d) = d|D(t-1) = d, T = t, G) - E(Y(d)|D(t-1) = d, T = t-1, G)$ does not depend on G.*

Assumption 1 imposes two conditional common trends conditions. First, it requires that for all $t \geq 1$, the evolution of the mean $Y(0)$ of units untreated at $T = t-1$ be the same in every group. Then, it requires that for all $t \geq 1$, the evolution of the mean $Y(1)$ of units treated at $T = t-1$ be the same in every group.

Theorem 4 *Suppose that D is binary and Assumption 10 holds.*

1. *If Assumptions 1-3 hold, then $W_{DID} = \Delta^S$.*

2. *If Assumptions 1' and 2 hold, then $W_{TC} = \Delta^S$.*

Theorem 4 shows that when "stable" groups exist for each pair of consecutive time periods, a weighted average of the Wald-DID (resp. Wald-TC) estimands proposed in de Chaisemartin and D'Haultfœuille (2018) identifies Δ^S under Assumptions 1-3 (resp. Assumptions 1' and 2). The averaging of those estimands is close to that proposed in section 1.2 of de Chaisemartin and D'Haultfœuille (2017), except that by reweighting observations at period $t-1$ and by considering slightly different weights $w_{g,t}$, the estimands proposed in this paper identify Δ^S, instead of a less interpretable weighted average of switchers' LATEs across periods.

Let us now give some intuition on the estimands W_{DID} and W_{TC}. Under the common trends, treatment monotonicity, and stable treatment effect assumptions, the evolution of the mean outcome in the "control group" ($G_t^* = 0$) identifies the evolution of the mean outcome we would have observed in the "treatment group" $G_t^* = 1$, if that group had not experienced any evolution of its treatment rate. Then, $DID_Y^*(1,t)$ identifies the treatment effect of the switchers in group $G_t^* = 1$ at $T = t$, times the proportion of switchers. $DID_D^*(1,t)$ identifies the share of switchers in that group, so $DID^*(1,t)/DID_D^*(1,t)$ identifies the LATE of switchers in group $G_t^* = 1$ at $T = t$. Similarly, $DID^*(-1,t)/DID_D^*(-1,t)$ identifies the LATE of switchers in group $G_t^* = -1$ at $T = t$. Finally, W_{DID} averages those estimands with weights equal to the proportion that each group accounts for in the total population of switchers, so it identifies the LATE of all switchers.

Turning to W_{TC}, let us start by giving the intuition underlying this estimand when the treatment is constant within each group × period cell and $G \perp\!\!\!\perp T$, as is for instance the case in Gentzkow et al. (2011). Then, let $NT_t = 1\{D_{G,t} = D_{G,t-1} = 0\}$ (resp. $AT_t = 1\{D_{G,t} = D_{G,t-1} = 1\}$) be an indicator for groups that are untreated (resp. treated) in periods $t-1$ and t, the "never treated" (resp. "always treated") groups. One can then show that

$$\begin{aligned} TCD^*(1,t)/DID_D^*(1,t) =& E(Y|G_t^* = 1, T = t) - E(Y|G_t^* = 1, T = t-1) \\ & - \left(E(Y|NT_t = 1, T = t) - E(Y|NT_t = 1, T = t-1)\right), \\ TCD^*(-1,t)/DID_D^*(-1,t) =& E(Y|AT_t = 1, T = t) - E(Y|AT_t = 1, T = t-1) \\ & - \left(E(Y|G_t^* = -1, T = t) - E(Y|G_t^* = -1, T = t-1)\right). \end{aligned}$$

When the treatment is constant within each group × period cell, $TCD^*(1,t)/DID_D^*(1,t)$ is a DID estimand comparing the evolution of the mean outcome between groups switching from being untreated to treated between $t-1$ and t, and groups that remain untreated between these two dates. Because these groups have the same treatment in $t-1$, under Assumption 1' this DID identifies the treatment effect in groups switching from being untreated to treated. Similarly, $TCD^*(-1,t)/DID^*_D(-1,t)$ is a DID estimand comparing the evolution of the mean outcome between groups that remain treated between $t-1$ and t, and groups switching from being treated to untreated between these two dates. Here again, these groups have the same treatment in $t-1$, so under Assumption 1' this DID identifies the treatment effect in groups switching from being treated to untreated. Finally, W_{TC} is a weighted average of those estimands.

Notice that in the staggered adoption design where $D = 1\{T \geq a_G\}$, W_{TC} has an even simpler expression. In this design, there is no group whose treatment diminishes between consecutive time periods, so W_{TC} is a weighted average of $TCD^*(1,t)/DID^*_D(1,t)$, with

$$\begin{aligned}\frac{TCD^*(1,t)}{DID^*_D(1,t)} =& E\left[Y|a_G = t, T = t\right] - E\left[Y|a_G = t, T = t-1\right] \\ & - \left(E(Y|a_G > t, T = t) - E(Q_0 Y|a_G > t, T = t-1)\right).\end{aligned}$$

In other words, W_{TC} is just a weighted average of the DIDs comparing the evolution of the outcome in groups that become treated at t and in groups not yet treated.

When the treatment is not constant within each group $\times$ period cell or G is not independent from T, the formula of W_{TC} is more complicated, but here is the intuition underlying it. Under Assumption 1', the trend affecting the $Y(0)$ (resp. $Y(1)$) of units with $D(t-1) = 0$ (resp. $D(t-1) = 1$) between $t-1$ and t is the same in every group. This trend is identified by the evolution of the mean of Y of untreated (resp. treated) units between $t-1$ and t in all "stable" groups with $G^*_t = 0$: under the treatment monotonicity assumption, one must have $D(t-1) = D(t)$ in those groups. Then, one can add the trend on $Y(0)$ (resp. $Y(1)$) to the outcome of untreated (resp. treated) units in group $G^*_t = 1$ in period $t-1$, and thus recover the mean outcome we would have observed in this group in period t if switchers had not changed treatment between the two periods. This is what $Y + \delta_D$ does. Therefore, $TCD^*(1,t)$ compares the mean outcome in group $G^*_t = 1$ at $T = t$ to the counterfactual mean we would have observed in that group at $T = t$ if switchers had remained untreated. Following the same logic as for W_{DID}, $TCD^*(1,t)/DID^*_D(1,t)$ identifies the LATE of switchers in group $G^*_t = 1$ at $T = t$ and W_{TC} identifies the LATE of all switchers.

It is worth noting that Theorem 4 can easily be extended to the case where the treatment is not binary, and to the case with covariates. Estimators of W_{DID} and W_{TC} are computed by the Stata package `fuzzydid`, see de Chaisemartin et al. (2018).

6 Discussion on estimation and inference

In this section, we discuss briefly how weights and the alternative estimands presented above can be estimated, and how inference can be conducted.

6.1 Inference on the weights and functions of the weights

We start by assuming that the population of groups only bears the $\overline{g}+1$ groups we observe, and that for each group $\times$ period, we observe an i.i.d. sample of the population of that group $\times$ period of size $n_{g,t}$. Let $\widehat{E}(D|G,T)$ and $\widehat{FD}_D(G,T)$ denote the empirical counterparts of $E(D|G,T)$ and $FD_D(G,T)$. Then let $\widehat{\varepsilon}_{fe,g,t}$ denote the residual of the regression of $\widehat{E}(D|G,T)$ on group and time fixed effects. We can then estimate $w_{fe,g,t}$ by

$$\widehat{w}_{fe,g,t} = \frac{\widehat{\varepsilon}_{fe,g,t}\sum_{(g',t')}\widehat{P}(G=g',T=t')\widehat{E}(D|G=g',T=t')}{\sum_{(g',t')}\widehat{P}(G=g',T=t')\widehat{\varepsilon}_{fe,g',t'}\widehat{E}(D|G=g',T=t')},$$

where $\widehat{P}(G=g,T=t)=n_{g,t}/\sum_{g',t'}n_{g',t'}$. The weights $w_{fd,g,t}$ can be estimated similarly. The weights $\omega_{fd,g,t}$ and $\omega_{fe,g,t}$ can also be estimated similarly, with one caveat. Those weights depend on $s_{g,t}=\text{sgn}(E(D_{g,t})-E(D_{g,t-1}))$, a discontinuous function of $(E(D_{g,t}),E(D_{g,t-1}))$. When the treatment is constant within each group $\times$ period, $s_{g,t}$ is known and does not need to be estimated. Otherwise, we consider the following estimator:

$$\widehat{s}_{g,t} = 1\{\widehat{E}(D_{g,t})-\widehat{E}(D_{g,t-1})>c_{n_{g,t}}\} - 1\{\widehat{E}(D_{g,t})-\widehat{E}(D_{g,t-1})<-c_{n_{g,t}}\}, \tag{12}$$

for some $c_{n_{g,t}}>0$. We introduce this threshold in order to obtain a consistent estimator of $s_{g,t}$, even in the case where $s_{g,t}=0$. For $k\in\{fe,fd\}$, $\widehat{w}_{k,g,t}$ is consistent and asymptotically normal as long as $n_{g,t}\to+\infty$, and $\widehat{\omega}_{k,g,t}$ is consistent and asymptotically normal as long as $n_{g,t}\to+\infty$, and $c_{n_{g,t}}\to 0$ and $\sqrt{n_{g,t}}c_{n_{g,t}}\to+\infty$ if $s_{g,t}$ has to be estimated. A similar result holds for estimators of functions of the weights such as $\underline{\sigma}^{TR}$ and $\underline{\sigma}^{S}$, because they are regular functions of $\widehat{\beta}_k$ and of the estimated weights. Inference can then be conducted using the bootstrap, by drawing samples from the samples of units in each group $\times$ period.

A limitation of the modelling framework outlined in the previous paragraph is that when the treatment is constant within each group $\times$ period, $\widehat{E}(D|G,T)=E(D|G,T)$ and $\widehat{FD}_D(G,T)=FD_D(G,T)$, so we for instance have $\widehat{\varepsilon}_{fe,g,t}=\varepsilon_{fe,g,t}$. Then, if the probabilities $(P(G=g,T=t))_{g,t}$ are also known,[8] the population weights are known and do not need to be estimated. Accordingly, any function of the weights is known without statistical uncertainty. Then, an alternative way of introducing statistical uncertainty is to assume that the $\overline{g}+1$ groups we observe are an i.i.d. sample from an infinite super-population of groups. Note that this modelling framework is in line with that underlying cluster-robust inference methods that are commonly used in DID analysis (see Bertrand et al., 2004). In that framework, we conjecture that it is possible to show that estimators of functions of the weights such as $\underline{\sigma}^{TR}$ and $\underline{\sigma}^{S}$ are still consistent and asymptotically normal, provided $\overline{g}+1\to+\infty$. Inference could then be conducted using the

[8] This is often the case. For instance, in an application where groups are US counties, the proportion each county accounts for in the US population is known.

bootstrap. Specifically, one could first draw samples of groups from the $\overline{g}+1$ groups in the sample, and then draw samples from the samples of units in each group × period.[9]

6.2 Alternative estimands

Our alternative estimands can also be estimated by plug-in estimators, where G_t^* is replaced by $\widehat{G}_t^* = \widehat{s}_{G,t}$, with $\widehat{s}_{g,t}$ defined in (12). For instance, we estimate the weights Q and Q_d by

$$\widehat{Q} = \frac{\widehat{r}(G|\widehat{G}^*_{T+1}, T+1, T+1)}{\widehat{r}(G|\widehat{G}^*_{T+1}, T+1, T)},\ \widehat{Q}_d = \frac{\widehat{r}_d(G|T+1, T+1)}{\widehat{r}_d(G|T+1, T)},$$

where

$$\widehat{r}(g|g^*, t, t') = \widehat{P}(G = g|\widehat{G}^*_{t'} = g^*, T = t),$$
$$\widehat{r}_d(g|g^*, t, t') = \widehat{P}(G = g|\widehat{G}^*_{t'} = g^*, T = t, D = d),$$

and the estimated probabilities on the right-hand sides are just sample proportions. Following the same reasoning as in the proof of Theorem S6 in de Chaisemartin and D'Haultfœuille (2017), we can show that the corresponding estimators of W_{DID} and W_{TC} are consistent and asymptotically normal. Again, inference can be conducted using the bootstrap.

7 Applicability, and applications

7.1 Applicability

We assess the pervasiveness of two-way fixed effects estimators in economics by conducting a review of all papers published in the American Economic Review (AER) between 2010 and 2012. Over these three years, the AER published 337 papers. This excludes papers and proceedings, comments, replies, and presidential addresses. Out of these 337 papers, 34 or 10.1% of them estimate Regression 1 or 2, or other two-way fixed effects regressions resembling closely Regression 1 or 2. When one withdraws from the denominator theory papers and lab experiments, the proportion of papers using these regressions raises to 19.5%. In the appendix, we review each paper and explain where it uses a two-way fixed effects estimator.

[9] This second step cannot be implemented when the data is aggregated at the group × period level.

Table 1: Two-way fixed effects papers published in the AER (2010-2012)

	2010	2011	2012	Total
Papers using two-way fixed effects estimators	5	15	14	34
% of published papers	5.2%	13.0%	11.2%	10.1%
% of empirical papers, excluding lab experiments	12.8%	24.6%	19.2%	19.7%

7.2 Applications

Enikolopov et al. (2011)

Enikolopov et al. (2011) study the effect of NTV, an independent TV channel introduced in 1996 in Russia, on the share of the electorate voting for opposition parties. NTV's coverage rate was heterogeneous across subregions: while a large fraction of the population received NTV in urbanized subregions, a smaller fraction received it in more rural subregions. The authors estimate the FE regression: they regress the share of votes for opposition parties in the 1995 and 1999 elections in Russian subregions on subregion fixed effects, an indicator for the 1999 election, and on the share of the population having access to NTV in each subregion at the time of the election. In 1995, the share of the population having access to NTV was equal to 0 in all subregions, while in 1999 it was strictly greater than 0 everywhere. Therefore, the authors' research design corresponds exactly to the heterogenous adoption design considered in Subsection 3.3. Enikolopov et al. (2011) find that $\widehat{\beta}_{fe} = 6.65$, with a standard error equal to 1.40. According to this regression, increasing the share of the population having access to NTV from 0 to 100% increases the share of votes for the opposition parties by 6.65 percentage points. Because $\overline{t} = 1$ and $T \perp\!\!\!\perp G$, $\widehat{\beta}_{fe} = \widehat{\beta}_{fd}$.

As no one was treated in 1995, the populations of treated units and switchers are the same, so the weights attached to β_{fe} under the common trends assumption and under the common trends, treatment monotonicity, and stable treatment effect assumptions are also the same. We estimate the weights $(P(G = g, T = 1|D = 1)w_{fe,g,1})_g$. 918 are strictly positive, while 1,020 are strictly negative. The negative weights sum to -2.26 (t-stat =-38.96).[10] Finally, we find that $\widehat{\underline{\sigma}}^{TR}_{fe}/\widehat{\beta}_{fe} = 0.148$ (95% level confidence interval=$[0.142, 0.153]$). Namely, β_{fe} and Δ^{TR} may be of opposite signs if the standard deviation of the $(\Delta^{TR}_{g,t})_{g,t}$ is equal to 15% of β_{fe}.

[10] To draw inference on the estimators we compute in this subsection, we use the bootstrap, clustered at the subregion level.

Therefore, the causal interpretation of β_{fe} relies on Assumption 4_{fe}. This assumption is not warranted. First, the effect of NTV is unlikely to be constant across Russian subregions: that effect could for instance be higher in more rural areas, as fewer other sources of independent information may be available there. Because the authors use aggregated data, we cannot use the overidentification test presented in Section 3 to test the constant treatment effect assumption. Instead, we estimate $\widehat{\beta}_{fe}$ again, weighting the regression by subregions' population. $\widehat{\beta}_{fe}$ is more than twice larger in the weighted than in the unweighted regression, and the difference between the two coefficients is statistically significant (t-stat=3.54). Second, the weights $(P(G = g, T = 1|D = 1)w_{fe,g,1})_g$ are not "randomly assigned" to subregions. For instance, those weights are strongly positively correlated with subregions' population (t-stat=14.15). Accordingly, the population of subregions receiving a positive weight is 46% larger than that of subregions receiving a negative weight.

Gentzkow et al. (2011)

Gentzkow et al. (2011) study the effect of newspapers on voters' turnout in US presidential elections between 1868 and 1928. They regress the first-difference of the turnout rate in county g between election years t and $t-1$ on state-year fixed effects and on the first difference of the number of newspapers available in that county. This regression corresponds to Regression 2, with state-year fixed effects as controls. Gentzkow et al. (2011) find that $\widehat{\beta}_{fd} = 0.0026$, with a standard error equal to 0.0009. According to this regression, one more newspaper increased voters' turnout by 0.26 percentage points. Because $\bar{t} > 1$, $\widehat{\beta}_{fe} \neq \widehat{\beta}_{fd}$. We find that $\widehat{\beta}_{fe} = -0.0011$, with a standard error equal to 0.0011.[11] $\widehat{\beta}_{fe}$ and $\widehat{\beta}_{fd}$ are significantly different (t-stat=3.64).

A large proportion of the weights attached to β_{fe} and β_{fd} under the common trends assumption are negative. We estimate the weights $(P(G = g, T = t|D = 1)w_{fd,g,t})$: 4,002 are strictly positive, 6,376 are strictly negative. The negative weights sum to -1.28 (t-stat =-9.43). Similarly, around 40% of the weights $(P(G = g, T = t|D = 1)w_{fe,g,t})_{g,t}$ are strictly negative. Therefore, under the common trends assumption the causal interpretation of β_{fe} and β_{fd} respectively relies on Assumption 4_{fe} and 4_{fd}. Those two assumptions are not warranted: the fact that $\widehat{\beta}_{fe}$ and $\widehat{\beta}_{fd}$ significantly differ implies that at least one of them must be violated.

Under the common trends, treatment monotonicity, and stable treatment effect assumptions, 25% of the weights attached to β_{fe} are still negative. On the other hand, all the weights attached to β_{fd} are positive: under those assumptions, β_{fd} identifies a convex combination of LATEs.

Still, that convex combination of LATEs is not equal to the LATE of all switchers, because it does not weight each group and period by its weight in the population of switchers. Moreover,

[11] To draw inference on the estimators we compute in this subsection, we use the bootstrap, clustered at the county level.

β_{fd} identifies a a convex combination of LATEs only if the stable treatment effect assumption holds, but this assumption may not be plausible in this context. This assumption requires that in counties with at least one newspaper in election year $t-1$, the effect of newspapers does not change between election years $t-1$ and t. However, newspapers' readership systematically decreases between consecutive elections. On average across pairs of consecutive elections and counties, and restricting the sample to counties with at least one newspaper in the first of the two consecutive elections, the fraction of a county's population reading the newspapers divided by the county's number of newspapers decreases by 1.0 percentage point between two consecutive elections (t-stat=-6.70). As newspapers tend to be less widely read in election-year t than in election-year $t-1$, their effect may decrease between consecutive elections.

The stable groups assumption holds in this application: between each pair of consecutive elections, there are counties where the number of newspapers did not change. We can then estimate W_{TC}, the estimand proposed in Section 5. That estimand identifies the LATE of all switchers, and contrary to β_{fd} it does not rely on any restriction on treatment effect heterogeneity. Instead, it relies on Assumption 1', a conditional common trends assumption.[12] We find that $\widehat{W}_{TC} = 0.0043$, with a standard error of 0.0014. $\widehat{W}_{TC}$ is 66% larger than, and significantly different from, $\widehat{\beta}_{fd}$ at the 10% level (t-stat=1.83). $\widehat{W}_{TC}$ is also of a different sign than $\widehat{\beta}_{fe}$.

8 Conclusion

Almost 20% of empirical articles published in the AER between 2010 and 2012 use regressions with groups and period fixed effects to estimate treatment effects. While it is well-known that such regressions identify the treatment effect of interest if that effect is constant and if the standard common trends assumption is satisfied, those regressions have not yet been studied in a model allowing for treatment effect heterogeneity. In this paper, we start by showing that under the common trends assumption alone, two pervasive two-way fixed effects regressions identify weighted sums of the ATTs in each group and at each period. Many of the weights attached to those regressions may be negative: in two empirical applications, we find that more than 50% of the weights are negative. When many weights are negative, two-way fixed effects regressions are not robust to heterogeneous treatment effects across groups and periods: the coefficient of the treatment variable in those regressions may for instance be negative while the treatment effect is positive for every unit in the population.

Then, we consider two supplementary assumptions. The first one requires that in each group

[12] de Chaisemartin and D'Haultfœuille (2017) conduct a placebo test of Assumption 1'. They find that conditional on their number of newspapers in $t-1$, counties with different evolutions of their number of newspapers from $t-1$ to t do not experience different evolutions of their turnout from $t-2$ to $t-1$. This suggests that Assumption 1' is plausible in this application.

and for each pair of consecutive periods, the average treatment effect of units treated at period $t-1$ be stable from $t-1$ to t. The second one requires that in each group and for each pair of consecutive periods, the treatment follows a monotonic evolution from $t-1$ to t. Under the common trends assumption and those two supplementary assumptions, we show that our two-way fixed effects regressions identify weighted sums of the LATEs of switchers in each group and at each period, where switchers are units whose treatment changes between two consecutive time periods. Here again, some of the weights may be negative. However, in some special cases the weights attached to the second regression we consider are all positive.

Finally, we propose a new estimand. This estimand identifies the LATE of all switchers, and it does not rely on any treatment effect homogeneity condition. It can be used in applications where there are groups whose exposure to the treatment does not change between each pair of consecutive time periods. In one of the two applications we revisit, the corresponding estimator is very different from the two two-way fixed effects estimators we consider.

References

Aaronson, D., Agarwal, S. and French, E. (2012), 'The spending and debt response to minimum wage hikes', *The American Economic Review* **102**(7), 3111–3139.

Abadie, A. (2005), 'Semiparametric difference-in-differences estimators', *Review of Economic Studies* **72**(1), 1–19.

Acemoglu, D., Cantoni, D., Johnson, S. and Robinson, J. A. (2011), 'The consequences of radical reform: The french revolution', *American Economic Review* **101**(7), 3286–3307.

Aizer, A. (2010), 'The gender wage gap and domestic violence', *The American economic review* **100**(4), 1847.

Algan, Y. and Cahuc, P. (2010), 'Inherited trust and growth', *The American Economic Review* pp. 2060–2092.

Anderson, S. T. and Sallee, J. M. (2011), 'Using loopholes to reveal the marginal cost of regulation: The case of fuel-economy standards', *American Economic Review* **101**(4), 1375–1409.

Angrist, J. D. (1998), 'Estimating the labor market impact of voluntary military service using social security data on military applicants', *Econometrica* **66**(2), 249–288.

Angrist, J. D., Graddy, K. and Imbens, G. W. (2000), 'The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish', *The Review of Economic Studies* **67**(3), 499–527.

Angrist, J. D. and Imbens, G. W. (1995), 'Two-stage least squares estimation of average causal effects in models with variable treatment intensity', *Journal of the American Statistical Association* **90**, pp. 431–442.

Athey, S. and Stern, S. (2002), 'The impact of information technology on emergency health care outcomes', *The RAND Journal of Economics* **33**(3), 399–432.

Bagwell, K. and Staiger, R. W. (2011), 'What do trade negotiators negotiate about? empirical evidence from the world trade organization', *American Economic Review* **101**(4), 1238–73.

Bajari, P., Fruehwirth, J. C., Kim, K. I. and Timmins, C. (2012), 'A rational expectations approach to hedonic price regressions with time-varying unobserved product attributes: The price of pollution', *The American Economic Review* **102**(5), 1898–1926.

Baum-Snow, N. and Lutz, B. F. (2011), 'School desegregation, school choice, and changes in residential location patterns by race', *American Economic Review* **101**(7), 3019–46.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004), 'How much should we trust differences-in-differences estimates?', *The Quarterly journal of economics* **119**(1), 249–275.

Besley, T. and Mueller, H. (2012), 'Estimating the peace dividend: The impact of violence on house prices in northern ireland', *The American Economic Review* **102**(2), 810–833.

Bloom, N., Sadun, R. and Van Reenen, J. (2012), 'Americans do it better: Us multinationals and the productivity miracle', *The American Economic Review* **102**(1), 167–201.

Borusyak, K. and Jaravel, X. (2016), 'Revisiting event study designs', *Browser Download This Paper* .

Brambilla, I., Lederman, D. and Porto, G. (2012), 'Exports, export destinations, and skills', *The American Economic Review* **102**(7), 3406–3438.

Bustos, P. (2011), 'Trade liberalization, exports, and technology upgrading: Evidence on the impact of mercosur on argentinian firms', *American Economic Review* **101**(1), 304–40.

Chandra, A., Gruber, J. and McKnight, R. (2010), 'Patient cost-sharing and hospitalization offsets in the elderly', *The American economic review* **100**(1), 193.

Chaney, T., Sraer, D. and Thesmar, D. (2012), 'The collateral channel: How real estate shocks affect corporate investment', *The American Economic Review* **102**(6), 2381–2409.

Dafny, L., Duggan, M. and Ramanarayanan, S. (2012), 'Paying a premium on your premium? consolidation in the us health insurance industry', *The American Economic Review* **102**(2), 1161–1185.

Dahl, G. B. and Lochner, L. (2012), 'The impact of family income on child achievement: Evidence from the earned income tax credit', *The American Economic Review* **102**(5), 1927–1956.

de Chaisemartin, C. and D'Haultfoeuille, X. (2015), 'Fuzzy Differences-in-Differences', *ArXiv e-prints, eprint 1510.01757v2* .

de Chaisemartin, C. and D'Haultfœuille, X. (2017), Supplement to "fuzzy differences-in-differences", Technical report.

de Chaisemartin, C. and D'Haultfœuille, X. (2018), 'Fuzzy differences-in-differences', *The Review of Economic Studies* **85**(2), 999–1028.
URL: *http://dx.doi.org/10.1093/restud/rdx049*

de Chaisemartin, C., D'Haultfœuille, X. and Guyonvarch, Y. (2018), Fuzzy differences-in-differences with stata, Technical report.

Dinkelman, T. (2011), 'The effects of rural electrification on employment: New evidence from south africa', *American Economic Review* **101**(7), 3078–3108.

Draca, M., Machin, S. and Witt, R. (2011), 'Panic on the streets of london: Police, crime, and the july 2005 terror attacks', *The American Economic Review* pp. 2157–2181.

Duggan, M. and Morton, F. S. (2010), 'The effect of medicare part d on pharmaceutical prices and utilization', *The American Economic Review* pp. 590–607.

Duranton, G. and Turner, M. A. (2011), 'The fundamental law of road congestion: Evidence from us cities', *American Economic Review* **101**(6), 2616–52.

Ellul, A., Pagano, M. and Panunzi, F. (2010), 'Inheritance law and investment in family firms', *The American Economic Review* pp. 2414–2450.

Enikolopov, R., Petrova, M. and Zhuravskaya, E. (2011), 'Media and political persuasion: Evidence from russia', *The American Economic Review* **101**(7), 3253.

Fang, H. and Gavazza, A. (2011), 'Dynamic inefficiencies in an employment-based health insurance system: Theory and evidence', *American Economic Review* **101**(7), 3047–77.

Faye, M. and Niehaus, P. (2012), 'Political aid cycles', *The American Economic Review* **102**(7), 3516–3530.

Forman, C., Goldfarb, A. and Greenstein, S. (2012), 'The internet and local wages: A puzzle', *The American Economic Review* pp. 556–575.

Gentzkow, M., Shapiro, J. M. and Sinkinson, M. (2011), 'The effect of newspaper entry and exit on electoral politics', *The American Economic Review* **101**(7), 2980.

Hornbeck, R. (2012), 'The enduring impact of the american dust bowl: Short-and long-run adjustments to environmental catastrophe', *The American Economic Review* **102**(4), 1477–1507.

Hotz, V. J. and Xiao, M. (2011), 'The impact of regulations on the supply and quality of care in child care markets', *American Economic Review* **101**(5), 1775–1805.

Imberman, S. A., Kugler, A. D. and Sacerdote, B. I. (2012), 'Katrina's children: Evidence on the structure of peer effects from hurricane evacuees', *The American Economic Review* **102**(5), 2048–2082.

Mian, A. and Sufi, A. (2011), 'House prices, home equity-based borrowing, and the us household leverage crisis', *American Economic Review* **101**(5), 2132–56.

Moser, P. and Voena, A. (2012), 'Compulsory licensing: Evidence from the trading with the enemy act', *The American Economic Review* pp. 396–427.

Simcoe, T. (2012), 'Standard setting committees: Consensus governance for shared technology platforms', *The American Economic Review* **102**(1), 305–336.

Słoczyński, T. (2014), 'New evidence on linear regression and treatment effect heterogeneity'.

Wang, S.-Y. (2011), 'State misallocation and housing prices: Theory and evidence from china', *American Economic Review* **101**(5), 2081–2107.

White, H. (1980), 'Using least squares to approximate unknown regression functions', *International Economic Review* pp. 149–170.

Zhang, X. M. and Zhu, F. (2011), 'Group size and incentives to contribute: A natural experiment at chinese wikipedia', *American Economic Review* **101**(4).

A Detailed literature review

We review here the 34 papers using two-way fixed effects estimators we found in our literature review. For each paper, we use the following presentation:
Authors (year), Title. *Where the two-way fixed effects estimator is used in the paper.*
Description of the two-way fixed effects estimator and how it relates to Regressions 1, 2 or 3, or to the Wald-DID estimator.

1. **Chandra et al. (2010), Patient Cost-Sharing and Hospitalization Offsets in the Elderly.** *Elasticities of care use to co-payment estimated after Tables 2 and 3.*
   The elasticity discussed after, say, Table 2 is estimated as the ratio of the effect of the Medicare reform on utilization, divided by the effect of the Medicare reform on co-payment. Both effects are estimated through standard DID regressions in Table 2. Therefore, the elasticity estimate is a Wald-DID.

2. **Duggan and Morton (2010), The Effect of Medicare Part D on Pharmaceutical Prices and Utilization.** *Tables 2 and 3.*
   In regression equation (1), the dependent variable is the change in the price of drug j between 2003 and 2006, and the explanatory variable is the Medicare market share for drug j in 2003. This regression corresponds to Regression 2.

3. **Aizer (2010), The Gender Wage Gap and Domestic Violence.** *Table 2.*
   In regression equation (2), the dependent variable is the log of female assaults among females of race r in county c and year t, and the explanatory variables are race, year, county, race × year, race × county, and county × year fixed effects, as well as the gender wage gap in county c, year t, and race r. This regression is a "triple difference" version of Regression 1.

4. **Algan and Cahuc (2010), Inherited Trust and Growth.** *Figure 4 and Table 6.*
   Figure 4 presents a regression of changes in income per capita from 1935 to 2000 on changes in inherited trust over the same period and a constant. This regression corresponds to Regression 2.

5. **Ellul et al. (2010), Inheritance Law and Investment in Family Firms.** *Table 7.*
   In the regressions presented in Table 7, the dependent variable is the capital expenditure of firm j in year t, and the explanatory variables are firm fixed effects, an indicator for whether year t is a succession period for firm j, and the interaction of this indicator with the level of investor protection in the country where firm j is located. This regression corresponds to Regression 1, with two periods (succession and no succession).

6. **Bustos (2011), Trade Liberalization, Exports, and Technology Upgrading: Evidence on the Impact of MERCOSUR on Argentinean Firms.** *Tables 3 to 12.*
   In regression equation (11), the dependent variable is the change in exporting status of firm i in sector j between 1992 and 1996, and the explanatory variable is the change in trade tariffs in Brasil for products in sector j over the same period. This regression corresponds to Regression 2.

7. **Anderson and Sallee (2011), Using Loopholes to Reveal the Marginal Cost of Regulation: The Case of Fuel-Economy Standards.** *Table 5 column 2.*
   In the regression in Table 5 column (2), the dependent variable is an indicator for whether a car sold is a flexible fuel vehicle, and the explanatory variables are state and month fixed effects, and the percent ethanol availability in each month $\times$ state. This regression corresponds to Regression 1.

8. **Bagwell and Staiger (2011), What Do Trade Negotiators Negotiate About? Empirical Evidence from the World Trade Organization.** *Table 3, OLS columns.*
   In regression equations (15a) and (15b), the dependent variable is the ad valorem tariff level bound by country c on product g, while the explanatory variables are country and product fixed effects, and two treatment variables which vary at the country $\times$ product level. These regressions are similar to Regression 1, except that they have two treatment variables.

9. **Zhang and Zhu (2011), Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia.** *Tables 3 and 4, columns 4-6.*
   In the regression in, say, Table 3 column (4), the dependent variable is the total number of contributions to Wikipedia by individual i at period t, regressed on individual fixed effects, an indicator for whether period t is after the Wikipedia block, and the interaction of this indicator and a measure of social participation by individual i. This regression corresponds to Regression 1 (treatment is equal to 0 before the block, and to social participation after it).

10. **Draca et al. (2011), Panic on the Streets of London: Police, Crime, and the July 2005 Terror Attacks.** *Table 2, Panel C, Columns 3-4.*
    In regression equation (7), the dependent variable is the change in crime rates between week t and the same week one year ago in borough b, and the explanatory variables are an indicator for whether week t is around the terrorist attacks in London, and the number of police forces in borough b in week t. The interaction of the time indicator and of whether borough b belongs to Theseus operation is used as the excluded instrument for police forces. This regression is equivalent to Regression 3 (borough fixed effects disappear

because of the first differencing with respect to the previous year, something the authors do to control for seasonality).

11. **Hotz and Xiao (2011), The Impact of Regulations on the Supply and Quality of Care in Child Care Markets.** *Table 7, Columns 4 and 5.*
In Regression Equation (1), the dependent variable is the outcome for market m in state s and year t, and the explanatory variables are state and year fixed effects and a measure of regulations in state s in year t. This regression corresponds to Regression 1.

12. **Mian and Sufi (2011), House Prices, Home Equity-Based Borrowing, and the US Household Leverage Crisis.** *Tables 2 and 3.*
Regression equations (1) and (2) are first-difference versions of Regression 3. In levels, the instrument would be the elasticity interacted with the year 2006. Because the data only bears two periods, the two regressions are algebraically equivalent.

13. **Wang (2011), State Misallocation and Housing Prices: Theory and Evidence from China.** *Table 5, Panel A.*
In regression equation (15), the dependent variable is a measure of the quantity of housing services in household i's residence in year t, while the explanatory variables are an indicator for period t being after the reform, a measure of mismatch in household i, and the interaction of the measure of mismatch and the time indicator. This regression is similar to Regression 1, except that it has a measure a mismatch in household i instead of household fixed effects.

14. **Duranton and Turner (2011), The Fundamental Law of Road Congestion: Evidence from US Cities.** *Table 5.*
In regression equation (4), the dependent variable is the change in vehicle kilometers traveled in MSA s between periods t and t-1, and the explanatory variable is the change in kilometers of roads in MSA s between periods t and t-1. This regression corresponds to Regression 2.

15. **Acemoglu et al. (2011), The Consequences of Radical Reform: The French Revolution.** *Table 3.*
In regression equation (1), the dependent variable is urbanization in polity j at time t, while the explanatory variables are time and polity fixed effects, and the number of years of French presence in polity j interacted with the time effects. This regression corresponds to Regression 1.

16. **Baum-Snow and Lutz (2011), School Desegregation, School Choice, and Changes in Residential Location Patterns by Race.** *Table 6.*

In the regression presented in, say, the first column of Table 6, the dependent variable is enrolment in schools of MSA j in year t, while the explanatory variables are time and MSA effects and the value of the dissimilarity index of schools in MSA j in year t. The excluded instrument for the dissimilarity index is an indicator for whether in period t, the MSA was desegregated. This regression is similar to Regression 3.

17. **Dinkelman (2011), The Effects of Rural Electrification on Employment: New Evidence from South Africa.** *Tables 4 and 5 columns 5-8, Table 8 columns 3-4, Table 9 column 2, and Table 10 columns 2, 4, and 6.*
Regression equation (4) is the first-difference version of Regression 3. In levels, the instrument would be the land gradient Z_j interacted with an indicator for the second wave of the panel. Because the data only bears two periods, the two regressions are algebraically equivalent.

18. **Enikolopov et al. (2011), Media and Political Persuasion: Evidence from Russia.** *Table 3.*
In regression equation (5), the dependent variable is the share of votes for party j in year t and subregion s, and the explanatory variables are subregion and time effects, and the share of people having access to NTV in subregion s in period t. This regression corresponds to Regression 1.

19. **Fang and Gavazza (2011), Dynamic Inefficiencies in an Employment-Based Health Insurance System: Theory and Evidence.** *Tables 2, 3, 5, and 6, Column 3.*
In regression equation (7), the dependent variable is the health expenditures of individual j working in industry i in period t and region r, and the explanatory variables are individual effects, region specific time effects, and the job tenure of individual j. The death rate of establishments in industry i in period t and region r is used as an instrument for the job tenure of individual j. Within each region, the regression has time effects and individual effects, and an instrument varying only across industry × periods. This regression is similar to Regression 3.

20. **Gentzkow et al. (2011), The Effect of Newspaper Entry and Exit on Electoral Politics.** *Tables 2 and 3.*
In regression equation (1), the dependent variable is the change in voter turnout in county c between elections year t and t-1, and the explanatory variables are state × year effects, and the number of newspapers in county c in year t. This regression corresponds to Regression 2, except that it allows for state specific trends.

21. **Bloom et al. (2012), Americans Do IT Better: US Multinationals and the Productivity Miracle.** *Table 2, Columns 6-8.*

In the regression in, say, column 6 of Table 2, the dependent variable is the log of output per worker in firm i in period t, while the explanatory variables are firms and time fixed effects, and the log of the amount of IT capital per employee ($\ln(C/L)$) as well as the interaction of $\ln(C/L)$ and an indicator for whether the firm is owned by a US multinational. This regression is similar to Regression 1, except that it has two treatment variables.

22. **Simcoe (2012), Standard Setting Committees: Consensus Governance for Shared Technology Platforms.** *Table 4, columns 1-3.*
In regression equation (5), the dependent variable is a measure of time to consensus for project i submitted to committee j, while the explanatory variables are an indicator for projects submitted to the standards track, a measure of distributional conflict, and the interaction of the standards track and distributional conflict. This regression is similar to Regression 1, except that it has a measure of distributional conflict instead of committee fixed effects.

23. **Moser and Voena (2012), Compulsory Licensing: Evidence from the Trading with the Enemy Act.** *Table 2, columns 3-8.*
In the regression equation in the beginning of Section III, the dependent variable is the number of patents by US inventors in patent class c at period t, and the explanatory variables are patent class and time fixed effects, and the interaction of period t being after the trading with the enemy act and a measure of treatment intensity. This regression corresponds to Regression 1 (treatment is equal to 0 before the act).

24. **Forman et al. (2012), The Internet and Local Wages: A Puzzle.** *Tables 2 and 4.*
In regression equation (1), the dependent variable is the difference between log wages in 2000 and 1995 in county i, and the explanatory variable is Internet investment by businesses in county i in 2000. This regression corresponds to Regression 2. Table 4 presents regressions where advanced internet investment is instrumented by a county level variable. This regression is the first-difference version of Regression 3. Because the data only bears two periods, these two regressions are algebraically equivalent.

25. **Besley and Mueller (2012), Estimating the Peace Dividend: The Impact of Violence on House Prices in Northern Ireland.** *Table 1, columns 3 and 5-7.*
In regression equation (1), the dependent variable is the price of houses in region r at time t, while the explanatory variables include region and time fixed effects, and the number of people killed because of the civil war in region r at time t-1. This regression corresponds to Regression 1.

26. **Dafny et al. (2012), Paying a Premium on Your Premium? Consolidation in the US Health Insurance Industry.** *Tables 2 and 5.*

In regression equation (1), the dependent variable is the change of the log premium for employer e in market m in year t, and explanatory variables are time and market fixed effects, and the change in various treatment variables (change in the fraction of self-insured employees...). This regression is similar to Regression 2, except that it has several treatment variables, and market and time fixed effects.

27. **Hornbeck (2012), The Enduring Impact of the American Dust Bowl: Short- and Long-Run Adjustments to Environmental Catastrophe.** *Table 2.* In regression equation (1), the dependent variable is, say, the change in log land value in county c between period t and 1930, and the explanatory variables are state × year effects, the share of county c in high erosion, and the share of county c in medium erosion. This regression is similar to Regression 1, except that it has two treatment variables and state-year fixed effects.

28. **Bajari et al. (2012), A Rational Expectations Approach to Hedonic Price Regressions with Time-Varying Unobserved Product Attributes: The Price of Pollution.** *Table 5.*
    In, say, the first regression equation in the bottom of page 1915, the dependent variable is the change in the price of house j between sales 2 and 3, and the explanatory variables are the change in various pollutants in the area around house j between sales 2 and 3. This regression is similar to Regression 2, except that it has several treatment variables.

29. **Dahl and Lochner (2012), The Impact of Family Income on Child Achievement: Evidence from the Earned Income Tax Credit.** *Table 3.*
    In the reduced-form of regression equation (4), the dependent variable is the change in test scores for child i between years a and a-1, while the explanatory variable is the change in the expected EITC income of her family based on her family income in year a-1. This regression corresponds to Regression 2. The first stage is the same regression but with the change in the income of the family of student i between years a and a-1. Overall, the 2SLS coefficient arising from regression equation (4) is a ratio of 2 weighted averages of Wald-DIDs.

30. **Imberman et al. (2012), Katrina's Children: Evidence on the Structure of Peer Effects from Hurricane Evacuees.** *Tables 3-6.*
    In regression equation (1), the dependent variable is the test score of student i in school j in grade g and year t, and the explanatory variables are grade, school, year, and grade × year effects, and the fraction of Katrina students received by school j in grade g and year t. Within each grade, this regression corresponds to Regression 1.

31. **Chaney et al. (2012), The Collateral Channel: How Real Estate Shocks Affect Corporate Investment.** *Table 5.*

In regression equation (1), the dependent variable is the value of investment in firm i and year t divided by the lagged book value of properties, plants, and equipments (PPE), and the explanatory variables are firm and time fixed effects and the market value of firm i in year t divided by its lagged PPE. This regression corresponds to Regression 1.

32. **Aaronson et al. (2012), The Spending and Debt Response to Minimum Wage Hikes.** *Tables 1, 2, and 5.*
In regression equation (1), the outcome variable is, say, income of household i at period t, and the explanatory variables include household and time fixed effects, and the minimum wage in the state where household i lives in period t. This regression corresponds to Regression 1.

33. **Brambilla et al. (2012), Exports, Export Destinations, and Skills.** *Table 5.*
In regression equation (7), the dependent variable is a measure of skills in the labor force employed by company i in industry j at period t, and the explanatory variables are firm and industry × period fixed effects, the ratio of exports to sales in firm i at period t, and the share of firm exports to high income destinations over total exports. To instrument this variable, the authors use an indicator for the years 1999 or 2000 (a large devaluation happened in Brazil in 1999) interacted with the share of exports of firm i to Brazil in 1998. This regression corresponds to Regression 3.

34. **Faye and Niehaus (2012), Political Aid Cycles.** *Table 3, columns 4 and 5, and Tables 4 and 5.*
In regression equation (2), the dependent variable is the amount of donations received by receiver r from donor d in year t, and the explanatory variables are donor × receiver fixed effects, an indicator for whether there is an election in country r in year t, a measure of alignment between the ruling political parties in countries r and d, and the interaction of the election indicator and the measure of alignment. This regression corresponds to Regression 1.

B Proofs

B.1 Two useful lemmas

For any random variable R, and for all $(g,g',t,t') \in \{0,...,\overline{g}\}^2 \times \{1,...,\overline{t}\}^2$, let

$$DID_R(g,g',t,t') = E(R_{g,t}) - E(R_{g,t'}) - \left(E(R_{g',t}) - E(R_{g',t'})\right).$$

Our lemma relates the $DID_Y(g,g',t,t')$ estimands to the $\text{ACR}^{TR}_{g,t}$ and $\text{ACR}^{S}_{g,t}$ parameters.

Lemma 1 *Assume that* $D \in \{0, , ..., \overline{d}\}$.

1. *If Assumption 1 holds, for all* $(g, g', t, t') \in \{0, ..., \overline{g}\}^2 \times \{1, ..., \overline{t}\}^2$

$$DID_Y(g, g', t, t') = E(D_{g,t})ACR^{TR}_{g,t} - E(D_{g,t'})ACR^{TR}_{g,t'} - \left(E(D_{g',t})ACR^{TR}_{g',t} - E(D_{g',t'})ACR^{TR}_{g',t'}\right).$$

2. *If Assumptions 1, 2, and 3O hold, , for all* $(g, g', t, t') \in \{0, ..., \overline{g}\}^2 \times \{1, ..., \overline{t}\}^2$

$$DID_Y(g, g', t, t-1) = (E(D_{g,t}) - E(D_{g,t-1}))\, ACR^S_{g,t} - (E(D_{g',t}) - E(D_{g',t-1}))\, ACR^S_{g',t}.$$

In the special case where the treatment is binary, Lemma 1 can be rewritten as follows.

Lemma 2 *Assume that D is binary.*

1. *If Assumption 1 holds, for all* $(g, g', t, t') \in \{0, ..., \overline{g}\}^2 \times \{1, ..., \overline{t}\}^2$

$$DID_Y(g, g', t, t') = E(D_{g,t})\Delta^{TR}_{g,t} - E(D_{g,t'})\Delta^{TR}_{g,t'} - \left(E(D_{g',t})\Delta^{TR}_{g',t} - E(D_{g',t'})\Delta^{TR}_{g',t'}\right).$$

2. *If Assumptions 1-3 hold, for all* $(g, g', t, t') \in \{0, ..., \overline{g}\}^2 \times \{1, ..., \overline{t}\}^2$

$$DID_Y(g, g', t, t-1) = (E(D_{g,t}) - E(D_{g,t-1}))\, \Delta^S_{g,t} - (E(D_{g',t}) - E(D_{g',t-1}))\, \Delta^S_{g',t}.$$

Proof of Lemma 1

1. We have

$$DID_Y(g, g', t, t') = E(Y_{g,t}) - E(Y_{g,t'}) - (E(Y_{g',t}) - E(Y_{g',t'})). \tag{13}$$

Moreover,

$$E(Y_{g,t}) = E(Y_{g,t}(0)) + E[Y_{g,t}(D) - Y_{g,t}(0)]. \tag{14}$$

The result follows by decomposing similarly the three other terms of $DID_Y(g, g', t, t')$, plugging these decompositions into (13), using Assumption 1, and finally using the definition of $\text{ACR}^{TR}_{g,t}$.

2. We prove the result when $E(D_{g,t}) \geq E(D_{g,t-1})$ and $E(D_{g',t}) \geq E(D_{g',t-1})$. The proof is similar in the three other cases. First,

$$\begin{aligned} E(Y_{g,t-1})) &= E(Y_{g,t-1}(0)) + \sum_{d=1}^{\bar{d}} P(D_{g,t-1}(t-1) = d)E(Y_{g,t-1}(d) - Y_{g,t-1}(0)|D(t-1) = d) \\ &= E(Y_{g,t-1}(0)) + \sum_{d=1}^{\bar{d}} P(D_{g,t}(t-1) = d)E(Y_{g,t}(d) - Y_{g,t}(0)|D(t-1) = d). \end{aligned} \tag{15}$$

where the second equality follows from Assumptions 2 and 3O. Similarly,

$$E(Y_{g,t})) = E(Y_{g,t}(0)) + \sum_{d=1}^{\bar{d}} P(D_{g,t}(t) = d)E(Y_{g,t}(d) - Y_{g,t}(0)|D(t) = d). \tag{16}$$

Combining Equations (15) and (16) yields

$$E(Y_{g,t}) - E(Y_{g,t-1}) = E(Y_{g,t}(0)) - E(Y_{g,t-1}(0)) + E\left[\sum_{d=1}^{\bar{d}} (Y_{g,t}(d) - Y_{g,t}(0)) \left(\mathbb{1}\{D_{g,t}(t) = d\} - \mathbb{1}\{D_{g,t}(t-1) = d\}\right)\right]. \quad (17)$$

Now, remark that

$$\begin{aligned} &\sum_{d=1}^{\bar{d}} (Y_{g,t}(d) - Y_{g,t}(0)) \left(\mathbb{1}\{D_{g,t}(t) = d\} - \mathbb{1}\{D_{g,t}(t-1) = d\}\right) \\ =&\sum_{d=1}^{\bar{d}} (Y_{g,t}(d) - Y_{g,t}(d-1)) \left(\mathbb{1}\{D_{g,t}(t) \geq d\} - \mathbb{1}\{D_{g,t}(t-1) \geq d\}\right) \\ =&\sum_{d=1}^{\bar{d}} (Y_{g,t}(d) - Y_{g,t}(d-1)) \mathbb{1}\{D_{g,t}(t) \geq d > D_{g,t}(t-1)\}. \quad (18) \end{aligned}$$

The first equality follows by summation by parts. The second uses the fact that under Assumption 2, $E(D_{g,t}) \geq E(D_{g,t-1})$ implies that $D_{g,t}(t) \geq D_{g,t}(t-1)$. Then,

$$\begin{aligned} &E(Y_{g,t}) - E(Y_{g,t-1}) - (E(Y_{g,t}(0)) - E(Y_{g,t-1}(0))) \\ =&\sum_{d=1}^{\bar{d}} P(D_{g,t}(t) \geq d > D_{g,t}(t-1)) E\left(Y_{g,t}(d) - Y_{g,t}(d-1) | D_{g,t}(t) \geq d > D_{g,t}(t-1)\right) \\ =&\sum_{d=1}^{\bar{d}} \left[P(D_{g,t} \geq d) - P(D_{g,t-1} \geq d)\right] E\left(Y_{g,t}(d) - Y_{g,t}(d-1) | D_{g,t}(t) \geq d > D_{g,t}(t-1)\right) \\ =&(E(D_{g,t}) - E(D_{g,t-1}))\text{ACR}^S_{g,t}. \quad (19) \end{aligned}$$

The first equality follows from plugging Equation (18) into Equation (17). The second uses the fact that under Assumption 2, $E(D_{g,t}) \geq E(D_{g,t-1})$ implies that $D_{g,t}(t) \geq D_{g,t}(t-1)$.

One can follow the same steps to show that

$$E(Y_{g',t}) - E(Y_{g',t-1}) - (E(Y_{g',t}(0)) - E(Y_{g',t-1}(0))) = (E(D_{g',t}) - E(D_{g',t-1}))\text{ACR}^S_{g',t}. \quad (20)$$

Finally, the result follows by taking the difference between Equations (19) and (20), and using Assumption 1.

Proof of Lemma 2

The first statement of the lemma follows from the first statement of Lemma 1, once noted that when D is binary, $ACR^{TR}_{g,t} = \Delta^{TR}_{g,t}$ for all $(g,t) \in \{0, ..., \overline{g}\} \times \{1, ..., \overline{t}\}$. The second statement follows from the second statement of Lemma 1, once noted that when D is binary, Assumptions 3 and 3O are the same, and $ACR^S_{g,t} = \Delta^S_{g,t}$ for all $(g,t) \in \{0, ..., \overline{g}\} \times \{1, ..., \overline{t}\}$.

B.2 Proof of Theorem 1

Proof of the first statement.

Proof for the fixed-effect regression.

It follows from the Frisch-Waugh theorem and the definition of $\varepsilon_{fe,G,T}$ that

$$\beta_{fe} = \frac{\text{cov}(\varepsilon_{fe,G,T}, Y)}{\text{cov}(\varepsilon_{fe,G,T}, E(D|G,T))}. \tag{21}$$

Note that the denominator must be different from 0, otherwise β_{fe} would not be well defined. As a result,

$$E(\varepsilon_{fe,G,T}E(D|G,T)) = \text{cov}(\varepsilon_{fe,G,T}, E(D|G,T)) \neq 0. \tag{22}$$

Now, by definition of $\varepsilon_{fe,G,T}$ again,

$$E[\varepsilon_{fe,G,T}|G] = 0,\ E[\varepsilon_{fe,G,T}|T] = 0. \tag{23}$$

Then,

$$\begin{aligned}
\text{cov}(\varepsilon_{fe,G,T}, Y) &= E[\varepsilon_{fe,G,T}E(Y|G,T)] \\
&= E\left[\varepsilon_{fe,G,T}DID_Y(G,0,T,0)\right] \\
&= E\left[\varepsilon_{fe,G,T}\left(E(D|G,T)\Delta^{TR}_{G,T} - E(D_{G,0})\Delta^{TR}_{G,0} - E(D_{0,T})\Delta^{TR}_{0,T} + E(D_{0,0})\Delta^{TR}_{0,0}\right)\right] \\
&= E\left[\varepsilon_{fe,G,T}E(D|G,T)\Delta^{TR}_{G,T}\right] \\
&= E\left[\varepsilon_{fe,G,T}D\Delta^{TR}_{G,T}\right]. \qquad (24)
\end{aligned}$$

The first equality follows from the law of iterated expectations. (23) implies that

$$E\left[\varepsilon_{fe,G,T}(-E(Y_{G,0}) - E(Y_{0,T}) + E(Y_{0,0}))\right] = 0,$$

hence the second equality. The third equality follows from the first point of Lemma 2. (23) implies that

$$E\left[\varepsilon_{fe,G,T}(-E(D_{G,0})\Delta^{TR}_{G,0} - E(D_{0,T})\Delta^{TR}_{0,T} + E(D_{0,0})\Delta^{TR}_{0,0})\right] = 0,$$

hence the fourth equality. The fifth equality follows from the law of iterated expectations. Combining (21), (22), and (24), we obtain

$$\begin{aligned}
\beta_{fe} &= \frac{E\left[\varepsilon_{fe,G,T}D\Delta^{TR}_{G,T}\right]}{E(\varepsilon_{fe,G,T}E(D|G,T))} \\
&= \frac{E\left[W_{fe}D\Delta^{TR}_{G,T}\right]}{E(D)} \\
&= E\left[W_{fe}\Delta^{TR}_{G,T}|D=1\right],
\end{aligned}$$

where the second equality follows from the definition of W_{fe}.

Proof for the first-difference regression.

With a slight abuse of notation, let, $E(Y_{G,T})$, $\Delta^{TR}_{G,T}$, and $E(D_{G,T})$ respectively denote $E(Y_{g,t})$, $\Delta^{TR}_{g,t}$, and $E(D_{g,t})$ evaluated at $(g,t)=(G,T)$.

It follows from Frisch-Waugh theorem and the definition of $\varepsilon_{fd,G,T}$ that

$$\beta_{fd} = \frac{\text{cov}(\varepsilon_{fd,G,T}, FD_Y(G,T)|T \geq 1)}{\text{cov}(\varepsilon_{fd,G,T}, FD_D(G,T)|T \geq 1)}. \tag{25}$$

Note that the denominator must be different from 0, otherwise β_{fd} would not be well defined. Now, by definition of $\varepsilon_{fd,G,T}$ again,

$$E[\varepsilon_{fd,G,T}|T] = 0. \tag{26}$$

Then,

$$\begin{aligned}
&\text{cov}(\varepsilon_{fd,G,T}, FD_Y(G,T)|T \geq 1)\\
=&E[\varepsilon_{fd,G,T} DID_Y(G,0,T,T-1)|T \geq 1]\\
=&E\left[\varepsilon_{fd,G,T}\left(E(D_{G,T})\Delta^{TR}_{G,T} - E(D_{G,T-1})\Delta^{TR}_{G,T-1} - E(D_{0,T})\Delta^{TR}_{0,T} + E(D_{0,T-1})\Delta^{TR}_{0,T-1}\right)|T \geq 1\right]\\
=&E\left[\varepsilon_{fd,G,T}\left(E(D_{G,T})\Delta^{TR}_{G,T} - E(D_{G,T-1})\Delta^{TR}_{G,T-1}\right)|T \geq 1\right]\\
=&\sum_{g=0}^{\overline{g}}\sum_{t=1}^{\overline{t}} E(\varepsilon_{fd,G,T}1\{G=g\}1\{T=t\}|T \geq 1)\left(E(D_{g,t})\Delta^{TR}_{g,t} - E(D_{g,t-1})\Delta^{TR}_{g,t-1}\right)\\
=&\sum_{g=0}^{\overline{g}}\sum_{t=0}^{\overline{t}} E\left[\varepsilon_{fd,G,T}1\{G=g\}\left(1\{T=t\} - 1\{T=t+1\}\right)|T \geq 1\right] E(D_{g,t})\Delta^{TR}_{g,t}\\
=&\sum_{g=0}^{\overline{g}}\sum_{t=0}^{\overline{t}} \left(\varepsilon_{fd,g,t}\frac{P(G=g,T=t)}{P(T \geq 1)} - \varepsilon_{fd,g,t+1}\frac{P(G=g,T=t+1)}{P(T \geq 1)}\right) E(D_{g,t})\Delta^{TR}_{g,t}\\
=&\frac{1}{P(T \geq 1)}\sum_{g=0}^{\overline{g}}\sum_{t=0}^{\overline{t}} P(G=g,T=t)\widetilde{w}_{fd,g,t}E(D_{g,t})\Delta^{TR}_{g,t}\\
=&\frac{1}{P(T \geq 1)}E\left[\widetilde{w}_{fd,G,T}D\Delta^{TR}_{G,T}\right].
\end{aligned}$$

Eq. (26) implies that

$$E[\varepsilon_{fd,G,T} - (E(Y_{0,T}) - E(Y_{0,T-1}))|T \geq 1] = 0,$$

hence the first equality. The second equality follows from the first point of Lemma 2. Eq. (26) implies again that

$$E[\varepsilon_{fd,G,T} - (E(D_{0,T})\Delta^{TR}_{0,T} - E(D_{0,T-1})\Delta^{TR}_{0,T-1})|T \geq 1] = 0,$$

hence the third equality. The fifth equality follows from a summation by part, the sixth holds because $\varepsilon_{fd,g,0} = 0$, the seventh follows from the definition of $\tilde{w}_{fd,g,t}$, and the eighth from the law of iterated expectations.

A similar reasoning yields

$$\text{cov}(\varepsilon_{fd,G,T}, FD_D(G,T)|T \geq 1) = \frac{1}{P(T \geq 1)} E\left[\widetilde{w}_{fd,G,T} D\right].$$

The end of the proof follows exactly as for $k = fe$.

Proof of the second statement.

We first prove Equation (5) and two other equalities that we use in the proof. Note first that for all $(g,t) \in \{0, ..., \overline{g}\} \times \{1, ..., \overline{t}\}$, we have

$$\begin{aligned}
|E(D_{g,t}) - E(D_{g,t-1})| =& |P(D(t) = 1|G = g, T = t) - P(D(t-1) = 1|G = g, T = t-1)| \\
=& |P(D(t) = 1|G = g, T = t) - P(D(t-1) = 1|G = g, T = t)| \\
=& P(S|G = g, T = t),
\end{aligned}$$

where the equalities follow from Assumption 2. Then,

$$\begin{aligned}
E[f(G,T)1_S] &= E[f(G,T)P(S|G,T)] \\
&= \sum_{g=0}^{\overline{g}} \sum_{t=1}^{\overline{t}} P(G = g, T = t) \left| E(D_{g,t}) - E(D_{g,t-1}) \right| f(g,t).
\end{aligned} \tag{27}$$

The first equality follows from the law of iterated expectations, and the second follows from the fact that $P(S|G = g, T = 0) = 0$ and from the previous display. Similarly,

$$P(S) = \sum_{g=0}^{\overline{g}} \sum_{t=1}^{\overline{t}} P(G = g, T = t) \left| E(D_{g,t}) - E(D_{g,t-1}) \right|. \tag{28}$$

(5) follows by dividing (27) by (28).

Proof for the fixed-effect regression.

We have

$$
\begin{aligned}
&\text{cov}(\varepsilon_{fe,G,T}, Y) \\
=&E[\varepsilon_{fe,G,T}(E(Y|G,T) - E(Y|G=0,T))] \\
=&E\left(\sum_{t=0}^{\bar{t}} E(\varepsilon_{fe,G,T}1\{T=t\}|G)\left[E(Y|G,T=t) - E(Y|G=0,T=t)\right]\right) \\
=&E\left(\sum_{t=1}^{\bar{t}} \left[\sum_{t'\geq t} E(\varepsilon_{fe,G,T}1\{T=t'\}|G)\right] DID_Y(G,0,t,t-1)\right) \\
=&E\left(\sum_{t=1}^{\bar{t}} E(\varepsilon_{fe,G,T}1\{T\geq t\}|G)\left[FD_D(G,t)\Delta^S_{G,t} - FD_D(0,t)\Delta^S_{0,t}\right]\right) \\
=&E\left(\sum_{t=1}^{\bar{t}} E(\varepsilon_{fe,G,T}1\{T\geq t\}|G)FD_D(G,t)\Delta^S_{G,t}\right) - \sum_{t=1}^{\bar{t}} E(\varepsilon_{fe,G,T}1\{T\geq t\})FD_D(0,t)\Delta^S_{0,t} \\
=&\sum_{g=0}^{\bar{g}}\sum_{t=1}^{\bar{t}} P(G=g,T=t)|FD_D(g,t)|\frac{s_{g,t}E[\varepsilon_{fe,G,T}1\{G=g,T\geq t\}]}{P(G=g,T=t)}\Delta^S_{g,t} \qquad (29) \\
=&E\left[\widetilde{\omega}_{fe,G,T}\Delta^S_{G,T}1_S\right]. \qquad (30)
\end{aligned}
$$

The first equality follows from the law of iterated expectation and (23). The third equality follows from summation by part and (23). The fourth equality follows from the second point of Lemma 2. (29) follows from (23). (30) follows from the definition of $\widetilde{\omega}_{fe,g,t}$ and from (27).

Similarly,

$$
\text{cov}(\varepsilon_{fe,G,T}, E(D|G,T)) = \sum_{g=0}^{\bar{g}}\sum_{t=1}^{\bar{t}} P(G=g,T=t)|FD_D(g,t)|\frac{s_{g,t}E[\varepsilon_{fe,G,T}1\{G=g,T\geq t\}]}{P(G=g,T=t)} \qquad (31)
$$

$$
= E\left[\widetilde{\omega}_{fe,G,T}1_S\right]. \qquad (32)
$$

Combining (21), (30), and (32) yields

$$
\beta_{fe} = \frac{E\left[\widetilde{\omega}_{fe,G,T}\Delta^S_{G,T}1_S\right]}{E\left[\widetilde{\omega}_{fe,G,T}1_S\right]}.
$$

Finally, the result follows from the definition of Ω_{fe}, after dividing the numerator and the denominator in the right-hand side of the previous display by $P(S)$.

Proof for the first-difference regression.

First,

$$\begin{aligned}
\text{cov}(\varepsilon_{fd,G,T}, FD_Y(G,T)|T \geq 1) =& E[\varepsilon_{fd,G,T} DID_Y(G,0,T,T-1)|T \geq 1] \\
=& E[\varepsilon_{fd,G,T} \left(FD_D(G,T)\Delta^S_{G,T} - FD_D(0,T)\Delta^S_{0,T}\right)|T \geq 1] \\
=& E[\varepsilon_{fd,G,T} FD_D(G,T)\Delta^S_{G,T}|T \geq 1] \\
=& E[\varepsilon_{fd,G,T} s_{G,T}|FD_D(G,T)|\Delta^S_{G,T}|T \geq 1] \\
=& \frac{E[\varepsilon_{fd,G,T} s_{G,T}\Delta^S_{G,T} 1_S]}{P(T \geq 1)}.
\end{aligned}$$

The first equality follows from (26). The second equality follows from the second point of Lemma 2. The third equality follows from (26). The fifth equality follows from (27).

Similarly,

$$\text{cov}(\varepsilon_{fd,G,T}, FD_D(G,T)|T \geq 1) = \frac{E[\varepsilon_{fd,G,T} s_{G,T} 1_S]}{P(T \geq 1)}.$$

The result follows combining (25) and the two previous displays.

B.3 Proof of Corollary 2

We prove only the first statement, as the second statement can be proven by following the exact same steps. If the assumptions of the corollary hold and if $\Delta^{TR} = 0$, then

$$\begin{cases} \beta_k &= E\left[W_k \Delta^{TR}_{G,T}|D=1\right], \\ 0 &= E\left[\Delta^{TR}_{G,T}|D=1\right]. \end{cases}$$

These two conditions and the Cauchy-Schwarz inequality imply

$$|\beta_k| = \left|\text{cov}\left(W_k, \Delta^{TR}_{G,T}\middle| D=1\right)\right| \leq V\left(W_k|D=1\right)^{1/2} V\left(\Delta^{TR}_{G,T}|D=1\right)^{1/2}.$$

Hence, $\sigma^{TR} \geq \underline{\sigma}^{TR}_k$.

Now, we prove that we can rationalize this lower bound. Let us define

$$\Delta^{TR}_{G,T} = \frac{\beta_k\left(W_k - 1\right)}{V\left(W_k|D=1\right)}.$$

Then

$$\Delta^{TR} = E\left[\Delta^{TR}_{G,T}|D=1\right] = \frac{\beta_k}{V\left(W_k|D=1\right)} E\left[W_k - 1|D=1\right] = 0.$$

Similarly,

$$\begin{aligned}
E[W_k \Delta^{TR}_{G,T}|D=1] &= \frac{\beta_k}{V\left(W_k|D=1\right)} E\left[W_k^2 - W_k|D=1\right] \\
&= \frac{\beta_k}{V\left(W_k|D=1\right)} V\left(W_k|D=1\right) \\
&= \beta_k.
\end{aligned}$$

This proves the result.

B.4 Proof of Proposition 1

Proof of the first statement.

Let $I_{g,t} = 1\{E(D_{g,t}) \neq 0\}$. We reason by contradiction, by showing that if one of the two conditions does not hold and if

$$w_{k,g,t}I_{g,t} = I_{g,t} \; \forall(g,t), \tag{33}$$

which is equivalent to $V(W_k|D=1)=0$, then the denominator $\text{Den}(w_k)$ of $w_{k,g,t}$ is equal to 0. This contradicts the fact that β_k is well-defined.

When $k = fe$, (33) implies that for all (g,t), there exists a constant C such that $\varepsilon_{fe,g,t}I_{g,t} = CI_{g,t}$. First suppose that there exists g_0 such that for all t, $I_{g_0,t} = 1$. Then

$$\begin{aligned}
C &= E\left[CI_{g_0,T}|G=g_0\right] \\
&= E\left[\varepsilon_{fe,g_0,T}I_{g_0,T}|G=g_0\right] \\
&= CE\left[\varepsilon_{fe,G,T}|G=g_0\right] \\
&= 0,
\end{aligned}$$

where the last equality follows by (23). Alternatively, suppose that there exists t_0 such that for all g, $I_{g,t_0} = 1$. By taking expectation over G instead of T and using (23) again, we also obtain $C = 0$. Then, in both cases,

$$\begin{aligned}
\text{Den}(w_k) &= E[E(D|G,T)\varepsilon_{fe,G,T}] \\
&= E[E(D|G,T)I_{G,T}\varepsilon_{fe,G,T}] \\
&= E[E(D|G,T)CI_{G,T}] \\
&= 0.
\end{aligned}$$

When $k = fd$, (33) implies that for all (g,t), there exists a constant C such that $\widetilde{w}_{fd,g,t}I_{g,t} = CI_{g,t}$. Then, if there exists g_0 such that for all t, $I_{g_0,t} = 1$, we get

$$\begin{aligned}
C &= E\left[CI_{g_0,T}|G=g_0\right] \\
&= E\left[\widetilde{w}_{fd,g_0,T}|G=g_0\right] \\
&= \frac{1}{P(G=g_0)}\sum_{t=0}^{\bar{t}}(P(G=g_0,T=t)\varepsilon_{fd,g_0,t} - P(G=g_0,T=t+1)\varepsilon_{fd,g_0,t+1}) \\
&= \frac{1}{P(G=g_0)}(P(G=g_0,T=\bar{t}+1)\varepsilon_{fd,g_0,0} - P(G=g_0,T=0)\varepsilon_{fd,g_0,\bar{t}+1}) \\
&= 0.
\end{aligned}$$

If there exists t_0 such that for all g, $I_{g,t_0} = 1$, we obtain similarly

$$\begin{aligned}
C &= E\left[CI_{G,t_0}|T=t_0\right] \\
&= E\left[\widetilde{w}_{fd,G,t_0}|T=t_0\right] \\
&= \frac{1}{P(T=t_0)}\sum_{g=0}^{\bar{g}}(P(G=g,T=t_0)\varepsilon_{fd,g,t_0} - P(G=g,T=t_0+1)\varepsilon_{fd,g,t_0+1}) \\
&= E\left[\varepsilon_{fd,G,T}|T=t_0\right] - \frac{P(T=t_0+1)}{P(T=t_0)}E\left[\varepsilon_{fd,G,T}|T=t_0+1\right] \\
&= 0,
\end{aligned}$$

where the last equality follows from (26). In both cases, one can then use the same argument as for $k = fe$ to show that the denominator of $w_{fd,g,t}$ is equal to 0.

Proof of the second statement.

Let $I'_{g,t} = 1\{E(D_{g,t}) \neq E(D_{g,t-1}\}$, so that $P(S|G=g,T=t) > 0$ if and only if $I'_{g,t} = 1$. As above, we reason by contradiction by showing that if the one of the conditions does not hold and if

$$\omega_{k,g,t}I'_{g,t} = I'_{g,t}\ \forall(g,t), \tag{34}$$

which is equivalent to $V(\Omega_k|S) = 0$, then the denominator $\text{Den}(\omega_k)$ of $\omega_{k,g,t}$ is equal to 0.

When $k = fe$, (34) together with the definition of $\omega_{fe,g,t}$ and the equalities $s_{g,t}I'_{g,t} = s_{g,t}$ and $s^2_{g,t}I'_{g,t} = I'_{g,t}$ imply that for some constant C and all (g,t),

$$E(\varepsilon_{fe,G,T}1\{G=g,T\geq t\})I'_{g,t} = CP(G=g,T=t)s_{g,t}.$$

Suppose that there exists t_0 such that $g \mapsto s_{g,t_0} = s$, with $s \in \{-1,1\}$. Then

$$\begin{aligned}
CP(T=t_0)s &= \sum_{g=0}^{\bar{g}} CP(G=g,T=t_0)s_{g,t_0} \\
&= \sum_{g=0}^{\bar{g}} E(\varepsilon_{fe,G,T}1\{G=g,T\geq t_0\}) \\
&= E(\varepsilon_{fe,G,T}1\{T\geq t_0\}) \\
&= 0
\end{aligned}$$

where the last equality follows by (23). One can then use the same reasoning as that used in the proof of the first statement to show that the previous equality implies that $\text{Den}(\omega_{fe}) = 0$.

When $k = fd$, (34) implies that for some C and all (g,t), $\varepsilon_{fd,G,T}I'_{G,T} = Cs_{G,T}$. Suppose that there exists t_0 such that $g \mapsto s_{g,t_0} = s$, where $s \in \{-1,1\}$. Then $Cs = E(\varepsilon_{fd,G,T}|T=t_0)$, which yields $C = 0$ by (26). Again, this implies that $\text{Den}(\omega_{fd}) = 0$.

B.5 Proof of Proposition 2

1. We have

$$P(W_{fe}\geq 0|D=1)=\sum_{g=0}^{\overline{g}}\sum_{t=0}^{\overline{t}}1\{w_{fe,g,t}\geq 0\}P(G=g,T=t|D=1).$$

Therefore, $P(W_{fe}\geq 0|D=1)=1$ is equivalent to having that $w_{fe,g,t}P(D=1|G=g,T=t)\geq 0$ for all $(g,t)\in\{0,...,\overline{g}\}\times\{0,...,\overline{t}\}$. A few lines of algebra show that $E(\varepsilon_{fe,G,T}E(D|G,T))=V(\varepsilon_{fe,G,T})$, which is strictly positive if β_{fe} is well-defined. Then, to study the sign of $w_{fe,g,t}P(D=1|G=g,T=t)$, it is sufficient to study the sign of $\varepsilon_{fe,g,t}1\{E(D_{g,t})>0\}$. Under Assumption 7, $G\perp\!\!\!\perp T$. If $G\perp\!\!\!\perp T$, one can show that

$$\varepsilon_{fe,g,t}=E(D_{g,t})-E(D_{g,.})-E(D_{.,t})+E(D).$$

Then, Assumptions 7 and 6 imply that for all $(g,t)\in\{0,...,\overline{g}\}\times\{0,...,\overline{t}\}$,

$$\begin{aligned}&\varepsilon_{fe,g,t}1\{E(D_{g,t})>0\}\\=&1\{t\geq a_g\}\left(1-\frac{\overline{t}+1-a_g}{\overline{t}+1}-\frac{1}{\overline{g}+1}\sum_{g'=0}^{\overline{g}}1\{a_{g'}\leq t\}+\frac{1}{\overline{t}+1}\sum_{t'=0}^{\overline{t}}\frac{1}{\overline{g}+1}\sum_{g'=0}^{\overline{g}}1\{a_{g'}\leq t'\}\right).\end{aligned}$$

$\varepsilon_{fe,g,t}1\{E(D_{g,t})>0\}=0$ for all (g,t) such that $t<a_g$. For all (g,t) such that $t\geq a_g$, $t\mapsto\varepsilon_{fe,g,t}1\{E(D_{g,t})>0\}$ is decreasing in t. Moreover, $g\mapsto\varepsilon_{fe,g,\overline{t}}$ is minimized at $g=g(e)$. Therefore, $\varepsilon_{fe,g,t}1\{E(D_{g,t})>0\}$ is strictly negative for some (g,t) if and only if

$$\frac{a_{g(e)}}{\overline{t}+1}-\frac{1}{\overline{g}+1}\sum_{g'=0}^{\overline{g}}1\{a_{g'}\leq\overline{t}\}+\frac{1}{\overline{t}+1}\sum_{t'=0}^{\overline{t}}\frac{1}{\overline{g}+1}\sum_{g'=0}^{\overline{g}}1\{a_{g'}\leq t'\}<0.$$

The proof is similar for Ω_{fe} and is therefore omitted.

2. Here as well, $P(W_{fd}\geq 0|D=1)=1$ is equivalent to having that $w_{fd,g,t}P(D=1|G=g,T=t)\geq 0$ for all $(g,t)\in\{0,...,\overline{g}\}\times\{0,...,\overline{t}\}$. Moreover, under the assumptions of the theorem, to study the sign of $w_{fd,g,t}P(D=1|G=g,T=t)$, it is sufficient to study the sign of $v_{g,t}=\tilde{w}_{fd,g,t}1\{E(D_{g,t})>0\}$. For all $(g,t)\in\{0,...,\overline{g}\}\times\{0,...,\overline{t}\}$,

$$\begin{aligned}&v_{g,t}\\=&1\{t\geq a_g\}(\varepsilon_{fd,g,t}-\varepsilon_{fd,g,t+1})\\=&1\{t\geq a_g\}\left[1\{t\geq 1\}(E(D_{g,t})-E(D_{g,t-1})-E(D_{.,t})+E(D_{.,t-1}))\right.\\&\left.-1\{t\leq\overline{t}-1\}(E(D_{g,t+1})-E(D_{g,t})-E(D_{.,t+1})+E(D_{.,t}))\right]\\=&1\{t\geq a_g\}\left[1\{t\geq 1\}1\{t=a_g\}-\frac{1\{t\geq 1\}}{\overline{g}+1}\sum_{g'=0}^{\overline{g}}1\{a_{g'}=t\}+\frac{1\{t\leq\overline{t}-1\}}{\overline{g}+1}\sum_{g'=0}^{\overline{g}}1\{a_{g'}=t+1\}\right].\end{aligned}$$

The first equality follows from the definition of $\tilde{w}_{fd,g,t}$, and from Assumptions 7 and 6. The second equality follows from the fact that for all $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}+1\}$,

$$\varepsilon_{fd,g,t} = 1\{1 \leq t \leq \overline{t}\}(E(D_{g,t}) - E(D_{g,t-1}) - E(D_{.,t}) + E(D_{.,t-1})).$$

The third equality follows from Assumptions 7 and 6.

If $a_g \geq \overline{t}$ for every $g \in \{0,...,\overline{g}\}$, it follows from the previous display that $v_{g,t} \geq 0$ for every $(g,t) \in \{0,...,\overline{g}\} \times \{0,...,\overline{t}\}$. Conversely, assume that at least one group adopts before $\overline{t}$. Then, $a_{g(e)} = t_0 < \overline{t}$. Now, we reason by contradiction. Assume that $v_{g(e),t_0+1},...,v_{g(e),\overline{t}} \geq 0$. $v_{g(e),\overline{t}} \geq 0$ implies $\frac{1}{\overline{g}+1}\sum_{g'=0}^{\overline{g}} 1\{a_{g'} = \overline{t}\} = 0$. Then, $v_{g(e),\overline{t}-1} \geq 0$ implies $\frac{1}{\overline{g}+1}\sum_{g'=0}^{\overline{g}} 1\{a_{g'} = \overline{t}-1\} = 0$. And so on and so forth. Finally, $v_{g(e),t_0+1} \geq 0$ implies $\frac{1}{\overline{g}+1}\sum_{g'=0}^{\overline{g}} 1\{a_{g'} = t_0+1\} = 0$. Therefore, all groups must have $a_g = a_{g(e)}$. But then, $E(D_{g,t}) - E(D_{g,t-1}) = 1\{t = a_{g(e)}\}$ if $a_{g(e)} \geq 1$, and $E(D_{g,t}) - E(D_{g,t-1}) = 0$ otherwise. This contradicts the fact that β_{fd} is well-defined. Therefore, at least one of the $(v_{g(e),t_0+1},...,v_{g(e),\overline{t}})$ must be strictly negative.

3. Here as well, $P(\Omega_{fd} \geq 0|S) = 1$ is equivalent to having that $\omega_{fd,g,t}P(S|G=g,T=t) \geq 0$ for all $(g,t) \in \{0,...,\overline{g}\} \times \{1,...,\overline{t}\}$. Moreover, to study the sign of $\omega_{fd,g,t}P(S|G=g,T=t)$, it is sufficient to study the sign of $s_{g,t}\varepsilon_{fd,g,t}P(S|G=g,T=t)$. For all $(g,t) \in \{0,...,\overline{g}\} \times \{1,...,\overline{t}\}$,

$$\begin{aligned}
&s_{g,t}\varepsilon_{fd,g,t} \\
=&s_{g,t}\left[E(D_{g,t}) - E(D_{g,t-1}) - E(D_{.,t}) + E(D_{.,t-1})\right] \\
=&1\{t = a_g\}\left[1 - E(D_{.,t}) + E(D_{.,t-1})\right] \geq 0.
\end{aligned}$$

The second equality follows from Assumption 6. This proves the result.

B.6 Proof of Proposition 3

First, notice that under Assumption 8, $S = \{D = 1\}$ and $s_{g,1} = 1$ for all g. Therefore, for all $(k,g) \in \{1,2,3\} \times \{0,...,\overline{g}\}$,

$$w_{k,g,1} = \omega_{k,g,1} = \frac{\varepsilon_{k,g,1}E(D)}{E(D\varepsilon_{k,G,T})}.$$

Moreover,

$$P(\Omega_k < 0|S) = P(W_k < 0|D=1,T=1) = P(W_k < 0|D=1).$$

Let us reason by contradiction and suppose that $P(W_k < 0|D=1,T=1) = 0$ is equal to zero. Let us also suppose that $E(D\varepsilon_{k,G,T}) > 0$ (the proof is symmetric if instead $E(D\varepsilon_{k,G,T}) < 0$).

Then:

$$
\begin{aligned}
0 &= P(W_k < 0|D = 1, T = 1) \\
&= P(\varepsilon_{k,G,1} < 0|D = 1, T = 1) \\
&= \frac{E\left[D1\{\varepsilon_{k,G,1} < 0\}|T = 1\right]}{E(D_{.,1})} \\
&= \frac{E\left[E(D|G, T = 1)1\{\varepsilon_{k,G,1} < 0\}|T = 1\right]}{E(D_{.,1})} \\
&\geq \frac{\left[\min_g E(D_{g,1})\right] P(\varepsilon_{k,G,1} < 0|T = 1)}{E(D_{.,1})}.
\end{aligned}
$$

The second equality follows from the fact that if $E(D\varepsilon_{k,G,T}) > 0$, W_k is equivalent to $1\{\varepsilon_{k,G,1} < 0\}$ conditional on $T = 1$. The fourth equality follows from the law of iterated expectations. The fifth equality follows from the fact that $E(D|G, T = 1) \geq \min_g E(D_{g,1})$.

The previous display implies that $P(\varepsilon_{k,G,1} < 0|T = 1) = 0$. By definition of $\varepsilon_{k,g,t}$, $E(\varepsilon_{k,G,1}|T = 1) = 0$. Hence, $P(\varepsilon_{k,G,1} = 0|T = 1) = 1$. But then

$$
E(D\varepsilon_{k,G,T}) = E(D\varepsilon_{k,G,1}1\{T = 1\}) = 0.
$$

This implies that β_{fe} is not well defined, a contradiction.

B.7 Proof of Theorem 2

The reasoning is exactly the same as in Theorem 1, except that we rely on Lemma 1 instead of Lemma 2.

B.8 Proof of Theorem 3

Proof of the first statement for β_{fe}.

First, remark that β^X_{fe} is the coefficient of $E(D|G, T)$ in the regression of $Y - X'\gamma_{fe}$ on group and time fixed effects and $E(D|G, T)$. Therefore, by the Frisch-Waugh theorem,

$$
\beta^X_{fe} = \frac{\text{cov}(\varepsilon_{fe,G,T}, Y - X'\gamma_{fe})}{\text{cov}(\varepsilon_{fe,G,T}, E(D|G, T))}.
$$

Then, reasoning as in the proof of Theorem 1, we obtain

$$
\text{cov}(\varepsilon_{fe,G,T}, Y - X'\gamma_{fe}) = E\left[\varepsilon_{fe,G,T} DID_{Y-X'\gamma_{fe}}(G, 0, T, 0)\right].
$$

Now, under Assumptions 1X$_{fe}$ and 9, we can follow the same steps as those used to establish the first point of Lemma 2 to show that

$$
DID_{Y-X'\gamma_{fe}}(g, 0, t, 0) = E(D_{g,t})\Delta^{TR}_{g,t} - E(D_{g,0})\Delta^{TR}_{g,0} - \left(E(D_{0,t})\Delta^{TR}_{0,t} - E(D_{0,0})\Delta^{TR}_{0,0}\right).
$$

Then, the proof follows exactly as that of the first statement of Theorem 1 for β_{fe}.

Proof of the first statement for β_{fd}.

First, remark that β_{fd}^X is the coefficient of $FD_D(G,T)$ in the regression of $FD_{Y-X'\gamma_{fd}}(G,T)$ on time fixed effects and $FD_D(G,T)$. Therefore, by Frisch-Waugh theorem,

$$\beta_{fd}^X = \frac{\text{cov}(\varepsilon_{fd,G,T}, FD_{Y-X'\gamma_{fd}}(G,T)|T \geq 1)}{\text{cov}(\varepsilon_{fd,G,T}, FD_D(G,T)|T \geq 1)}.$$

Then, reasoning as in the proof of Theorem 1, we obtain

$$\text{cov}(\varepsilon_{fd,G,T}, FD_{Y-X'\gamma_{fd}}(G,T)|T \geq 1) = E\left[\varepsilon_{fd,G,T} DID_{Y-X'\gamma_{fd}}(G,0,T,T-1)|T \geq 1\right].$$

Now, as above with $k = fe$,

$$DID_{Y-X'\gamma_{fd}}(g,0,t,t-1) = E(D_{g,t})\Delta_{g,t}^{TR} - E(D_{g,t-1})\Delta_{g,t-1}^{TR} - \left(E(D_{0,t})\Delta_{0,t}^{TR} - E(D_{0,t-1})\Delta_{0,t-1}^{TR}\right).$$

Then, the proof follows exactly as that of the first statement of Theorem 1 for β_{fd}.

Proof of the second statement.

We only sketch the proof of the result for β_{fe}. As above, the reasoning is similar to that in the proof of the second statement of Theorem 1, but replacing Y by $Y - X'\gamma_{fe}$. In particular, under Assumptions $1X_{fe}$, 3, and 9, one can follow the same steps as those used to establish the second point of Lemma 2 to show that

$$DID_{Y-X'\gamma_{fe}}(g,0,t,t-1) = \left(E(D_{g,t}) - E(D_{g,t-1})\right)\Delta_{g,t}^S - \left(E(D_{0,t}) - E(D_{0,t-1})\right)\Delta_{0,t}^S.$$

B.9 Proof of Theorem 4

Proof of the first statement.

For all $(g^*,t) \in \{-1,0,1\} \times \{1,...,\bar{t}\}$, let $\mathcal{G}_{g^*,t} = \{g : \text{sgn}\left(FD_D(g,t)\right) = g^*\}$. First, note that for all $t \geq 1$,

$$E(Y|G_t^* = 1, T = t) = \sum_{g \in \mathcal{G}_{1,t}} r(g|1,t,t)E(Y_{g,t}).$$

Similarly,

$$\begin{aligned} E(QY|G_t^* = 1, T = t-1) &= \sum_{g \in \mathcal{G}_{1,t}} r(g|1,t,t-1)\frac{r(g|1,t,t)}{r(g|1,t,t-1)}E(Y_{g,t-1}) \\ &= \sum_{g \in \mathcal{G}_{1,t}} r(g|1,t,t)E(Y_{g,t-1}). \end{aligned}$$

Hence,

$$\begin{aligned}&DID_Y^*(1,t)\\&=\sum_{g\in\mathcal{G}_{1,t}} r(g|1,t,t)\left[E(Y_{g,t})-E(Y_{g,t-1})-(E(Y|G_t^*=0,T=t)-E(QY|G_t^*=0,T=t-1))\right].\end{aligned}$$

Moreover, reasoning as above,

$$E(Y|G_t^*=0,T=t)-E(QY|G_t^*=0,T=t-1)=\sum_{g'\in\mathcal{G}_{0,t}} r(g'|0,t,t)\left(E(Y_{g',t})-E(Y_{g',t-1})\right).$$

Thus,

$$DID_Y^*(1,t)=\sum_{(g,g')\in\mathcal{G}_{1,t}\times\mathcal{G}_{0,t}} r(g|1,t,t)r(g'|0,t,t)DID_Y(g,g',t,t-1), \tag{35}$$

where $DID(g,g',t,t-1)$ is defined as above in Lemma 1. By definition of $\mathcal{G}_{0,t}$, we have that under Assumption 2, for all $g'\in\mathcal{G}_{0,t}$, $P(S_{g',t})=0$. Then, by Lemma 2, we have, for all $(g,g')\in\mathcal{G}_{1,t}\times\mathcal{G}_{0,t}$,

$$DID_Y(g,g',t,t-1)=P(S_{g,t})\Delta^S_{g,t}.$$

Combining this equation with (35), we obtain

$$DID_Y^*(1,t)=\sum_{g\in\mathcal{G}_{1,t}} r(g|1,t,t)P(S_{g,t})\Delta^S_{g,t}.$$

Then, because $P(G_t^*=1,T=t)r(g|1,t,t)=P(G=g,T=t)$ for all $g\in\mathcal{G}_{1,t}$,

$$w_{fe,t}\frac{DID_Y^*(1,t)}{DID_D^*(1,t)}=\frac{\sum_{g\in\mathcal{G}_{1,t}}P(G=g,T=t)P(S_{g,t})\Delta^S_{g,t}}{\sum_{t'=1}^{\bar{t}}\sum_{g'\in\{-1,1\}}g'DID_D^*(g',t')P(G_t^*=g',T=t')}.$$

Reasoning as above, we obtain

$$\begin{aligned}DID_Y^*(-1,t)&=-\sum_{g\in\mathcal{G}_{-1,t}} r(g|-1,t,t)P(S_{g,t})\Delta^S_{g,t},\\DID_D^*(1,t)&=\sum_{g\in\mathcal{G}_{1,t}} r(g|1,t,t)P(S_{g,t}),\\DID_D^*(-1,t)&=-\sum_{g\in\mathcal{G}_{-1,t}} r(g|-1,t,t)P(S_{g,t}).\end{aligned}$$

Combining all these and the fact that $P(S_{g,t})=0$ for all $g\in\mathcal{G}_{0,t}$ yields

$$\sum_{g^*\in\{-1,1\}} w_{g^*,t}\frac{DID_Y^*(g^*,t)}{DID_D^*(g^*,t)}=\frac{\sum_{g=0}^{\overline{g}}P(G=g,T=t)P(S_{g,t})\Delta^S_{g,t}}{\sum_{t'=1}^{\bar{t}}\sum_{g'=0}^{\overline{g}}P(G=g',T=t)P(S_{g',t})}.$$

The result follows by summing over $t\in\{1,...,\overline{t}\}$ and by definition of Δ^S.

Proof of the second statement.

First, reasoning as above,

$$E(Y|G^*_t = 1, T = t) - E(QY|G^*_t = 1, T = t-1) = \sum_{g\in\mathcal{G}_{1,t}} r(g|1,t,t)[E(Y_{g,t}) - E(Y_{g,t-1})] \quad (36)$$

Moreover,

$$\begin{aligned}
&E(Y_{g,t}) - E(Y_{g,t-1})\\
=&E(Y_{g,t}(1) - Y_{g,t}(0)|S)P(S_{g,t}) + E(Y_{g,t}(1)|D(t-1) = 1)P(D(t-1) = 1|G = g, T = t)\\
-&E(Y_{g,t-1}(1)|D(t-1) = 1)P(D(t-1) = 1|G = g, T = t-1)\\
+&E(Y_{g,t}(0)|D(t-1) = 0)P(D(t-1) = 0|G = g, T = t)\\
-&E(Y_{g,t-1}(0)|D(t-1) = 0)P(D(t-1) = 0|G = g, T = t-1)\\
=&\Delta^S_{g,t}P(S_{g,t}) + E(Y_{g,t}(1) - Y_{g,t-1}(1)|D(t-1) = 1)E(D_{g,t-1})\\
&+ E(Y_{g,t}(0) - Y_{g,t-1}(0)|D(t-1) = 0)(1 - E(D_{g,t-1})). \qquad (37)
\end{aligned}$$

where the first equality follows from Assumption 2.3 and the second from Assumption 2.1 and 2.2. Now, by a similar reasoning as that used to obtain (36),

$$\begin{aligned}
E(Q\delta_D|G^*_t = 1, T = t-1) &= E(Q(D\delta_1 + (1-D)\delta_0|G^*_t = 1, T = t-1)\\
&= \sum_{g\in\mathcal{G}_{1,t}} r(g|1,t,t)[E(D_{g,t-1})\delta_1 + (1 - E(D_{g,t-1}))\delta_0]. \qquad (38)
\end{aligned}$$

Next, we have, for any $g \in \mathcal{G}_{1,t}$,

$$\begin{aligned}
\delta_d &= E(Y|D = d, G^*_t = 0, T = t) - E(Q_dY|D = d, G^*_t = 0, T = t-1)\\
&= \sum_{g'\in\mathcal{G}_{0,t}} r_d(g'|t,t)E(Y_{g',t}(d) - Y_{g',t-1}(d)|D(t-1) = d)\\
&= \sum_{g'\in\mathcal{G}_{0,t}} r_d(g'|t,t)E(Y_{g,t}(d) - Y_{g,t-1}(d)|D(t-1) = d)\\
&= E(Y_{g,t}(d) - Y_{g,t-1}(d)|D(t-1) = d).
\end{aligned}$$

The first equality follows from a similar reasoning as that used to obtain (36). The third follows from Assumption 1'.

Therefore, in view of (36), (37) and (38), we obtain

$$\begin{aligned}
&E(Y|G^*_t = 1, T = t) - E(QY|G^*_t = 1, T = t-1)\\
=&\sum_{g\in\mathcal{G}_{1,t}} r(g|1,t,t)\Delta^S_{g,t}P(S_{g,t}) + E(Q\delta_D|G^*_t = 1, T = t-1).
\end{aligned}$$

Rearranging the previous display and using the definition of $TCD^*(1,t)$ yields

$$TCD^*(1,t) = \sum_{g \in \mathcal{G}_{1,t}} r(g|1,t,t)P(S_{g,t})\Delta^S_{g,t}.$$

A similar reasoning yields

$$TCD^*(-1,t) = -\sum_{g \in \mathcal{G}_{-1,t}} r(g|-1,t,t)P(S_{g,t})\Delta^S_{g,t}.$$

The rest of the proof is the same as that of the first statement.