Proportional Treatment Effects in Staggered Settings: An Approach for Poisson.*

Ninon Moreau-Kastler†

January 22, 2025

Abstract

I propose an approach to measure proportional treatment effects for multiplicative difference-in-differences models. TWFE (two-way fixed effect) linear estimators do not recover difference-in-differences estimates in the presence of staggered treatment. I show that this issue extends to Poisson-Pseudo Maximum Likelihood estimators. In the linear case, robust estimators exist to recover correct DiD estimates, but these approaches do not extend to PPML, as aggregation is challenging in the non-linear case. This paper develops an estimator robust to TWFE staggered bias for PPML, which recovers a quantity with a similar interpretation as in the canonical 2-by-2 model: the Ratio-of-Ratios.

*I thank Kirill Borusyak, Peter Egger, Martin Mugnier, Farid Toubal and Morgan Ubeda for helpful comments and advices.

†EU Tax Observatory, Paris School of Economics.

1 Introduction

Applied economists are often interested in studying variables which take only positive values and are non-normally distributed. Such outcomes can be trade flows, sales or employment for example. Public policies or economic shocks generate changes in these outcomes in magnitudes that will often vary across small or large countries, firms or sectors. In such cases, researchers are interested in proportional treatment effects, or semi-elasticities: the change in the outcome in percentage generated by the treatment. Common practice among empirical researchers has been to use log transformations of the outcome, mixed with a linear two-way fixed effect model when selection into treatment is non-random: a method I refer to as TWFE log-OLS.

The TWFE PPML estimator presents several advantages over TWFE log-OLS. It can include observations with zero in the outcome, and can easily estimate high-dimensional fixed effects models. It is not biased in settings where the treatment changes the level of the outcome and the variance of the error term, contrary to log-OLS. In settings where treatment effects are heterogeneous across units, TWFE log-OLS and TWFE PPML recover different quantities of interest and imply different parallel trends. The TWFE PPML estimator targets the percentage change in the outcome, and relies on the assumption that the growth rate in the outcome of the two groups should have been the same without treatment.[1]

On the other hand, in the presence of treatment heterogeneity and staggered treatment, empirical researchers have been recently concerned that two-way fixed effects estimators do not recover desired aggregate difference-in-differences quantity of interest. The estimator recovers "forbiden comparisons" and weights negatively some treatment effects, potentially yielding estimates of the wrong sign.[2] In this paper, I show that the same issue plagues the TWFE PPML estimator. Using a simple example with two indivuals treated at different times, I show that the estimated quantity differs significantly from the estimation target when there are heterogeneous treatment effects by time and individuals.

Robust estimators have been developed in the linear case, recovering correct DiD estimates for cohort and time cells and aggregating them correctly (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Wooldridge, 2021). They are challenging for non-linear estimators such as PPML: they rest on averaging correct linear treatment effects, which cannot be extended to non-

[1] TWFE log-OLS captures changes in the average treatment parameter.

[2] See De Chaisemartin and D'Haultfoeuille (2023) for a review of this literature.

linear treatment effects. In particular, the multiplicative difference-in-differences model from PPML targets the proportional change in the average, which cannot be recovered by a weighted sum of proportional changes on cohorts g at time t:

$$\sum_{i,t} \nu_{g,t} \frac{E[y_{i(g,t)}(1)|D=1] - E[y_{i(g,t)}(0)|D=1]}{E[y_{i(g,t)}(0)|D=1]} \neq \frac{E[y_{it}(1)|D=1] - E[y_{it}(0)|D=1]}{E[y_{it}(0)|D=1]} \quad (1)$$

This paper develops an estimator which recovers a proportional treatment effect (semi-elasticity), corresponding to the correct mutiplicative difference-in-differences (ratio-of-ratios), even when treatment is staggered and its effect heterogeneous. This estimator rests on the idea that by analogy with the linear case, the TWFE PPML estimator recovers the ratio-of-ratios in the canonical 2 times and 2 groups setting (Ciani and Fisher, 2019).

Using a parallel trend in growth rates, a counterfactual outcome can be estimated by multiplying the pre-treatment outcome of the treated group by the growth rate of the control group's outcome. The estimator then recovers a correct average treatment effect in level, scaled by the counterfactual average outcome of the treated group. The estimated quantity corresponds to the growth rate of the average outcome caused by the treatment, or treatment effect semi-elasticity, which is exactly the quantity estimated by TWFE PPML in the canonical setting. I show that this estimator can be computed using either a fully saturated model or an imputation estimator (Wooldridge, 2023; Borusyak et al., 2024).

This paper closely relates to two other attempts to reconcile staggered setting best pratices in the non linear case. Wooldridge (2023) extend the idea that allowing for all margins of treatment heterogeneity allows the TWFE estimator to recover the correct treatment effect estimates, even for nonlinear difference-in-differences. He further provides evidence that a fully saturated model is equivalent to an imputation estimator, and that it allows to easily estimate treatment effects for non-linear models. However, his proposed estimator does not allow to recover a higher level proportional treatment effect than for cohort-time cells. In this paper, I clearly state the quantity of interest of the canonical set-up, and provides a realiable methodology to recover this quantity in the staggered setting, using some results from Wooldridge (2023). My estimator is suited to recover a proportional treatment effect (semi-elmasticity) at any aggregation scale. I also show that the TWFE PPML estimator is biased in the staggered case, with some treatment effect contributing negatively to the estimate.

Nagengast and Yotov (2023) apply Wooldridge (2023)'s fully saturated model to estimate tariff trade creation effects in the gravity setting. They propose a quantity aggregating propor-

tional individual treatment effects from this model interpretable as a proportional treatment effects. In a setting were treatment effect are small, and treatment heterogeneity occurs only accross cohorts and time, their estimator approximates well the average treatment effect, close to the quantity recovered by the log-OLS estimator. However, in a more general case, the estimated quantity is closer to the average model parameter, which differs from the average multiplicative treatment effect in a nonlinear setting. It further breaks the equivalence relationship between the TWFE PPML estimator and the ratio-of-ratios in the canonical 2x2 setting. In this paper, I propose a non-linear estimator, allowing for any general form of treatment heterogeneity, that yields an estimate interpretable as a semi-elasticity, and derived from the ratio-of-ratio estimator.

In section 2, I present a setting in which the researcher has an incentive to recover a treatment effect semi-elasticity. I present a data generating process suited to use TWFE PPML estimators. In section 3, I present the 2x2 canonical setting with two time periods and two groups. I present the quantity of interest for the researcher: the change in percentage in the outcome induced by the treatment. I discuss required assumptions for identification of this quantity using sample moments, and the results from Ciani and Fisher (2019) that the TWFE PPML estimator recovers the ratio-of-ratios. I then present the drawbacks from using the TWFE log-OLS in this setting.

In section 4, I move to the multiperiod setting and heterogeneous treatment timing case. I discuss potential estimates to recover correct multiplicative difference-in-differences (ratio-of-ratios) estimates. I show that estimators aggregating treatment effects estimated separately for each cohort-time cell, such as what is done in the linear case, are not always well suited for the non-linear models. I provide an estimator recovering the correct ratio-of-ratios analogous to the canonical setting. I show that this estimator can be recovered etiher through an imputation or fully saturated model procedure.

I compare my estimator to the true quantity of interest against alternative estimators in section 5. I confirm the result from Ciani and Fisher (2019) that the TWFE PPML recovers the ratio-of-ratios and that the TWFE log-OLS estimator recovers the average parameter in the canonical setting. In staggered treatment timing case, I show that TWFE PPML is biased from the true quantity of interest, but that TWFE log-OLS displays an even bigger bias. I show that the proposed estimator of this paper correctly estimates the true ratio-of-ratios, even when treatment both is heterogeneous accross time and individuals, and induces a change in the outcome variance. In contrast, Nagengast and Yotov (2023) estimator recovers the average model parameter,

but only when there is no individual heterogeneity within treated cohorts.

I study the effect of information-exchange-on-request on bank deposits held in tax havens (Johannesen and Zucman, 2014; Menkhoff and Miethe, 2019) as a real set-up to assess the performance of my estimator. Papers in Public Economics tests wether treaties of automatic exchange of information regarding bank account owners decreased cross-border deposits owned in tax havens. The setting motivates the use of a nonlinear estimator and the estimation of a treatment effect semi-elasticity. First, the outcome under study strongly motivates the use of the PPML estimator. Bilateral deposits are censored to positive values only. Treatment is likely to generate a change in the outcome level (total deposits) and the variance of the outcome (for example if bigger tax havens tend to sign bilateral exchanges of information). Country pairs display very different baseline cross-border owned deposits, which motivates the researcher's interest for a treatment effect expressed as a semi-elasticity. Second, treatment (treaties passed) is staggered and likely to be heterogeneous by time and country-pairs, providing the ideal setting to test for robustness for recent bias of TWFE estimators. I find that the author's estimate have a small positive staggered treatment bias. However, the treated cohort display very large treatment effect heterogeneity, which causes the difference-in-difference estimates of the log-linearized model (log-OLS) to differ by a lot from the ratio-of-ratio (PPML) estimates. More precisely, even though treaties tend to cause a large a negative effect on tax havens deposit *on average*, their effect on the average volume of deposits held offshore is weaker as some large country-pairs react positively or do not react by much. I show that in this case, the proposed imputation estimator of this paper recovers a similar quantity to the TWFE PPML estimator, while the aggregation estimator from Nagengast and Yotov (2023) recovers an intermediate quantity between log-OLS and PPML, with an interpretation depending on the time structure of the treatment.

This paper relates to several trends of the literature in applied econometrics. It relates first to a literature motivating the use of PPML estimators for multiplicative model estimation (Silva and Tenreyro, 2006; Cohn et al., 2022; Chen and Roth, 2023), which can account for observations with zeros in the dependant variable, and do not suffer from bias arising from log-linearization. I show that in presence of strong treatment heterogeneity, the log-OLS estimator can yield estimates of the opposite sign by cumulating several types of biases. I contribute to the literature on the interpretation of models estimating semi-elasticities (Kennedy, 1981; Jan van Garderen

and Shah, 2002) and on non-linear difference-in-differences (Angrist, 2001; Ciani and Fisher, 2019; Wooldridge, 2023). I show that for non-linear difference-in-differences models, in presence of large treatment heterogeneity, the interpretation of the PPML estimator can differ by a lot from the interpretation of the log-OLS estimator. This is due to the fact that the average mutiplicative effect does not correspond to the multiplicative effect on the average. Finally, I contribute to the literature on the estimation of ATE with difference-in-differences strategy in the presence of heterogeneous treatment effects (De Chaisemartin and d'Haultfoeuille, 2020; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Borusyak et al., 2024; De Chaisemartin and D'Haultfoeuille, 2023; Nagengast and Yotov, 2023). I propose a new estimator robust to staggered treatment bias that recovers the ratio-of-ratios, a quantity similar to one yield by the TWFE PPML estimator in the two-by-two canonical case.

The rest of the paper proceeds as follows. Section 2 briefly presents an empirical model illustrating a data generating process of interest to researchers willing to estimate a proportional treatment effect (semi-elasticity). Section 3 presents the 2x2 canonical setting of multiplicative differences and the two-way fixed effect poisson-pseudo maximum likelihood estimator (TWFE PPML). Section 4 presents the staggered treatment case, the setting induced bias of TWFE PPML and a robust estimator to recover the ratio-of-ratios. Section 5 displays simulations comparing existing estimators in the canonical and staggered cases. Section 6 explores the question of Menkhoff and Miethe (2019) as an empirical application. Section 7 concludes.

2 Model

We observe individuals subject to a policy or a economic shock, and an outcome. We are interested in the causal effect of this policy on this outcome. We suppose that the ouctome under study y_{igt} has the following conditional mean:

$$E[y_{igt}|D_{it}] = exp(\alpha_i + \beta_t + \delta D_{igt}) \tag{2}$$

The researcher observes:

$$y_{igt} = exp(\alpha_i + \beta_t + \delta_{it} D_{igt})\eta_{igt} \tag{3}$$

With i the individual (person, firm or country), g the group (or cohort) of individuals treated on the same year, and t the year. D_{igt} takes the value 1 if individual i from group g is treated and time t, zero otherwise. The terms α_i and β_t respectively represent individual-specific determinants fixed in time and a time trend common to all individuals, driving y_{igt}. The term η_{igt} captures remaining individual-time varying heterogeneity such that $E[\eta_{igt}|D_{it}, \alpha_i, \beta_t] = 1$. The coefficients δ_{it} capture the heterogeneous treatment effects across i and t. We want to estimate the proportional effect of the policy on the outcome.

In case the policy under study has a homogeneous effects across individuals and time, the DGP updates to the following model:

$$y_{igt} = exp(\alpha_i + \beta_t + \delta D_{igt})\eta_{igt} \tag{4}$$

The model can be extended to include a vector of control variables X_{igt}:

$$y_{igt} = exp(\alpha_i + \beta_t + \delta_{it} D_{igt} + X'_{igt}\gamma)\eta_{igt} \tag{5}$$

Researchers often consider that the same DGP can be represented the following linear model, using a log transformation:

$$\ln y_{igt} = \alpha_i + \beta_t + \delta D_{igt} + \ln \eta_{igt} \tag{6}$$

With $E[\ln \eta_{it}|D_{it}] = E[\varepsilon_{it}|D_{it}] = 0$ which imposes stricter conditions on the error term (Silva and Tenreyro, 2006).

3 The 2x2 canonical setting

I begin by describing the simple canonical setting. The researcher observes two groups of countries $G = 0, 1$, at two periods $t = 0, 1$. Group 1 is treated at period 1 (i.e. the policy is imple-

mented), and that group 0 is never treated. For each group and time period, the researcher observes the outcome y_{igt}.

3.1 Quantity of interest and identification

In the case of multiplicative models, the researcher is often interested in the proportional treatment effect. When the researcher searches a treatment effect in percentage, or semi-elasticity, the multiplicative difference-in-difference targets (Angrist, 2001):

$$\frac{E[y_{igt}(1)] - E[y_{igt}(0)]}{E[y_{igt}(0)]} \tag{7}$$

This quantity is the change in the outcome induced by the treatment, or the ATE, as a proportion of the non treated outcome level. It is the change of the expected outcome variable in percentage of the expected outcome in the absence of treatment: a semi-elasticity. This is also the quantity that Chen and Roth (2023) advise to target when the researcher wants to include zeros and recover a treatment effect in percentage. In case of constant treatment effect across individuals, this quantity correspond to $exp(\delta) - 1$ from our model.

In the sample, we can only estimate the average treatment on the treated, the ATT, and its proportional counterpart, the proportional treatment on the treated (PTT). The PTT corresponds to the quantity of interest from (7) and is the ATT normalized by the expected value of the non treated outcome:

$$\begin{aligned} PTT &= \frac{E[y_1(1)|G=1] - E[y_1(0)|G=1]}{E[y_1(0)|G=1]} = \frac{ATT}{E[y_1(0)|G=1]} \\ &= \frac{E[exp(\alpha_i + \beta_t + \delta_i)\eta_{it}|G=1] - E[exp(\alpha_i + \beta_t)\eta_{it}|G=1]}{E[exp(\alpha_i + \beta_t)\eta_{it}|G=1]} \\ &= \frac{E[exp(\alpha_i + \beta_t)(exp(\delta_i) - 1)|G=1]}{E[exp(\alpha_i + \beta_t)|G=1]} \end{aligned} \tag{8}$$

It is important to note that using the linear model from equation 6 does not always target the same quantity of interest, especially when treatment effect is heterogeneous (Ciani and Fisher, 2019). Rewriting the linear difference-in-difference target:

$$\begin{aligned} &E[\ln y_1(1)|G=1] - E[\ln y_1(0)|G=1] = E[\alpha_i + \beta_t + \delta_i + \varepsilon_i|G=1] - E[\alpha_i + \beta_t + \varepsilon_i|G=0] \\ &= E[\delta_i|G=1] \neq \ln E[exp(\delta_i)|G=1] \end{aligned} \tag{9}$$

Because of Jensen's inequality. The target of the linear model is the approximated average effect when treatment effects δ_i are small. The two model estimation targets are the same only when treatment effect is homogeneous: $\delta_i = \delta, \forall i$.

$E[y_1(1)|G=1]$ can be directly estimated from coresponding moments in the data, but not $E[y_1(0)|G=1]$ which is by definition never observed. Further assumptions are needed to estimate the ATT and PTT.

3.1.1 Identifying assumptions

A1: No anticipation assumption On average, among the eventually treated group there are no anticipatory changes that affect the potential outcomes prior to the intervention.

$$E[y_0(1) - y_0(0)|G=1] = 0 \tag{10}$$

A2: Unconditional multiplicative parallel trend assumption (MPT) This assumption states that in the absence of treatment, changes in percentages of expected outcome should have been the same in the two groups. The averages of the two groups would have known the same growth in the absence of treatment.[3]

$$\frac{E[y_1(0)|G=1]}{E[y_0(0)|G=1]} = \frac{E[y_1(0)|G=0]}{E[y_0(0)|G=0]} \tag{11}$$

Using model 3 notations, another version of this assumption is:

$$\forall it, E[y_{it}(0)|D_{it},] = exp(\alpha_i + \beta_t) \tag{12}$$

Note that in this case the parallel trend is on the growth of the averages and not on the average growths. The linear model therefore also differs in term of parallel trend assumption:

$$\begin{aligned} &E[\ln y_1(0) - \ln y_0(0)|G=1] = E[\ln y_1(0) - \ln y_0(0)|G=0] \\ \Leftrightarrow &E[\ln y_1(0)|G=1] - E[\ln y_0(0)|G=1] = E[\ln y_1(0)|G=0] - E[\ln y_0(0)|G=0] \end{aligned} \tag{13}$$

This assumption states that in the absence of treatment, the expected log of the outcome in the treated group should have changed by the same log points as the non-treated group. It implies that the average approximated growth rate should have been the same in the two groups. With treatment heterogeneity, the two models do not imply the same parallel trend assumption.

A visual exploration can be undertaken on pre-trends to check which assumption seems more reliable. In the case of the multiplicative model, the pretrend should be similar when

[3]This assumption is also called the index parallel trend assumption by Wooldridge (2023), for GLM difference-in-differences.

the researcher plots the logarithm of the total outcome for treated and control groups. In case of the log-linear model, the pretrend should be similar when the researcher plots the average logarithm of the outcome for treated and control groups.

A2.B: Conditional multiplicative parallel trend assumption (MPT) The previous assumption can be amended to include covariates. If the multiplicative parallel trend holds conditionally, we assume that in the absence of treatment and conditional on the change in the outcome induced by covariates X_{it}, the treated group would have followed the same trend as the untreated one. Using model 3 notations, this is equivalent to assuming:

$$\forall it, E[y_{it}(0)|D_{it}, X_{it}] = exp(\alpha_i + \beta_t + X'_{it}\gamma) \tag{14}$$

3.1.2 Identification

$E[y_1(0)|G = 1]$ can be expressed as a function of terms that can be estimated using the multiplicative parallel trend assumption (A.2):

$$E[y_1(0)|G = 1] = \frac{E[y_1(0)|G = 0] \times E[y_0(0)|G = 1]}{E[y_0(0)|G = 0]}$$

I inject (11) in (8) and recover the PTT expressed as a ratio of ratios (by analogy to a difference-in-differences in the linear case):

$$PTT = \frac{E[y_1(1)|G = 1]}{E[y_0(0)|G = 1]} / \frac{E[y_1(0)|G = 0]}{E[y_0(0)|G = 0]} - 1 \tag{15}$$

The expression of the ATT follows, which expresses the treatment effect in units of the outcome variable:

$$\begin{aligned} ATT &= E[y_1(1) - y_1(0)|G = 1] \\ &= E[y_1(1)|G = 1] - E[y_1(0)|G = 1] \\ &= E[y_1(1)|G = 1] - \frac{E[y_1(0)|G = 0] \times E[y_0(0)|G = 1]}{E[y_0(0)|G = 0]} \end{aligned} \tag{16}$$

3.2 Estimation

3.2.1 Corresponding sample moments

The PTT and ATT can be estimated from their corresponding sample moments. In the following expressions, I denote G_i a binary variable taking the value 1 if the individual i belongs to the

treated group, and $y_{i,t}$ the outcome of country i at time t. There are n individuals in the sample. Estimates of both quantities $\widehat{\tau}$ and $\widehat{RoR}$ are:

$$\widehat{\tau} = \frac{\sum_{i=1}^{n} G_i(y_{i,1})}{\sum_{i=1}^{n} G_i} - \frac{\frac{\sum_{i=1}^{n}(1-G_i)(y_{i,1})}{\sum_{i=1}^{n}(1-G_i)} \times \frac{\sum_{i=1}^{n} G_i(y_{i,0})}{\sum_{i=1}^{n} G_i}}{\frac{\sum_{i=1}^{n}(1-G_i)(y_{i,0})}{\sum_{i=1}^{n}(1-G_i)}}$$

$$\widehat{\tau} = \frac{1}{\sum_{i=1}^{n} G_i}\Big(\sum_{i=1}^{n} G_i(y_{i,1}) - \frac{\sum_{i=1}^{n}(1-G_i)(y_{i,1})}{\sum_{i=1}^{n}(1-G_i)(y_{i,0})} \times \sum_{i=1}^{n} G_i(y_{i,0})\Big) \tag{17}$$

To estimate the ATT, the average outcome of the treated group in period 0 is multiplied by the growth rate of the non treated group between the two periods. It recovers the counterfactual outcome of the treated group if the multiplicative parallel trend assumption holds (i.e. parallel trend assumption in the growth rate).

The proportional treatment effect is recovered by computing a ratio of ratios (RoR), relying again on the multiplicative parallel trend assumption.

$$\begin{aligned}\widehat{RoR} &= \frac{\frac{\sum_{i=1}^{n} G_i(y_{i,1})}{\sum_{i=1}^{n} G_i}}{\frac{\sum_{i=1}^{n} G_i(y_{i,0})}{\sum_{i=1}^{n} G_i}} \Big/ \frac{\frac{\sum_{i=1}^{n}(1-G_i)(y_{i,1})}{\sum_{i=1}^{n}(1-G_i)}}{\frac{\sum_{i=1}^{n}(1-G_i)(y_{i,0})}{\sum_{i=1}^{n}(1-G_i)}} - 1 \\ &= \frac{\sum_{i=1}^{n} G_i(y_{i,1})}{\sum_{i=1}^{n} G_i(y_{i,0})} \Big/ \frac{\sum_{i=1}^{n}(1-G_i)(y_{i,1})}{\sum_{i=1}^{n}(1-G_i)(y_{i,0})} - 1\end{aligned} \tag{18}$$

This is the ratio-of-ratios (RoR) estimator.

3.2.2 Equivalence of TWFE PPML and ROR estimator

In the linear canonical setting, there is a direct equivalence between the moments used to recover the difference-in-differences and the quantity computed by the two-way fixed effect OLS estimator. This property largely motivated the use of TWFE estimator by the applied literature.

A similar analogy holds in the multiplicative model case. In the non-linear case, the TWFE PPML estimator in the 2-by-2 setting computes the RoR and maintains an equivalence. This result is provided by Ciani and Fisher (2019) and I verify this property in following in simulations and estimations.

In the canonical setting, we have:

$$exp(\widehat{\delta_{PPML}}) - 1 = \frac{\sum_{i=1}^{n} G_i(y_{i,1})}{\sum_{i=1}^{n} G_i(y_{i,0})} \Big/ \frac{\sum_{i=1}^{n}(1-G_i)(y_{i,1})}{\sum_{i=1}^{n}(1-G_i)(y_{i,0})} - 1 \tag{19}$$

The same quantity as in (18) that converges in probability, under the identification assumptions, to the quantity of interest (8):

$$\frac{E[y_1(1)|G=1] - E[y_1(0)|G=1]}{E[y_1(0)|G=1]} \tag{20}$$

This property, and the tractability of the PPML estimator with high dimensional fixed effects motivates the use of this estimator.

3.2.3 Other reasons for the researcher to choose TWFE PPML over TWFE log-OLS

Appart from the different estimation target and parallel trend assumption, there are other reasons why an applied researcher might prefer to use TWFE PPML over TWFE log-OLS.

Zeros The most well understood issue in the economics literature is the exclusion of zeros by log-OLS. The model cannot be estimated on observations with zeros in the dependent variable, excluding them from the estimation sample. This is because a proportional change for the extensive margin is not defined (Chen and Roth, 2023). TWFE PPML solves this issue by estimating the change in the average. This quantity weights predicted individual proportional changes by their $predicted$ counterfactual outcome share in total predicted outcome. Intuitivelly, PPML provides small weights to individuals zeros or small observations by predicting small counterfactual outcomes, because these individuals are the most likely to display extreme proportional changes.

Heteroskedasticity bias Back to the case of multiplicative difference-in-differences, considering a log-normal error term once again, the estimated coefficient of the log-OLS model is (Ciani and Fisher, 2019):

$$\widehat{\delta}_{log-OLS} = \delta - \frac{\sigma^2_{11}}{2} + \frac{\sigma^2_{10}}{2} + \frac{\sigma^2_{01}}{2} - \frac{\sigma^2_{00}}{2} \tag{21}$$

With σ^2_{gt} the conditional variance of $\ln \eta_{it}$ in group g at time t. If the variance fo the error term is constant within groups across time in the absence of treatment, we have that the bias is caused by the change in variance induced by the treatment is:

$$\widehat{\delta}_{log-OLS} = \delta + \frac{\sigma^2_{10} - \sigma^2_{11}}{2} \tag{22}$$

The second term is negative if the variance of the error term increases with treatment, and vice versa (second order sochastic dominance).

Cohn et al. (2022) provide an intuition of how the log-OLS bias can be magnified for empirical models including fixed effects. In a multiplicative model, the group fixed effects affect both the level of the outcome and the variance of the error term, as they are scaling parameters. Based on simulations, the authors show that the inclusion of fixed effects magnifies the log-OLS bias if the fixed effects account for a bigger share of the variation in y_{igt} than in the heteroskedastic standard deviation of the error term $\sigma_\eta(x)$.

4 Multiperiod setting and heterogeneous timing

I turn to the multiperiod and multicohort setting. There are now T time periods starting at $t = 1$, and g cohorts of individuals treated at different times. Once a cohort gets treatment, it is considered treated until the end on the panel (treatment is not reversible). The two assumptions made earlier are generalized to this setting:

A1: No anticipation assumption On average, among the eventually treated group there are no anticipatory changes that affect the potential outcomes prior to the intervention.

$$E[y_{g,t}(1) - y_{g,t}(0)|G = 1] = 0 \quad \forall t < q \tag{23}$$

With q the time of treatment for cohort g.

A2: Multiplicative parallel trend assumption

$$\forall it, E[y_{it}(0)|D_{it}] = exp(\alpha_i + \beta_t) \tag{24}$$

Which implies that across units i, for all periods t and t', $\frac{E[Y_{it'}(0)]}{E[Y_{it}(0)]}$is the same. Again, this is equivalent to assuming that in the absence of treatment, the growth rate of the average outcome in the treated group between two time periods would have been the same than in the non-treated group.

Researchers have been tempted to extend the equivalence between the DiD and TWFE estimators to the multiperiod setting. A recent literature in econometrics (see De Chaisemartin and D'Haultfoeuille, 2023) shows that TWFE estimators in a multiperiod multigroup setting can lead to biased estimates of the ATT because of a model making too strict assumptions on treatment homogeneity. When units are treated at different times and treatment effect are heterogeneous

across time periods, this type of model makes wrong comparisons between treated and control groups, and estimates a quantity that averages treatment effects with negative weights.

I show with a simple example that this problem also arises in the multiplicative case with PPML, and discuss approaches to recover a proportional treatment effect in staggered settings.

4.1 PPML TWFE bias

We take a simple example to show that the TWFE PPML estimator is biased under the same conditions as TWFE OLS. There are two individuals $i = A, B$ observed at three time periods $t = 1, 2, 3$. Individual A is treated in period $t = 2$ and individual B is treated in period $t = 3$, such that B is the control group for individual A in $t = 2$. The treatment effect is proportional, such that there is a multiplicative parallel trend holding (i.e. in growth rate):

$$y_{it} = exp(\alpha_i + \beta_t + \delta_{it} D_{it})\eta_{it} \tag{25}$$

If treatment effect is homogeneous, there is $\delta_{A2} = \delta_{A3} = \delta_{B3} = \delta$. If we have heterogeneous treatment effect then $\delta_{A2} \neq \delta_{A3} \neq \delta_{B3}$. The quantity of interest is then:

$$PTT = \frac{E[y_{it}(1)|D=1] - E[y_{it}(0)|D=1]}{E[y_{it}(0)|D=1]} = \sum_{i,t,D_{it}=1} \frac{E(y_{it}(0))}{\sum_{i,t,D_{it}=1} E(y_{it}(0))} \Big(exp(\delta_{it}) - 1 \Big) \tag{26}$$

Which is a weighted sum of cohort and time specific treatment effects $exp(\delta_{it}) - 1$. The weights ω_{it} correspond to the share of the counterfactual outcome in the total size of counterfactual observations.[4] The TWFE PPML estimator assumes the following model:

$$y_{it} = exp(\alpha_i + \beta_t + \delta D_{it})\eta_{it} \tag{27}$$

Solving the system yields the TWFE PPML estimator for the proportional treatment effect $exp(\delta) - 1$:

$$exp(\widehat{\delta}_{PPML}) - 1 = \frac{y_{A2}(y_{B1} + y_{B3}) - y_{B2}(y_{A1} + y_{A3}))}{y_{B2}(y_{A1} + y_{A3}))} \tag{28}$$

With homogeneous treatment effect, using expected values of outcome realization, this quantity should yield:

$$\frac{E(y_{A2}(y_{B1} y_{B3})) - E(y_{B2}(y_{A1}) + y_{A3}))}{E(y_{B2}(y_{A1} y_{A3}))} = exp(\delta) - 1 \tag{29}$$

[4] The PPML estimator weights more cells with large counterfactual outcomes and reduces weigths associated to cells with the smaller counterfactual outcomes which are the most susceptible to display the most extreme proportional changes.

With treatment heterogeneity, the quantity estimated by TWFE PPML becomes:

$$\frac{E(y_{A2}(y_{B1}y_{B3})) - E(y_{B2}(y_{A1}) + y_{A3}))}{E(y_{B2}(y_{A1}y_{A3}))} = exp(\delta_{A2}) \times \frac{1 + exp(\delta_{B3} + \beta_3)}{1 + exp(\delta_{A3} + \beta_3)} - 1 \tag{30}$$

The TWFE PPML recovers here the growth rate of the "good" comparison period ($t = 2$), scaled by the differential in growth rate between the two groups in the second period. This scaling will be bigger if the common trend in this later period is large ($exp(\beta_3)$ is high). There is an analogy with the problem encountered in the linear case, with some treatment effects contributing potentially with a potential negative weight to the quantity of interest.

4.2 Robust estimators for TWFE PPML

The issue generated by the TWFE estimators comes from the fact that it imposes constraints that are too strong on the model in the staggered setting. Recent papers solve this issue in the linear case by allowing for the most flexible model given the data structure (Sun and Abraham, 2021; Borusyak et al., 2024; Wooldridge, 2021).

Wooldridge (2023) extends this idea to the non-linear case. With g_{iq} and indicator variable taking the value one if idividual i is treated in period q, one can estimate the following model using poisson-pseudo maximum likelihood:

$$E[y_{it}|g_{iq}, ..., g_{iT}] = exp\Big[\sum_{r=q}^{T}\sum_{l=0}^{T-r} \delta_{rs}\Big(D_{it} \times g_{ir} \times \mathbb{1}\{t - r = l\}\Big) + \alpha_i + \beta_t\Big] \tag{31}$$

In this model:

$$\begin{aligned} \delta_{gt} &= log(E(y_{igt}(1)|D = 1)) - log(E(y_{igt}(0)|D = 1)) \\ \Leftrightarrow exp(\delta_{gt}) - 1 &= \frac{E(y_{igt}(1)|D = 1) - E(y_{igt}(0)|D = 1)}{E(y_{igt}(0)|D = 1)} \end{aligned} \tag{32}$$

So estimating δ_{gt} recovers the correct estimation target: the proportional treatment effect on cohort g and time t. The researcher is often interested in a more aggregated estimation target, which is not covered by Wooldridge (2023). The next section explain why aggregating cohort-time treatment effects as in the linear case presents some caveats with a multiplicative model, and the next section presents the proposed approach of this paper.

4.2.1 Aggregation estimators in the non-linear case

Robust estimators have been developed for the linear case to reccover aggregate treatment effects (De Chaisemartin and d'Haultfoeuille, 2020; Callaway and Sant'Anna, 2021; Sun and Abraham,

2021; Borusyak et al., 2024; Wooldridge, 2021). These estimators rely on recovering treatment effects for correct "building blocks" (i.e. cohorts) and aggregating them over the desired sample to recover an estimate of the ATT. For example De Chaisemartin and d'Haultfoeuille (2020); Callaway and Sant'Anna (2021); Sun and Abraham (2021) compute two-by-two DiD estimators for each existing combination of treated cohort and time period. For three treated cohorts $1, 2, 3$, they would for example compute the treatment effect on cohort 1 by computing the DiD using a control group of never treated on the time period, then the DiD of cohort 2 using a similar control group, ... And aggregate all estimated DiD to recover the average effect. Given that the models used are linear, the ATT can be easily retrieved aggregating linear treatment effects.

Translated in the multiplicative setting, one could also compute the two-by-two estimates of $PTT_{g,t}$ by group and time period, and average this effect to recover an aggregate treatment effect. This would yeld an estimator of the form:

$$\sum \nu_{g,t} \widehat{RoR}_{g,t} \tag{33}$$

With $\nu_{g,t}$ a weight associated to observations in g, t, chosen by the researcher depending on the estimation target.

Using the fully interacted model above, we know that estimating coefficients δ_{rs} recovers the multiplicative model estimation target for each cohort-time cell: $exp(\widehat{\delta_{rs}^{PPML}}) - 1$ is the multiplicative effect on the average of cohort g at time t. Such an "aggregation" estimator with the same spirit as the ones from the linear case is:

$$exp\Big(\sum_{g}^{G} \sum_{t}^{T} \nu_{g,t} (\widehat{\delta_{g,t}^{PPML}}) \Big) - 1 \tag{34}$$

If treatment is homogeneous within cohort-time cells, this estimator approximates the average proportional effect on the treated. The estimator can be easily implemented using the ppmlhdfe stata command when the number of parameters to estimate gets big: interaction coefficients can be estimated at fixed effect, appropriately rescaled and aggregated to recover (34).[5]

The researcher should bear in mind that this quantity presents two caveats in the multiplicative case. First, its interpretation is different than the ratio-of-ratios estimator in the canonical case when treatment effect is homogeneous within cells. The interpretation will be closer to the one recovered by the log-linear model, which is a transformation of the average parameter of the

[5]See Correia et al. (2020) for ppmlhdfe command.

model, and approximate the average proportional effect. But the quantity in (35) would most of the time differ from the proportional change in the average because of Jensen's inequality.

$$exp\Big(\sum_{g}^{G}\sum_{t}^{T}\nu_{g,t}(\delta_{g,t})\Big) - 1 = exp\Big(\sum_{g}^{G}\sum_{t}^{T}\nu_{g,t}(log\Big(\frac{E[y_{gt}(1)|g] - E[y_{gt}(0)|g]}{E[y_{gt}(0)|g]}\Big) + 1)\Big) - 1 \tag{35}$$

When comparing estimates of the static TWFE PPML and this aggregation estimator, one should be concious about the fact that the difference between the two is due to a different quantity estimated and the staggered setting bias. Compare to the staggered robust estimators available for the log-linear model, this estimator has the advantage to be robust to the heteroskedasticity bias developed above. It also means that this estimator is not always suited to check the robustness of the parallel trend assumption on leads coefficients. Coefficients are estimated using the multiplicative model, which implies a parallel trend assumption in growth rate of the average outcome. It could be that this parallel trend holds in the average growth rate but not in the average parameter.

Second and more worryingly, if $\delta_{igt} \neq \delta_{gt},\ \forall i \in g$, the quantity recovered by this estimator might not have an interpretable meaning because of Jensen's inequality. If there is treatment effect heterogeneity within a cohort-time cell g, t, the estimated coefficient $\widehat{\delta_{g,t}^{PPML}}$ will recover the proportional treatment effect on the average of cell g, t. The estimator in (34) recovers the average over cells of multiplicative treatment effect on the average of cells, an intermediate quantity between the estimated parameter (log-OLS) and the estimated growth rate of the average (ratio-of-ratios). It would moreover arbitrarily depend on the structure of the panel and the treatment timings.

4.2.2 Proposed imputation estimator

The next section proposes a new estimator for proportional treatment effects, recovering a semi-elasticity derived from the ratio-of-ratios estimator. This estimator is robust to any type of treatment heterogeneity in a staggered treatment setting. It is an imputation estimator in the spirit Borusyak et al. (2024), based on the idea that one can specify the correct counterfactual model. Wooldridge (2023) shows that this approach is equivalent to the fully interacted model and I derive the equivalent interaction estimator in appendix.

Under our identification assumptions, on the expected conditional mean of the counterfactual outcome is: $E[y_{igt}(0)|D_{igt} = 1] = exp(\alpha_i + \beta_t)$. The parameters α_i and β_t can be estimated

on the sample of never-treated and not-yet treated observations. One can then predict the counterfactual outcomes for the treated sample, using estimates of these estimates:

$$\widehat{y_{igt}(0)} = exp(\widehat{\alpha_i} + \widehat{\beta_t})$$

Wooldridge (2023) states that:

$$\widehat{\tau_{g,t}} = \sum_{i \in g} y_{igt}(1) - \widehat{y_{igt}(0)}$$

Estimates the ATT in level for cohort g and time t, and can also be recovered by predicting the treatment average partial effect for cell g, t. Contrary to coefficients $\widehat{\delta_{g,t}}$, this is a linear effect that can be aggregated linearly without loss of interpretability. I recover the average treatment effect in level on the full treated sample, which is equivalent to computing the difference between the observed outcome and the predicted one on the treated sample:

$$\widehat{\overline{\tau}} = \sum_{i, D_{igt}=1} N_{g,t} \widehat{\tau_{g,t}} = \sum_{i, D_{igt}=1} y_{igt}(1) - \widehat{y_{igt}(0)} \tag{36}$$

To recover the proportional treatment effect, or treatment semi-elasticity, this quantity can be scaled by the total counterfactual outcome to recover the following estimator:

$$\begin{aligned} \widehat{RoR}_{imput} &= \frac{\widehat{\overline{\tau}}}{\sum_{i, D_{igt}=1} \widehat{y_{igt}(0)}} \\ &= \frac{\sum_{i, D_{igt}=1} y_{igt}(1)}{\sum_{i, D_{igt}=1} \widehat{y_{igt}(0)}} - 1 = \frac{\frac{1}{N_{D=1}} \sum_{i, D_{igt}=1} y_{igt}(1)}{\frac{1}{N_{D=1}} \sum_{i, D_{igt}=1} \widehat{y_{igt}(0)}} - 1 \end{aligned} \tag{37}$$

This estimator is based on the ratio of the average of observed and counterfactual outcomes. Its interpretation is similar to the TWFE PPML and RoR estimator in the canonical setting: the percentage change in the average outcome due to treatment. The numerator converges in probability to the expected value of the treated outcome in the treated group. The denominator, under the assumption the parallel trend is valid, converges to the expected value of the untreated outcome in the treated group. It is obtained by multiplying the average outcome of the treated groups in the pre-treatment period by the growth rate of the non-treated group after treatment. This quantity should converge to the true PTT, provided that the denominator does not reach zero. This is unlikely to take place: the PPML model only predict strictly positive values. Moreover for the model to predict counterfactual outcomes very close to zero, it means that the researcher faces a DGP in which treatment affect mainly the extensive margin, and therefore is more suited for a binary outcome model. The estimator $\widehat{RoR}_{imput}$ can be easily computed for a less aggregated level, such as cohort or relative time.

4.2.3 Special cases for the imputation estimator

Categorial parallel trends Empirical researchers often choose to specify categorial parallel trends, or parallel trends holding across some groups of the population. For example, if treated and control firms are compared over time within the same region or sector. In the TWFE model, this translates in specifying time fixed effects disaggregated by the desired catagory c:

$$y_{igct} = exp(\alpha_i + \beta_{ct} + \delta D_{igt})\eta_{igt}$$

The correct estimation of the treatment effect in level and counterfactual outcome requires to slightly adjust the counterfactual model and to estimate more parameters. The imputation procedure only requires to estimate $y_{ict} = exp(\alpha_i + \beta_{ct})\eta_{ict}$ on the treated sample for pre-treatment periods and the never-treated to get $\hat{\alpha}_i$ and $\hat{\beta_{jt}}$. The predicted counterfactual outcome on the treated sample:

$$\widehat{y_{igt}(0)} = exp(\widehat{\alpha_i} + \widehat{\beta_{ct}})$$

$$\widehat{\overline{\tau}} = \sum_{i \in \omega_1} y_{igt}(1) - \widehat{y_{igt}(0)}$$

Going through the imputation process is less computationally intensive than the interaction approach, especially when the number of categories c increase, and is numerically equivalent.

Control variables The mutiplicative parallel trend often holds conditionally on a set of control variables (assumption A2.B). Wooldridge (2023) explains how to include time-constant controls in the fully saturated model and keep the equivalence with the imputation estimator:

$$\begin{aligned} E(y_{ipt}|g_{it}, ..., g_{iT}) = exp\Big[& \sum_{s=2}^{T} (fs_t X_i)\pi_{Ct} + \sum_{g=q}^{T} \sum_{l=0}^{T-r} (D_{it} \times g_{ig} \times \mathbb{1}\{t-g=l\})\delta_{gl} \\ & + \sum_{p=1}^{P} \sum_{g=q}^{T} \sum_{l=0}^{T-r} (\mathbb{1}\{c=s\} \times D_{it} \times g_{ig} \times \mathbb{1}\{t-g=l\})\zeta_{pgl} \\ & + \sum_{p=1}^{P} \sum_{g=q}^{T} \sum_{l=0}^{T-r} (\dot{X_{ig}} \times D_{it} \times g_{ig} \times \mathbb{1}\{t-g=l\})\xi_{gl} + \alpha_{ij} + \beta_{jt} \Big] \end{aligned}$$

Coefficients π_{Ct} capture the divergence from the parallel trend due to control variable X_{ij}. Coefficients ξ_{gl} the divergence from the average treatment effect in cohort g at time l due to variables X_{ij}. Control variables are centered on the treated sample: $\dot{X} = X - E(X|D=1)$. This normalization ensures that δ_r has the desired interpretation $log(E[y_r(1)|g=1]) - log(E[y_r(0)|g=1])$

among the treated. The ATT is recovered as before by predicting average partial effects of treatment on the desired sample. The idea behind this model is to allow full heterogeneity in treatment effect across the level of control variables X_i in the sample. Limitation to time constant covariates can be quite restrictive, especially if researchers wish to control for within cohorts time varying shocks that could counfound the treatment effect estimation.

I propose instead to use the imputation process by allowing for a more flexible counterfactual model (such as Borusyak et al. (2024)). The equivalence with the interaction approach breaks in this case. The researcher assumes the DGP to approximated by:

$$y_{igt} = exp(\alpha_i + \beta_t + \delta_{it} D_{igt} + X'_{igt}\gamma)\eta_{igt} \tag{38}$$

With X_{igt} a set of individual specific, time-varying, variables impacting the outcome y_{igt}. Under the conditional parallel trend (14), the counterfactual model can be estimated by:

$$\widehat{y_{igt}(0)} = exp(\widehat{\alpha_i} + \widehat{\beta_{ct}} + X'_{igt}\widehat{\gamma})$$

Which requires to estimate $\widehat{\alpha_i}$, $\widehat{\beta_{ct}}$ and $\widehat{\gamma}$ on the sample such that $D_{it} = 0$. We recover the proportional treatment effect then as:

$$\widehat{RoR}_{imput} = \frac{\sum_{i\in\omega_1} y_{igt}(1) - \widehat{y_{igt}(0)}}{\sum_{i\in\omega_1} \widehat{y_{igt}(0)}} = \frac{\widehat{\bar{\tau}}_{imput}}{\sum_{i\in\omega_1} \widehat{y_{igt}(0)}}$$

Triple differences In a triple difference approach, researchers observes treated cohorts that differ along two additional dimensions, denoted as j and p, which are used to select control groups. These dimensions can represent sectors and products, or regions, and are used to correct the potential bias of a simple difference-in-differences estimator by cancelling out this bias using a supplementary dimension (Olden and Møen, 2022). The expected conditional mean takes the following form:

$$E[y_{i(jp)gt}|G, D_{it}] = exp(\alpha_i + \beta_{jt} + \beta_{pt} + \delta_{it} D_{igt}) \tag{39}$$

The new parallel trend assumption becomes:

$$E[y_{i(jp)gt}(0)|G] = exp(\alpha_i + \beta_{jt} + \beta_{pt}) = exp(\alpha_i) \times \underbrace{exp(\beta_{jt}) \times exp(\beta_{pt})}_{Relative\ growth\ rate} \tag{40}$$

Here, $exp(\beta_{jt})$ and $exp(\beta_{pt})$ denote the relative growth rates associated with the two dimensions. If j is a state and p a product this assumption states that the relative growth rate between treated

and non treated products in the treated state should have been the same that in the non treated states in the absence of treatment.

Using the imputation approach simplifies the analysis compared to interaction models, which require interacting all the cohort time interactions with the p and j dimensions to break down δ_{gs} coefficients. With the imputation approach, the expected outcome y_{ijpgt} is estimated on the not-yet and never-treated samples as $exp(\alpha_i+\beta_{jt}+\beta_{pt})\eta_{ijt}$. The imputed counterfactual outcome $\widehat{y_{ijpgt}(0)}$ is then calculated as $exp(\widehat{\alpha_i} + \widehat{\beta_{jt}} + \widehat{\beta_{pt}})$. The estimate is recovered as above.

5 Simulations

I simulate data to compare estimators presented in the previous section.

5.1 Common treatment timing

5.1.1 Data generating process

I generate a panel of 10,000 individuals observed for three time periods $t = 1, 2, 3$. The outcome y_{it} follows a multiplicative data generating process:

$$y_{it} = exp(\mu_t + \alpha_i + \delta_i D_{it})\eta_{it}$$

With μ_t the time effects, α_i the individual effects, D_{it} the treatment status and δ_i the treatment effect, and η_{it} a log-normal error term such that $E[\eta_{it}|\mu_t, \alpha_i, D_{it}] = 1$. Individuals are treated in period 3, such that treatment timing is homogeneous. I generate some selection into treatment status so that I need to implement a difference-in-differences strategy to recover the causal effect of treatment.

I also introduce heteroskedasticity in the error term as a function of observables: in one case variance is function of individual fixed effects, and in the other it is a function of the treatment status. This second case should jeopardize retrieval of causal treatment effect via log-OLS (Ciani and Fisher, 2019). Finally, I simulate a homogeneous treatment effect and a case with heterogeneity in treatment effect accross individuals. The average treatment effect is positive as $\delta > 0$ when it is homogeneous. Heterogeneity across individuals is normally distributed, such that the average growth rate corresponds to the growth rate of the average. The standard error is set such that a small portion of individuals could face true negative treatment effect. Heterogeneity

in treatment effect allows to compare estimators in their capacity to estimate the average treatment effect. The observed outcome y_{it} is always strictly positive such that there is no difference between estimators driven by zeros in the outcome. I simulate the data 1000 times.

Table 1: Common timing: simulation cases

Case	$V(\eta_{it}\|.)$	Treatment effect parameter
1	α_i	$\delta = 0.31$
2	$0.2D_{it}$	$\delta = 0.31$
3	α_i	$\delta_i = 0.31 + \nu_i, \quad \nu_i \sim \mathcal{N}(0, 0.5)$
4	$0.2D_{it}$	$\delta_i = 0.31 + \nu_i, \quad \nu_i \sim \mathcal{N}(0, 0.5)$

5.1.2 Simulations results

Table 2 displays the true distribution of the treatment effect and the distribution of estimators accross simulations. The upper panel displays cases 1 and 2 with homogeneous treatment effect and the lower panel cases 3 and 4 with heterogeneous treatment effect. In those later case, I provide the distribution of the average parameter $exp(\overline{\delta_i}) - 1$ and the true RoR or growth rate of the average $\frac{E[y_{1igt}|G=1]-E[y_{0igt}|G=1]}{E[y_{1igt}|G=1]}$. I estimate the treatment effect using TWFE log-OLS, TWFE PPML and the imputation estimator. Densities of estimators are displayed in figure B1a in appendix.

Two estimators are unbiased and strictly equivalent: TWFE PPML and the imputation estimator. For each simulation, they provide the same estimate. I compare TWFE PPML and TWFE log-OLS. In case of homogeneous treatment and individual heteroskedasticity, in the left upper panel, log-OLS is the most efficient unbiased estimator. PPML is centered close to the true value of the effect but display twice a larger variance. Assumption that the treatment has no effect on the variance of the outcome is quite restrictive. When I introduce heterogeneity varying by treatment status, log-OLS estimates display a downward bias, while TWFE PPML and imputation are unbiased.

More worriying results come from the case of heterogeneous treatment effects, an assumption difficult to rule out in empirical applications. In both cases, PPML estimates are quite close

Table 2: 1000 simulations: canonical setting

Homogeneous treatment effect $exp(\delta) - 1$

	Case 1 $V(\eta_{it}\|.) = \alpha_i$				Case 2 $V(\eta_{it}\|.) = 0.2D_i$			
Estimator	Mean	St.D.	Min	Max	Mean	St.D.	Min	Max
$exp(\delta) - 1$	0.363	0	0.363	0.363	0.363	0	0.363	0.363
TWFE PPML/RoR	0.370	0.0810	0.112	0.658	0.362	0.0509	0.203	0.543
Imputation	0.370	0.0810	0.112	0.658	0.362	0.0509	0.203	0.543
TWFE Log-OLS	0.364	0.0357	0.259	0.475	0.293	0.0266	0.200	0.367

Heterogeneous treatment effect $exp(\delta_i) - 1$

	Case 3 $V(\eta_{it}\|.) = \alpha_i$				Case 4 $V(\eta_{it}\|.) = 0.2D_i$			
Estimator	Mean	St.D.	Min	Max	Mean	St.D.	Min	Max
$exp(\overline{\delta_i}) - 1$	0.363	0.0104	0.333	0.398	0.364	0.00991	0.334	0.396
True RoR	0.545	0.0125	0.508	0.585	0.545	0.0118	0.505	0.583
TWFE PPML/RoR	0.544	0.0961	0.233	0.981	0.545	0.0635	0.332	0.768
Imputation	0.544	0.0961	0.233	0.981	0.545	0.0635	0.332	0.768
TWFE Log-OLS	0.363	0.0366	0.258	0.496	0.294	0.0281	0.195	0.379

to the true value of the growth rate. Estimates are less precise when heteroskedasticity is correlated with individual effects (Case 3) rather than treatment (Case 4). Turning to TWFE log-OLS, we observe that the model recovers the exponential of the average parameter, a quantity that cannot be interpreted in terms of semi-elasticity. The magnitude of the treatment effect differs by a lot from the PPML estimates and the true growth rate of the average. Moreover, in case of treatment related heteroskedasticity (Case 4 in the lower right panel) there is an added bias.

Simulations indicate that TWFE PPML is more adapted to the estimation of semi-elasticity treatment effects. Log-OLS estimator suffers from two drawbacks: first the heteroskedasticity bias discussed by Silva and Tenreyro (2006) and Ciani and Fisher (2019). Second, it cannot retrieve an interpretable average treatment effect in case of treatment heterogeneity, but rather

the average parameter. To these problems adds-up the new TWFE issues introduced by the recent literature (De Chaisemartin and D'Haultfoeuille, 2023).

5.2 Staggered treatment

5.2.1 Data generating process

I generate a panel of 10,000 individuals observed for fifteen time periods $t = 1, ..., 15$. The outcome y_{igt} follows a multiplicative data generating process:

$$y_{igt} = exp(\mu_t + \alpha_i + \delta_{it} D_{igt})\eta_{igt}$$

With μ_t the time effects, α_i the individual effects, D_{igt} the treatment status and δ_{it} the treatment effect. Finally η_{igt} a log-normal error term such that $E[\eta_{igt}|\mu_t, \alpha_i, D_{igt}] = 1$. Individuals are treated in different period starting at $t = 10$, such that treatment is staggered, and cohorts are indexed by g. Treatment effect is heterogeneous by time and individual: the setting gathers the conditions under which the TWFE bias arises.

I use two types of treatment heterogeneity. In the first case, heterogeneity depends on the time period t. In the second case, I introduce individual heterogeneity normally distributed across individuals, on top of time heterogeneity. Heterogeneity in treatment effect is now distributed such that the growth rate of the average outcome is different from the average growth rate of the outcome due to the treatment for each cell (g, t) Again, I introduce heteroskedasticity in the error term as a function of observables: in one case variance is function of individual fixed effects, and in the other it is a function of the treatment status. The observed outcome y_{igt} is always strictly positive such that there is no difference between estimators driven by zeros in the outcome. I simulate the data 1000 times.

5.2.2 Simulation results

Results of simulations are displayed in table 4. The first two lines of each table panel represent a different quantity of interest based on the true model. The first line recovers the exponential of the average parameter δ_{it} minus one. This quantity approximate the average growth rate of the treated outcome, but lacks interpretability if coefficients are large. It is the quantity to which log-OLS converges in the canonical setting when there is no heteroskedasticity bias. The second line displays the growth rate of the average treated outcome, which is the true treatment

Table 3: Staggered treatement timing: simulation cases

Case	$V(\eta_{it}\|.)$	Treatment effect parameter
1	α_i	$\delta_t = log(\|t - 12.5\|)$
2	$0.2D_{igt}$	$\delta_t = log(\|t - 12.5\|)$
3	α_i	$\delta_{it} = log(\|t - 12.5\|) + \nu_i, \quad \nu_i \sim \mathcal{N}(0, 0.5)$
4	$0.2D_{igt}$	$\delta_{it} = log(\|t - 12.5\|) + \nu_i, \quad \nu_i \sim \mathcal{N}(0, 0.5)$

semi-elasticity. This is the quantity recovered by the ratio-of-ratios, and the quantity to which PPML converges in the canonical setting. I compare five estimators: TWFE Log-OLS, TWFE PPML, the proposed imputation estimator, the aggregation estimator for PPML, and the log-linear estimator by Borusyak et al. (2024) robust to staggered settings. Densities of estimators are displayed in figure B1b in appendix.

The upper left panel presents the case with time constant heteroskedasticity and treatment heterogeneity by time period. Even in the absence of treatment induced heteroskedasticity, the TWFE log-OLS estimator falls behind all true quantities of interest. It is now lower than $exp(\overline{\delta_t}) - 1$ because of the staggered setting bias. I turn to the TWFE PPML estimator, which does not suffer from the heteroskedasticity bias and the treatment heterogeneity bias. The estimated coefficient is far below the true treatment semi-elasticity. This indicates that TWFE PPML suffers from a bias in settings with staggered treatment and treatment effect heterogeneity, as TWFE OLS in the linear setting.

In contrast, the imputation estimator recovers a quantity that is close up to 0.1 percentage points from the true ratio-of-ratios. In figure B1b, I observe that the kernel density of the estimator over the 1000 simulations is centered around the true growth rate of the average. The aggregation estimator is below this value, and identifies the true average parameter, such as the estimator from Borusyak et al. (2024).

In the right upper panel, I introduce treatment induced heteroskedasticity. The OLS estimator is now taking negative values for a large number of simulations, and the mean is at -0.00861. In case the true value of the parameter and the growth rate of the average are large (0.233 and 0.776), using the TWFE log-OLS estimator can lead to statistically non significant parameters

and potentially negative estimates. The imputation estimator recovers the true growth rate of the average, or treatment semi-elasticity. The aggregation estimator recovers the average parameter. The estimator from Borusyak et al. (2024) now suffers from the heteroskedasticity bias.

In the lower panels I introduce normally distributed individual heterogeneity on top of time heterogeneity. Individual treament heterogeneity is centered around zero such that the true parameter average value stays the same. In both cases, only the imputation/interaction estimators are unbiased. They recover quantities close to the true treatment semi-elasticity. In the case with individual heteroskedasticity (left panel), the precision of the estimator is reduced. The difference between log-OLS and the true growth rate is magnified in this case. TWFE PPML partly accounts for the increased growth rate but is still biased by the staggered setting.

The aggregation estimator is now different from the average parameter. When there is treatement effect heterogeneity within cohort-time cells, this estimator recovers a quantity intermediate between the average parameter and the growth rate. This is because it aggregates within cells average growth rates (semi-elasticities specific to cells g, t) accross cells. It averages estimates of true RoRs at the g, t level, weighting them by the share of cell g, t in the treated sample. The estimator from Borusyak et al. (2024) identifies its quantity of interest when there is no heteroskedasticity bias.

Table 4: 1000 simulations: staggered treatment

Heterogeneous treatment effect by time $exp(\delta_t) - 1$

	Case 1 $V(\eta_{it}\|.) = \alpha_i$				Case 2 $V(\eta_{it}\|.) = 0.2D_i$			
Estimator	Mean	St.D.	Min	Max	Mean	St.D.	Min	Max
$exp(\overline{\delta_t}) - 1$	0.233	0.00354	0.220	0.247	0.233	0.00254	0.226	0.240
True RoR	0.776	0.0133	0.734	0.824	0.776	0.00941	0.748	0.809
Imputation	0.780	0.0879	0.398	1.096	0.775	0.0550	0.559	0.954
Aggregation	0.234	0.0471	0.0721	0.395	0.232	0.0291	0.141	0.332
Robust log-OLS	0.233	0.0161	0.178	0.282	0.170	0.0121	0.128	0.203
TWFE PPML	0.198	0.0567	0.0166	0.470	0.194	0.0360	0.0790	0.344
TWFE Log-OLS	0.0452	0.0125	-0.000729	0.0877	-0.00861	0.00968	-0.0364	0.0215

Heterogeneous treatment effect by time and individuals $exp(\delta_{it}) - 1$

	Case 3 $V(\eta_{it}\|.) = \alpha_i$				Case 4 $V(\eta_{it}\|.) = 0.2D_i$			
Estimator	Mean	St.D.	Min	Max	Mean	St.D.	Min	Max
$exp(\overline{\delta_{it}}) - 1$	0.233	0.00592	0.216	0.251	0.233	0.00440	0.221	0.247
True RoR	1.011	0.0307	0.909	1.113	1.013	0.0228	0.939	1.112
Imputation	1.010	0.102	0.432	1.359	1.016	0.0675	0.770	1.228
Aggregation	0.394	0.0529	0.185	0.592	0.397	0.0345	0.274	0.494
Robust log-OLS	0.233	0.0174	0.178	0.295	0.169	0.0146	0.121	0.217
TWFE PPML	0.340	0.0693	0.121	0.703	0.340	0.0472	0.164	0.499
TWFE Log-OLS	0.0448	0.0130	0.00833	0.0951	-0.00802	0.01000	-0.0350	0.0279

In figure 1, I recover the dynamic of treatment versions of the various estimators in one simulation. All coefficients are expressed relative to $t-1$ to avoid as much as possible differences in interpretation (Roth, 2024).[6] I compare their ability to recover the dynamic of the quantities of interest. I study them in the full heterogeneity case, with treatment induced heteroskedasticity and heterogeneous treatment effect accross time and individuals (Case 4 of table 4). I estimate leads and lags for the TWFE log-OLS and PPML, the imputation and interaction estimators and the aggregation estimator. To derive more easily confidence intervals, I plot $log(PTT + 1)$ for

[6] I do not include the estimator from Borusyak et al. (2024) now as it as a different interpretation on leads.

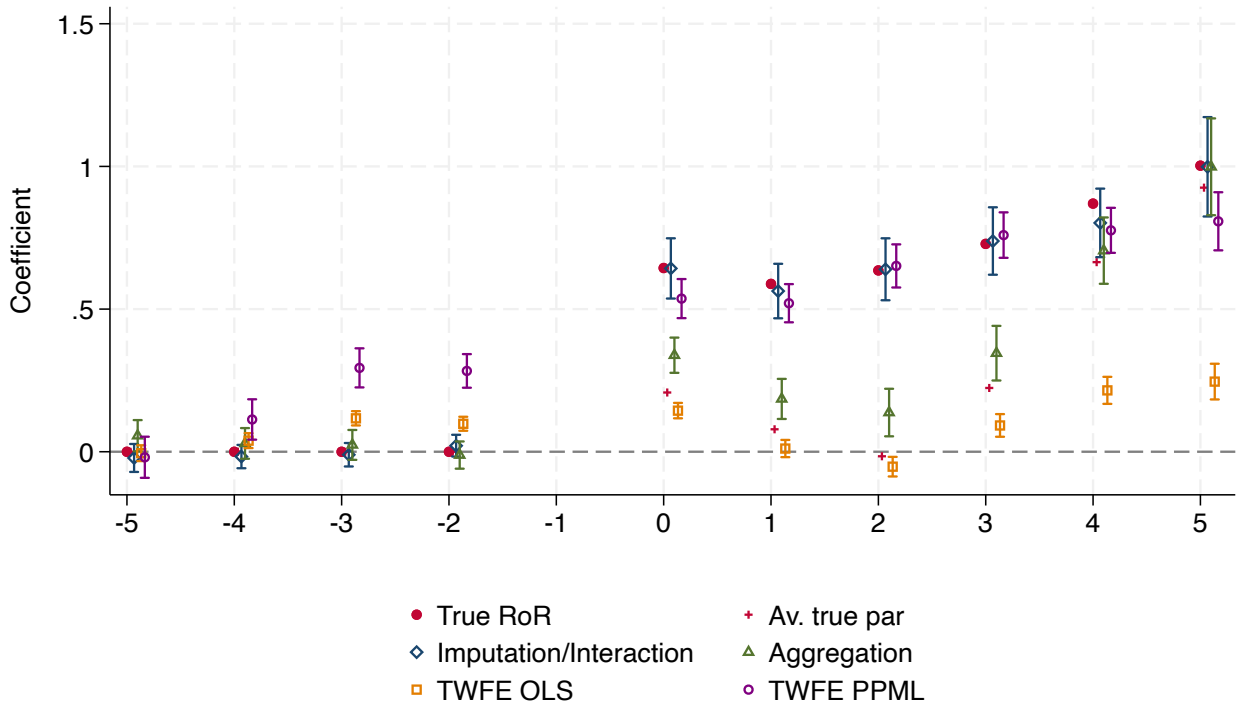
each estimator, which corresponds to $\widehat{\delta}$ in the case of the TWFE log-OLS estimator. Red markers are set for the true value of the parameter and the RoR.

As expected, the imputation estimator is an unbiased estimator of dynamic percentage changes in the average. It closely matches the true RoR, the quantity to which the PPML estimator converges in the canonical setting. The TWFE PPML estimator is biased donward for treatment effects at $t = 0$ and $t = 1$. For following periods, it is close to the true RoR. TWFE log-OLS strongly underestimate the RoR, and cannot recover the true parameter, except for $t = 0, 1, 2$. The aggregation parameter recovers an intermediate quantity between the average parameter and the true treatment semi-elasticity (RoR). It converges to the true RoR in the later period, when there are less treated cohorts: in $t = 5$, when only the first cohort is treated, it computes the same quantity as the imputation estimator, because it covers only one (g, t) cell now.

The imputation and aggregation estimators display close to zero and non statistically significative coefficients for leads. Coefficients on the leads are more worrying for TWFE estimators. But both TWFE PPML and TWFE log-OLS display false positive coefficients on pre-trend. This strongly discourages the use of these estimators, as they might lead to suspect existing pre-trends in settings with staggered treatment, by displaying false positive coefficients.

Figures B2, B3 and B4 in the appendix display event studies for unique simulations generated in the other cases. In cases 1 and 2, when there is no heterogeneity in treatement effect within cohort-time cells (in simulations, driven by individual treatment heterogeneity), the aggregation estimator correctly identifies the average true parameter. In case 3 it computes an intermediate quantity in between the average true parameter. TWFE log-OLS always fails to identify the average true parameter, because of either the heteroskedasticity bias or the staggered treatment timing bias. Estimator from Borusyak et al. (2024) correct for this bias and fails with treatment incuded heteroskedasticity. TWFE PPML displays large false positive pre-trends and diverges from the true growth rate. Only the imputation estimator correctly identifies the true growth rate before and after the policy change.

Figure 1: Case 4: Event study



Note: 95% confidence intervals. Case 4: Heteroskedasticity function of treatment status, no individual treatment effect heterogeneity. To ease the derivation of confidence intervals, I plot $log(PTT + 1)$ for each estimator.

6 Application

6.1 Set-up

I apply my estimator to the empirical question of Menkhoff and Miethe (2019) and Johannesen and Zucman (2014). The authors investigate whether the G20 crackdown on tax havens in 2009 reduced bank deposits that individuals held offshore to avoid taxes.

Following this summit, many tax havens were compelled to sign bilateral treaties implementing automatic exchange of information on bank account holders. These treaties, signed for example between France and Switzerland in 2009, make it mandatory for banks in both countries to report accounts held by each others' citizens to the tax authorities of their home countries. The signature and implementation of treaties vary across country pairs. Using data from the Bank of International Settlements (BIS) from 2003 to 2011 Menkhoff and Miethe (2019) and Johannesen and Zucman (2014) explore whether a treaty signed between a tax haven and another country reduces deposits held by citizens of the home country in the tax haven. This is likely to be the case if those deposits are held for tax or regulation evasion purpose.

I replicate the findings of Menkhoff and Miethe (2019) while the replication of Johannesen and Zucman (2014) is available in Appendix. The two sets of authors use the same identification strategy, but Menkhoff and Miethe (2019), use a more conservative definition of treatment, benefiting from a few additional years of perspective: they only consider new TIEAs and DTCs implementing the OECD's banking transparency standards. Moreover the BIS sample data they use is publicly available, contrary to Johannesen and Zucman (2014) who had access to extended confidential data, facilitating replication and comparison of results.

The authors estimate the following model:

$$\log(Deposit_{ijq}) = \alpha + \beta Signed_{ijq} + \gamma_{ij} + \theta_q + \epsilon_{ijq} \tag{41}$$

With $Deposit_{ijq}$ the deposits held by citizens of country i in tax haven j at time q. The treatment variable $Signed_{ijq}$ takes the value one when a treaty is signed between i and j at time q. Fixed effects for country pairs γ_{ij} and time θ_q are included. The authors use a two-way fixed effect log-linearized model, using as a control group all non-haven to haven dyads which did not sign a treaty during the time frame under study. In the following section, I present results of log-linear and non-linear TWFE estimators, and estimators of models allowing for full heterogeneity in the β coefficient across cohorts and month.

6.2 Results

I estimate equation (41) with five different estimators: the TWFE log-OLS estimator, the linear estimator from Borusyak et al. (2024), the TWFE PPML estimator, the proposed imputation estimator, and the aggregation estimator similar to the strategy of Nagengast and Yotov (2023).

Results are displayed in Table 5. Column (1) presents the replication of Menkhoff and Miethe (2019) results using their methodology. On average, the signature of a treaty reduced deposits held in the partner tax havens by 31.9%.[7] In column (2), I restrict the sample and remove the few country-pairs that are always treated to avoid forbidden comparisons. The results remain. In column (3), I use the estimator from Borusyak et al. (2024) to recover difference-in-difference estimates corrected for any staggered treatment bias. The effect is slightly bigger than before, indicating that the staggered treatment bias upward the treatment effect estimate of TWFE.

Columns (4), (5), and (6) display the non-linear estimations. In column (5), the TWFE PPML estimate that treaties signed decreased deposits held in tax havens by 13.2%.[8] Interpreting the quantity of interest, the signature of bilateral treaties of automatic exchange of information reduced the average volume of deposits held in tax havens. Column (6) implements my proposed estimator, for a same quantity of interest, and robust to staggered bias: it recovers a drop in deposits by 16.5%. The comparison of columns (5)-(6) points to an upward bias because staggered treatment, as columns (2)-(3).

There is a large difference between the results derived from the difference-in-differences estimator (OLS) and the ratio-of-ratios (PPML). The difference between column (1) to (3) and (5)-(6) comes from the fact that the difference-in-difference estimator approximates an average effect, while the ratio-of-ratios recovers the effect of treatment on the average, which differ for non-linear DGPs. The average effect of treaties across country pairs is larger that the effect of the set of treaties on average deposits held in tax havens.

This empirical pattern can be understood by looking at the joint distribution of treatment effects and deposit volumes across treatment cells. In figure C1 in the appendix, I display the raw coefficients from the full interaction model (see equation 31). Each coefficient recovers the treatment effect on the average for cell (ij, q). Cohorts (country-pairs ij treated at the same time) are displayed in the same color. We observe that even though cells (ij, q) display a large

[7] $(exp(-0.384) - 1) \times 100 \approx -31.9$

[8] $(exp(-0.141) - 1) \times 100 \approx -13.2$

negative treatment effect *on average*, most of the cells exhibiting the strongest effects are small country-pairs in term of volume of tax haven deposits held. On the contrary, there are some cohorts exhibiting at the same time a weak or positive treatment effect, and a large volume of tax haven deposits, explaining the lower ratio-of-ratios, or change *in the average*.

The result of column (4) goes further in reconciling both results by showing that the aggregation estimator lies between them. It displays the estimate from an aggregation estimator used in Nagengast and Yotov (2023). Without treatment-related heteroskedasticity and under limited treatment effect heterogeneity, this estimator recovers a quantity in between the DiD and RoR: the average parameter. Its semi-elasticity interpretation lies in between these quantities of interest. In column (4), I estimate that on average, when a set of countries sign new treaties with some tax havens on the same month, their deposits held in these tax haven drop by 23.9% (average change of the averages).[9]

Imputation (6) and aggregation (4) estimators both recover treatment effect robusts to staggered treatment timing but aggregate heterogeneous treatment effect in different ways. Recovering the treatment effect on the average, the imputation estimator, just as the PPML estimator, gives a bigger weight to treatment effect in units that affect more the average: country pairs with bigger deposit volumes. The aggregation estimator weights cohorts differently. In this setting, the interpretation of the aggregation estimator is different than the average change, and is dependent on the time structure of the treatment.

7 Conclusion

This paper reconciles two significant empirical issues encountered by applied economists when estimating treatment effects in non-linear models, using difference-in-differences methodologies. First, the log-OLS estimator is biased in the presence of heteroskedasticity and treatment-induced changes in outcome variance. Even in the absence of bias, the recovered estimate might not be the researcher's prefer quantity in presence of large treatment effect heterogeneity. Second, the traditional two-way fixed effects estimators do not accurately recover difference-in-differences estimates when treatment timing is staggered and effect is heterogeneous.

To reconcile both issues, I propose a novel estimator that recovers of a proportional treatment

[9] $(exp(-0.273) - 1) \times 100 \approx -23.9$

Table 5: Staggered treatment robustness: Static estimator

	Linear estimators			Non-linear estimators		
	TWFE OLS (replication)	TWFE OLS	Borusyak et al. (2024)	Aggregation	TWFE PPML	Imputation
	(1)	(2)	(3)	(4)	(5)	(6)
Coef	-0.384***	-0.383***	-0.402***	-0.273**	-0.141**	-0.180**
S.e.	(0.09)	(0.09)	(0.074)	(0.11)	(0.078)	(0.091)
N	17267	16244	16244	16244	16244	16244
Control group	All	Never treated & Not yet treated				
Country-pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes

Column (1): Replication of Menkhoff and Miethe (2019). Standard errors adjusted for clustering by country-pairs. Standard errors for the imputation and aggregation estimators are computed through 500 bootstrap replications. No control variables included.

effect (semi-elasticity) even in cases of staggered treatment timing and heterogeneous treatment effects. Leveraging interpretation of the TWFE PPML estimator in the canonical 2x2 setting, I develop an approach that accurately estimates the ratio-of-ratios, ensuring an interpretable treatment effect estimates similar to the canonical setting. The specified model can account for any kind of heterogeneity in the treatment effect, under parallel trend and no anticipation assumption. Moreover, it can account for a parallel trend assumption conditional on some covariates.

Through empirical validation and simulations, I compare the proposed estimator compared to existing approaches. From simulations in with staggered treatment timing and heterogeneous treatment effects, it appears that the interaction estimator proposed in this paper is the most suited to recover the correct treatment change in the average. In all studied cases, its density is centered around the true ratio-of-ratios, the quantity of interest of the 2-by-2 canonical setting. It is also robust to any type of treatment effect heterogeneity. The aggregation estimator used seems more suited to recover the average parameter, and performs better that Borusyak et al. (2024) when there is a treatment induced heteroskedasticity, but requires to make assumptions on the structure of treatment heterogeneity for interpretation. It is up to the empirical researcher to think about what is her preferred quantity to recover. The use of TWFE log-OLS and TWFE PPML is strongly discouraged in this setting, with the former potentially yielding negative estimates when the true treatment effect is positive and of a large magnitude.

I apply my estimator to the empirical question of Johannesen and Zucman (2014) and Menkhoff and Miethe (2019). The authors investigate wether bilateral treaties of automatic exchange of information decreased deposits held in tax havens' banks. I show that their results are robust to correcting for staggered treatment and using non-linear estimator. I show that using a TWFE PPML estimator in their set-up provides an lower treatment effect estimate, which can be rationalized by the fact that it aggregates differently the strong heterogeneity in treaties effects accross country pairs. My proposed estimator recover a close estimate, showing that eventhough the treaties on average decreased deposits held offshore in tax havens, the average volume of deposits held in tax havens changed by a lower magnitude. Furthermore, by applying the proposed estimator to the empirical question of cross-border deposit behavior in response to automatic exchange of information treaties, I showcase its practical relevance to answer empirical questions.

This paper contributes to the understanding of multiplicative difference-in-differences models and the interpretation of semi-elasticities and provides a valuable tool for researchers grappling with heterogeneous treatment effects and staggered treatment timings. In future research, further exploration of the precision of this estimator would be valuable. In particular, derivating analytical formulas for the variance of estimated parameters would represent substantial precision gains for empirical researchers.

References

Angrist, J. D. (2001). Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. *Journal of business & economic statistics*, 19(1):2–28.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event study designs: Robust and efficient estimation. *Review of Economic Studies*.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

Chen, J. and Roth, J. (2023). Logs with Zeros? Some Problems and Solutions*. *The Quarterly Journal of Economics*, 139(2):891–936.

Ciani, E. and Fisher, P. (2019). Dif-in-dif estimators of multiplicative treatment effects. *Journal of Econometric Methods*, 8(1):20160011.

Cohn, J., Liu, Z., and Wardlaw, M. (2022). Count (and count-like) data in finance. *Journal of Financial Economics*.

Correia, S., Guimarães, P., and Zylkin, T. (2020). Fast poisson estimation with high-dimensional fixed effects. *The Stata Journal: Promoting communications on statistics and Stata*, 20(1):95–115.

De Chaisemartin, C. and d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96.

De Chaisemartin, C. and D'Haultfoeuille, X. (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *Econometrics Journal*.

Jan van Garderen, K. and Shah, C. (2002). Exact interpretation of dummy variables in semilogarithmic equations. *The Econometrics Journal*, 5(1):149–159.

Johannesen, N. and Zucman, G. (2014). The end of bank secrecy? an evaluation of the g20 tax haven crackdown. *American Economic Journal: Economic Policy*, 6(1):65–91.

Kennedy, P. E. (1981). Estimation with correctly interpreted dummy variables in semi logarithmic equations. *American Economic Review*.

Menkhoff, L. and Miethe, J. (2019). Tax evasion in new disguise? examining tax havens' international bank deposits. *Journal of Public Economics*, 176:53–78.

Nagengast, A. and Yotov, Y. (2023). Staggered difference-in-differences in gravity settings: Revisiting the effects of trade agreements. *Deutsche Bundesbank Discussion Paper*.

Olden, A. and Møen, J. (2022). The triple difference estimator. *The Econometrics Journal*, 25(3):531–553.

Roth, J. (2024). Interpreting event-studies from recent difference-in-differences methods. *arXiv preprint arXiv:2401.12309*.

Silva, J. S. and Tenreyro, S. (2006). The log of gravity. *The Review of Economics and statistics*, 88(4):641–658.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

Wooldridge, J. (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Available at SSRN 3906345*.

Wooldridge, J. M. (2023). Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal*.

Appendix

A. Supplementary results

Heteroskedasticity bias Silva and Tenreyro (2006) explain that using log-OLS models to estimate multiplicative economic relationship will produce biased estimates if there is heteroskedasticity in the multiplicative error term driven by the explanatory variables.

Cohn et al. (2022) derive more explicitely the relationship between patterns of heteroskedasticity and the sign of the bias. They take the following multiplicative model as an illustration case:

$$y = e^{x\beta}\eta$$

With x a covariate and η a multiplicative error term with conditional expectation equal to one. If x if normal with mean 0 and η is log-normal with standard deviation $\sigma_\eta(x) = exp(\delta x)$, we have:

$$\widehat{\beta_{OLS}} = \beta - \frac{\delta}{2}$$

If the variance of the error term increases with x (δ>0), the estimate is downward biased, and vice versa. If $\delta/\beta > 2$, the estimate is of the wrong sign. Note that the distorsion of variance linked to x must be big enough to generate this bias.

More generally, authors derive that:

$$\frac{\partial E[log(y)|x]}{\partial x_j} = \beta_j + \frac{\partial E[log(\eta)|x]}{\partial x_j} \tag{42}$$

Maximum likelihood The TWFE PPML estimation by maximum of log-likelihood implies the following first order conditions:

$$\begin{cases} \sum_{i,t,D_{it}=1}(y_{it} - \hat{y_{it}}) = 0 \\ \sum_{i=j,t}(y_{jt} - \hat{y_{jt}}) = 0 \\ \sum_{i,t=l}(y_{il} - \hat{y_{il}}) = 0 \end{cases}$$

Table A1: Staggered setting - Simple example

$\mathbb{E}[Y_{it}]$	$i = A$	$i = B$
$t = 1$	$exp(\alpha_A)$	$exp(\alpha_B)$
$t = 2$	$exp(\alpha_A + \beta_2 + \delta_{A2})$	$exp(\alpha_B + \beta_2)$
$t = 3$	$exp(\alpha_A + \beta_3 + \delta_{A3})$	$exp(\alpha_B + \beta_3 + \delta_{B3})$

Which translates in this system of equation in the simple case in table A1:

$$\begin{cases} y_{A2} + y_{A3} + y_{B3} = exp(\hat{\alpha_A} + \hat{\beta}_2 + \hat{\delta}) + exp(\hat{\alpha_A} + \hat{\beta}_3 + \hat{\delta}) + exp(\hat{\alpha_B} + \hat{\beta}_3 + \hat{\delta}) \\ y_{A1} + y_{A2} + y_{A3} = exp(\hat{\alpha_A}) + exp(\hat{\alpha_A} + \hat{\beta}_2 + \hat{\delta}) + exp(\hat{\alpha_A} + \hat{\beta}_3 + \hat{\delta}) \\ y_{B1} + y_{B2} + y_{B3} = exp(\hat{\alpha_B}) + exp(\hat{\alpha_B} + \hat{\beta}_2) + exp(\hat{\alpha_B} + \hat{\beta}_3 + \hat{\delta}) \\ y_{A1} + y_{B1} = exp(\hat{\alpha_A}) + exp(\hat{\alpha_B}) \\ y_{A2} + y_{B2} = exp(\hat{\alpha_A} + \hat{\beta}_2 + \hat{\delta}) + exp(\hat{\alpha_B} + \hat{\beta}_2) \\ y_{A3} + y_{B3} = exp(\hat{\alpha_A} + \hat{\beta}_3 + \hat{\delta}) + exp(\hat{\alpha_B} + \hat{\beta}_3 + \hat{\delta}) \end{cases}$$

This yields:

$$\begin{cases} exp(\hat{\beta}_2 + \hat{\delta}) = \frac{Y_{A2}}{exp(\hat{\alpha_A})} \\ exp(\hat{\alpha_A}) = \frac{(y_{A1}+y_{A3})\times(y_{A1}+y_{B1})}{y_{A1}+y_{B1}+y_{A3}+y_{B3}} \\ exp(\hat{\alpha_B}) = \frac{(y_{B1}+y_{B3})\times(y_{A1}+y_{B1})}{y_{A1}+y_{B1}+y_{A3}+y_{B3}} \\ y_{A1} + y_{B1} = exp(\hat{\alpha_A}) + exp(\hat{\alpha_B}) \\ y_{A2} + y_{B2} = exp(\hat{\alpha_A} + \hat{\beta}_2 + \hat{\delta}) + exp(\hat{\alpha_B} + \hat{\beta}_2) \\ exp(\hat{\beta}_3) = \frac{y_{A3}+y_{B3}}{(y_{A1}+y_{B1})exp(\hat{\delta})} \end{cases}$$

Quantity of interest If we have, heterogeneous treatment effect $\delta_{A2} \neq \delta_{A3} \neq \delta_{B3}$. The quantity of interest is then:

$$
\begin{aligned}
PTT &= \frac{E[y_{it}(1)|D=1] - E[y_{it}(0)|D=1]}{E[y_{it}(0)|D=1]} \\
&= \frac{(1/3)\Big(E(y_{A2}(1)) + E(y_{A3}(1)) + E(y_{B3}(1))\Big) - (1/3)\Big(E(y_{A2}(0)) + E(y_{A3}(0)) + E(y_{B3}(0))\Big)}{(1/3)\Big(E(y_{A2}(0)) + E(y_{A3}(0)) + E(y_{B3}(0))\Big)} \\
&= \frac{\Big(exp(\alpha_A + \beta_2 + \delta_{A2}) + exp(\alpha_A + \beta_3 + \delta_{A3}) + exp(\alpha_B + \beta_3 + \delta_{B3})\Big)}{\Big(exp(\alpha_A + \beta_2) + exp(\alpha_A + \beta_3) + exp(\alpha_B + \beta_3)\Big)} \\
&\quad - \frac{\Big(exp(\alpha_A + \beta_2) + exp(\alpha_A + \beta_3) + exp(\alpha_B + \beta_3)\Big)}{\Big(exp(\alpha_A + \beta_2) + exp(\alpha_A + \beta_3) + exp(\alpha_B + \beta_3)\Big)} \\
&= \sum_{i,t,D_{it}=1} \frac{E(y_{it}(0))}{\sum_{i,t,D_{it}=1} E(y_{it}(0))}\Big(exp(\delta_{it}) - 1\Big) = \sum_{i,t,D_{it}=1} \omega_{it}\Big(exp(\delta_{it}) - 1\Big)
\end{aligned}
\tag{43}
$$

Equivalence of the imputation and saturated approaches Wooldridge (2023) proposes to recover estimates of ATTs in level which converges to $\tau_{rs} = y_{rs}(1) - y_{rs}(0)$ by predicting the average partial effect of the treatment variable D_{it} over the desired treated sample, evaluated for the right value of cohort and time dummies. For time period and cohorts r, s, it computes:

$$
\begin{aligned}
\widehat{\tau}_{inter,rs} =& E(\widehat{y}|D_{it}=1, g_{is}=1, f_{s_t}=1, \forall (k,l) \neq (r,s)\, g_{ik}=0; f_{l_t}=0) \\
&- E(\widehat{y}|D_{it}=0, g_{is}=1, f_{s_t}=1, \forall (k,l) \neq (r,s)\, g_{ik}=0; f_{l_t}=0) \\
=& N_{rs}^{-1} \sum_{i=1}^{N} D_{irs}[exp(\widehat{\alpha_i} + \widehat{\beta t} + \widehat{\delta_{rs}}) - exp(\widehat{\alpha_i} + \widehat{\beta t})]
\end{aligned}
$$

With N_{rs} the number of observations for cohort r at time s and D_{irs} an indicator variable if the observation belongs to cohort r observed at time s. Again, I can re-write the model to compute the average partial effect across the entire treated sample with $\mathbb{1}\{t - r = l\}$ to get the ATT l time periods after treatment:

$$
\widehat{\tau}_{inter,rl} = N_{rl}^{-1} \sum_{i=1}^{N} D_{irl}[exp(\widehat{\alpha_i} + \widehat{\beta t} + \widehat{\delta_{rl}}) - exp(\widehat{\alpha_i} + \widehat{\beta t})] \tag{44}
$$

Interestingly, Wooldridge (2023) notes that this quantity is numerically equivalent to the imputation estimator from equation (36) on the same sample. It also has the advantage to have known analytical expressions for standard errors. As in the previously section, I propose to

scale this quantity by the predicted counterfactual outcome in the absence of treatment on the same subsample:

$$\widehat{RoR}_{inter} = \frac{\widehat{\tau}_{inter}}{\sum_{i\in\omega_1} \widehat{y_{igt}(0)}} = \frac{\widehat{\tau}_{imput}}{\sum_{i\in\omega_1} \widehat{y_{igt}(0)}} = \widehat{RoR}_{imput} \tag{45}$$

In the case with group specific parallel trends, the fully saturated model should write:

$$\begin{aligned} E[y_{ipt}|g_{iq},...,g_{iT}] = exp\Big[& \sum_{g=q}^{T} \beta_g g_{ig} + \sum_{s=2}^{T} \gamma_s fs_t + \sum_{p=1}^{P} \mathbb{1}\{c=p\}\kappa_p + \sum_{p=1}^{P}\sum_{s=2}^{T} (fs_t \times \mathbb{1}\{c=p\})\pi_{pt} \\ & + \sum_{g=q}^{T}\sum_{l=0}^{T-r} (D_{it} \times g_{ig} \times \mathbb{1}\{t-g=l\})\delta_{gl} \\ & + \sum_{p=1}^{P}\sum_{g=q}^{T}\sum_{l=0}^{T-r} \mathbb{1}\{c=p\} \times (D_{it} \times g_{ig} \times \mathbb{1}\{t-g=l\})\zeta_{pgl} \\ & + \alpha_i + \beta_{ct}\Big] \end{aligned}$$

Coefficients γ_s and π_{pt} control for the drug-specific parallel trend. Coefficients δ_{gs} and ζ_{pgs} control for the full heterogeneity of the treatment effect, by cohort, drugs and time. Variables $g_{ig}, fs_t, \mathbb{1}\{c=s\}, \mathbb{1}\{c=s\} \times fs_t$ are dropped because they are colinear with fixed effects α_i and β_{ct}, and we are left with the model to estimate:

$$\begin{aligned} E[Y_{ipt}|g_{iq},...,g_{iT}] = exp\Big[& \sum_{g=q}^{T}\sum_{l=0}^{T-r} (D_{it} \times g_{ig} \times \mathbb{1}\{t-g=l\})\delta_{gl} \\ & + \sum_{p=1}^{P}\sum_{g=q}^{T}\sum_{l=0}^{T-r} \mathbb{1}\{c=p\} \times (D_{it} \times g_{ig} \times \mathbb{1}\{t-g=l\})\zeta_{pgl} + \alpha_i + \beta_{ct}\Big] \end{aligned} \tag{46}$$

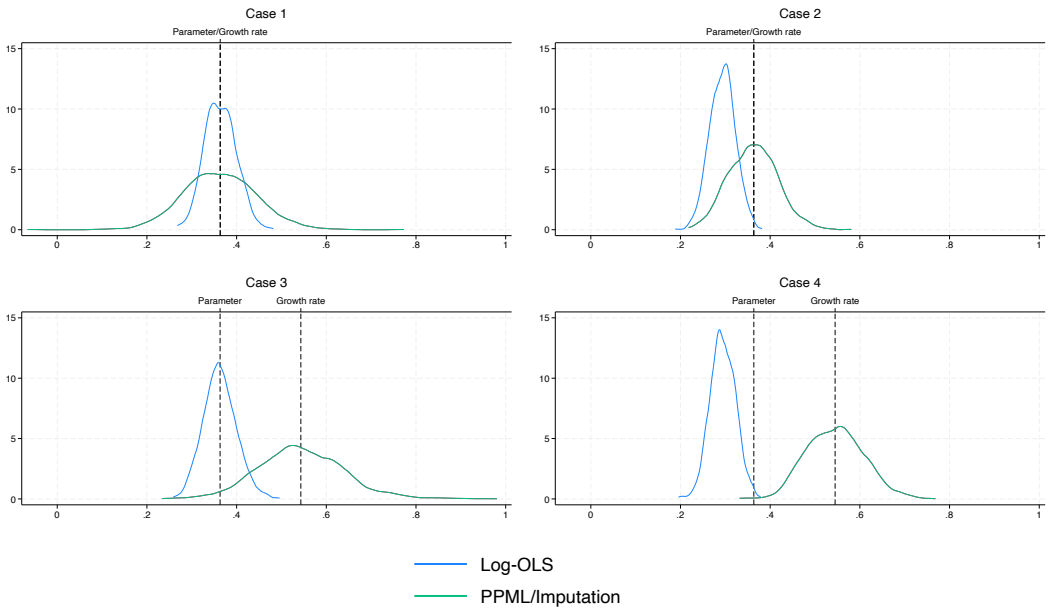
The average treatment effect ATT_{inter} is estimated as in the simple case, by predicting the treatement average partial effect on the treated sample. As before, the proportional treatment effect estimate is:

$$\widehat{RoR}_{inter} = \frac{\widehat{\tau}_{inter}}{\sum_{i\in\omega_1} \widehat{y_{igt}(0)}}$$

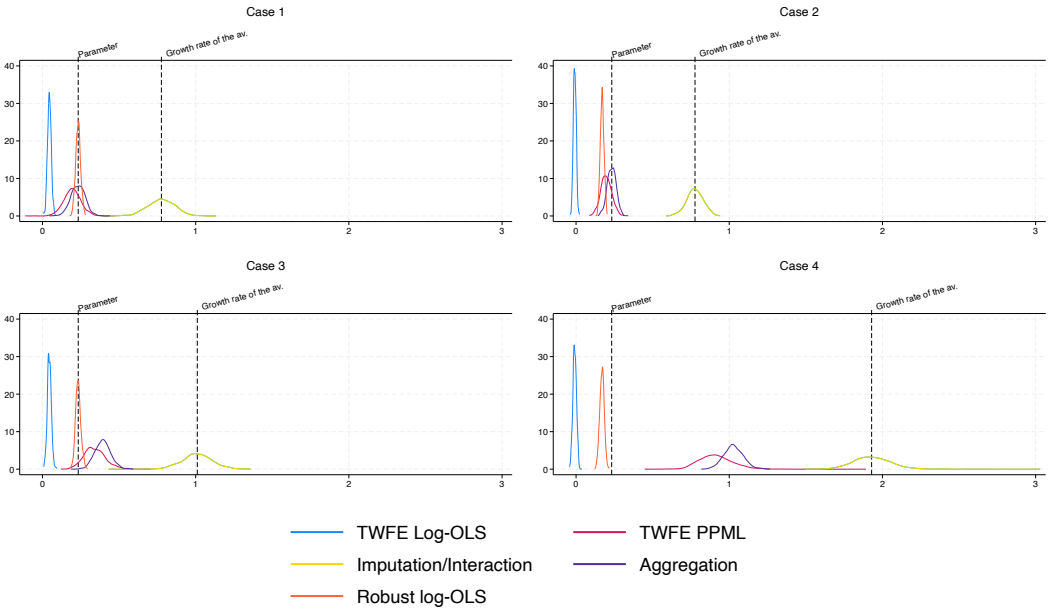
B. Simulations

Figure B1: Simulations: density

(a) Common treatment timing

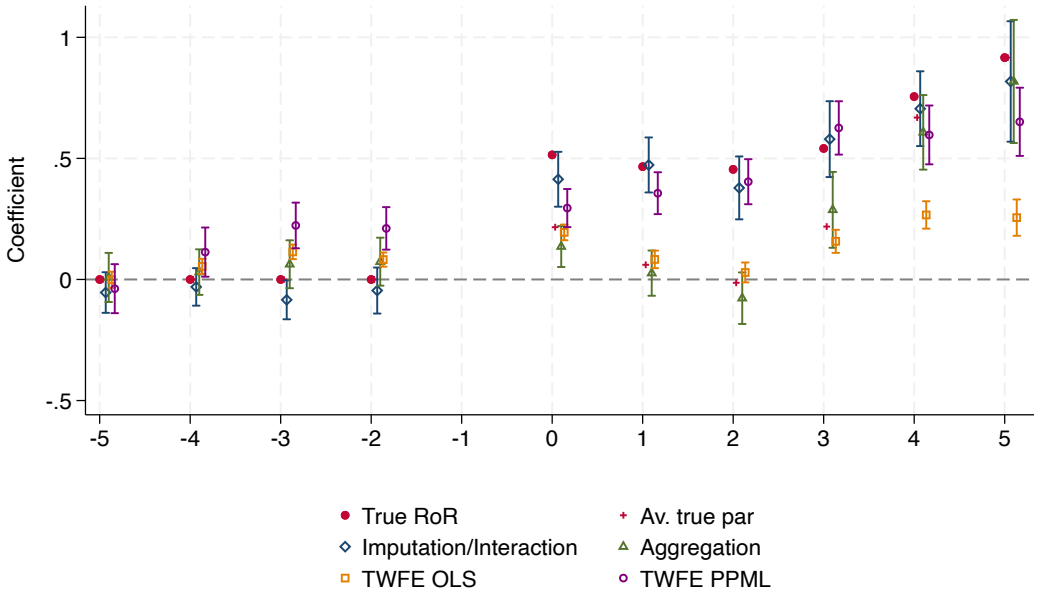


(b) Staggered treatment timing



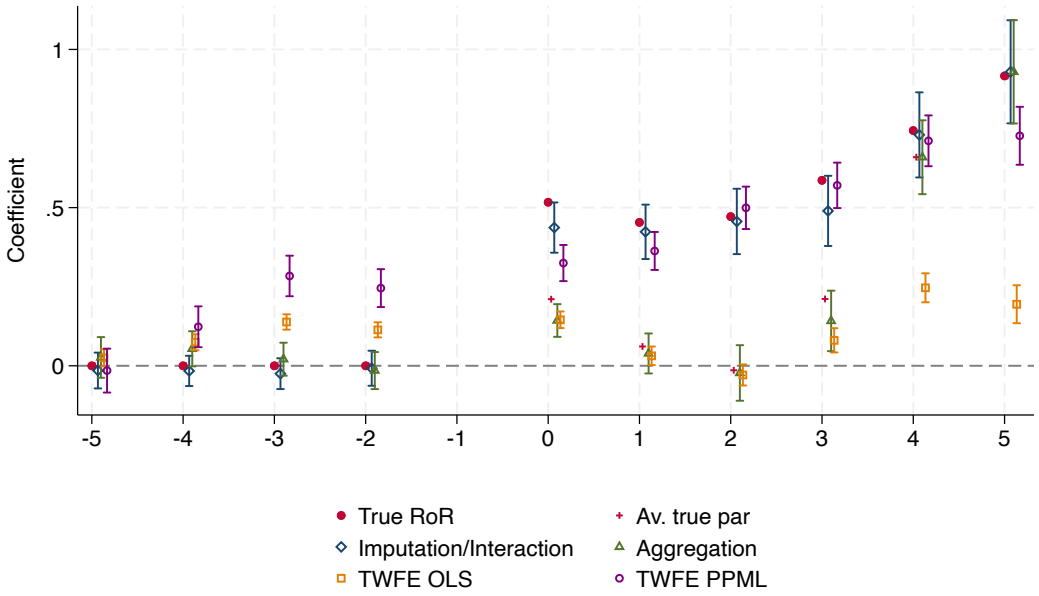
Note: Case 1: No heteroskedasticity function of treatment status, no individual treatment effect heterogeneity. Case 2: Heteroskedasticity function of treatment status, no individual treatment effect heterogeneity. Case 3: No heteroskedasticity function of treatment status, individual treatment effect heterogeneity. Case 4: Heteroskedasticity function of treatment status, no individual treatment effect heterogeneity. Robust log-OLS from Borusyak et al. (2024).

Figure B2: Case 1: Event study



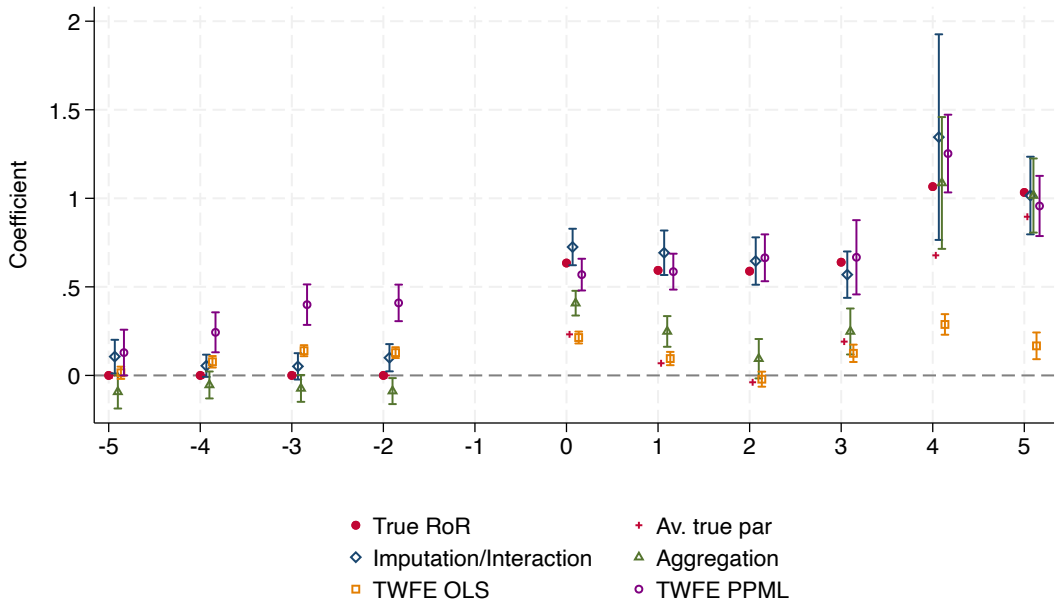
Note: 95% confidence intervals. Case 1: No heteroskedasticity function of treatment status, no individual treatment effect heterogeneity. To ease the derivation of confidence intervals, I plot $log(PTT + 1)$ for each estimator.

Figure B3: Case 2: Event study



Note: 95% confidence intervals. Case 2: Heteroskedasticity function of treatment status, no individual treatment effect heterogeneity. To ease the derivation of confidence intervals, I plot $log(PTT + 1)$ for each estimator.

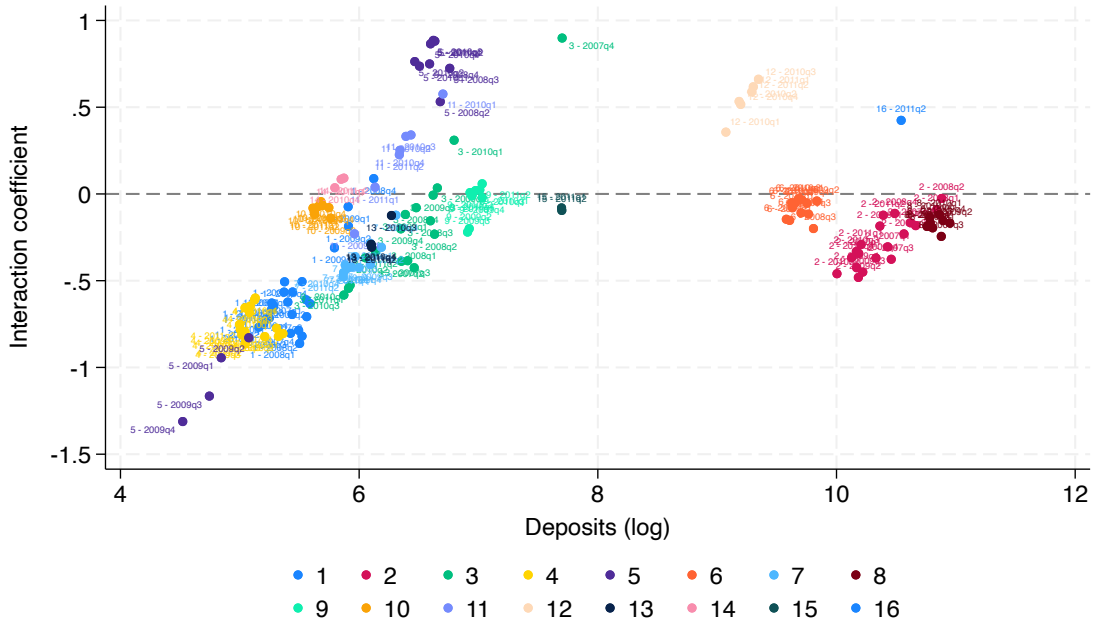
Figure B4: Case 3: Event study



Note: 95% confidence intervals. Case 3: No heteroskedasticity function of treatment status, individual treatment effect heterogeneity. To ease the derivation of confidence intervals, I plot $log(PTT + 1)$ for each estimator.

A C. Application

Figure C1: Interaction coefficients (Menkhoff and Miethe, 2019)



Note: Colors in the legend correspond to the different treated cohorts. Each dot correspond to a coefficient of the interaction from the aggregation estimator.

Table C1 presents the replication results of Johannesen and Zucman (2014), which includes TIEAs, new double tax conventions (DTCs) and domestic legislation changes in tax havens that enable information exchange through existing treaties. The treatment definition is less conservative than Menkhoff and Miethe (2019), and their objective was to evaluate all legal changes endorsed by the OECD that could have generated a decrease in bank secrecy.

Following the author's methodology in column (1), on average, a legl change in banking secrecy reduced deposits held in the partner tax havens by 12.4%.[10] This treatment effect is lower than the one estimated by Menkhoff and Miethe (2019), because the treatment definition is less conservative due to available information at the time of the study. In column (2), removing the few country-pairs that are always treated, the results remain, but slightly less precise. In column (3), the estimator from Borusyak et al. (2024) recovers an effect is slightly bigger than before and more precise. In column (4), the aggregation estimator used in Nagengast and Yotov

[10] $(exp(-0.133) - 1) \times 100 \approx -12.4$

(2023) estimates that the signature of a treaty reduces by 12.3% the deposit held in a partner tax haven.[11]

In column (5), the TWFE PPML estimate is close to zero and not statistically significant. Interpreting the results, the signature of bilateral treaties of automatic exchange of information did not reduce the average volume of deposits held in tax havens. Column (6) implements my proposed estimator and recovers a result close to column (4). Column (3) indicates that there is a negative effect of legal changes regarding bank secrecy on deposits in tax havens on average across country pairs, whereas column (6) indicates that the average volume of deposits held in tax havens was not reduced by these OECD endorsed legal changes in the sample. Figure C2 decomposes the cohort-time treatment effects of the non-linear model.

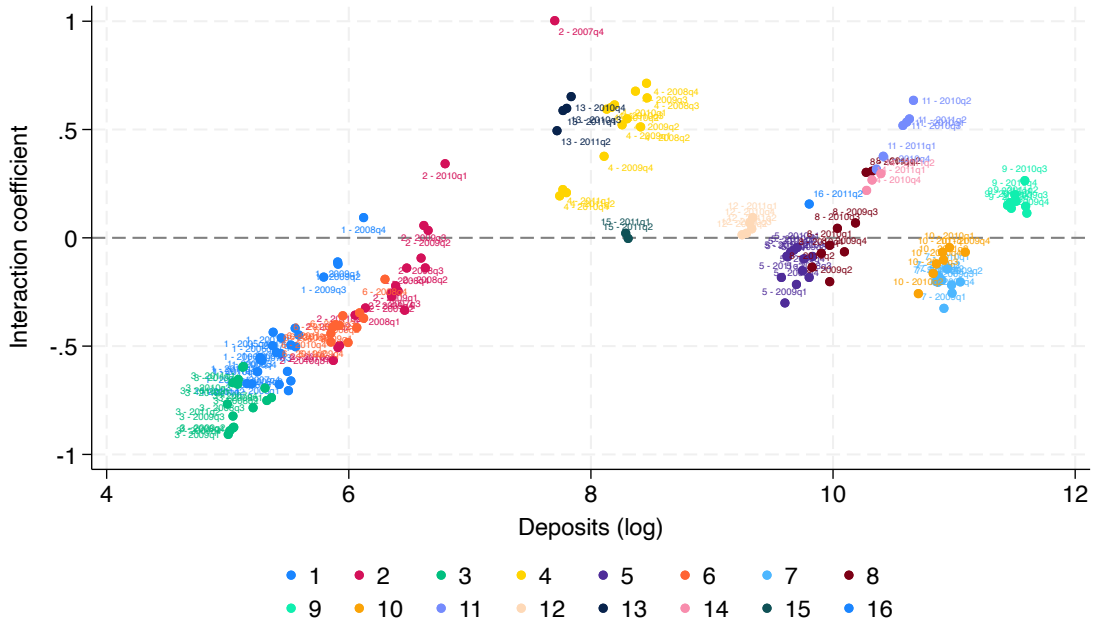
Table C1: Staggered treatment robustness: Static estimator

	Linear estimators			Non-linear estimators		
	TWFE OLS (replication)	TWFE OLS	Borusyak et al. (2024)	Aggregation	TWFE PPML	Imputation
	(1)	(2)	(3)	(4)	(5)	(6)
Coef	-0.133**	-0.129*	-0.158***	-0.131**	0.042	0.057
S.e.	(0.061)	(0.061)	(0.054)	(0.065)	(0.067)	(0.073)
N	16523	16430	16430	16430	16430	16430
Control group	All	Never treated & Not yet treated				
Country-pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes

Column (1): Replication of Johannesen and Zucman (2014) using data from Menkhoff and Miethe (2019), differences with original results can arise from differences in underlying data. Standard errors adjusted for clustering by country-pairs. Standard errors for the imputation and aggregation estimators are computed through 500 bootstrap replications. No control variables included.

[11] $(exp(-0.131) - 1) \times 100 \approx -12.3$

Figure C2: Interaction coefficients Johannesen and Zucman (2014)



Note: Colors in the legend correspond to the different treated cohorts. Each dot correspond to a coefficient of the interaction from the aggregation estimator.