WORKING PAPERS

N° 17-824

June 2017

"Should We Combine Difference In Differences with Conditioning on Pre-Treatment Outcomes?"

Sylvain Chabé-Ferret

Toulouse School of Economics

Should We Combine Difference In Differences with Conditioning on Pre-Treatment Outcomes?*

Sylvain Chabé-Ferret†
Toulouse School of Economics, Inra and Iast

June 27, 2017

Abstract

Applied researchers often combine Difference In Differences (DID) with conditioning on pre-treatment outcomes when the Parallel Trend Assumption (PTA) fails. I examine both the theoretical and empirical basis for this approach. I show that the theoretical argument that both methods combine their strengths – DID differencing out the permanent confounders while conditioning on pre-treatment outcomes captures the transitory ones – is incorrect. Worse, conditioning on pre-treatment outcomes might increase the bias of DID. Simulations of a realistic model of earnings dynamics and selection in a Job Training Program (JTP) show that this bias can be sizable in practice. Revisiting empirical studies comparing DID with RCTs, I also find that conditioning on pre-treatment outcomes increases the bias of DID. Taken together, these results suggest that we should not combine DID with conditioning on pre-treatment outcomes but rather use DID conditioning on covariates that are fixed over time. When the PTA fails, DID applied symmetrically around the treatment date performs well in simulations and when compared with RCTs. Matching on several observations of pre-treatment outcomes also performs well in simulations, but evidence on its empirical performance is lacking.

Keywords: Difference in Differences - Matching - Selection Model - Treatment Effects.

JEL codes: C21, C23.

*I thank Alberto Abadie, Erich Battistin, Marco Caliendo, Habiba Djebbari, Koen Jochmans, Arthur Lewbel, Thierry Magnac, Nour Meddahi, Steve Pischke, Roland Rathelot, Jeff Smith, Nicolas Treich and Ken Troske for very constructive comments and suggestions on earlier versions of this work. I also thank participants at the French Econometrics Conference in Toulouse, the European Summer Symposium in Labor Economics at Ammersee, the IZA/IFAU Conference on Active Labor Market Programs Evaluation in Uppsala, the COMPIE conference in Rome and the Royal Economic Society Conference in Manchester and seminar participants at Aix-Marseille School of Economics for their useful feedback and suggestions. All remaining errors are my own.

†Correspondence to: Sylvain Chabé-Ferret, Toulouse School of Economics, 21 Allée de Brienne, 31015 Toulouse Cedex 6, France. Email: sylvain.chabe-ferret@tse-fr.eu. Tel: +33 (0)5 61 12 88 28. Fax:+33 (0)5 61 12 85 20.

1 Introduction

This paper studies the conditions under which the combination of Difference In Differences (DID) with conditioning on pre-treatment outcomes is a valid estimator of the effect of an intervention.[1] Combining DID with conditioning on pre-treatment outcomes is often used in empirical studies when evaluating Job Training Programs (JTPs),[2] but also other types of programs.[3]. This approach is especially used when the Parallel Trend Assumption (PTA) of DID fails, that is when the pre-treatment trends in outcomes between the treated and control groups are not parallel.

There are two informal arguments in favor of combining DID with conditioning on pre-treatment outcomes when the PTA fails, a theoretical one and an empirical one. The theoretical argument suggests that combining DID with conditioning on pre-treatment outcomes combines the strengths of both methods: DID differences out the permanent confounders while conditioning on pre-treatment outcomes captures the transitory ones.[4] A case in point is the evaluation of Job Training Programs (JTPs). Participants in JTPs have permanently lower earnings than non-participants but also experience a transitory decrease in earnings just before entering the program – a stylized fact known as Ashenfelter's dip.[5] Conditioning on pre-treatment outcomes captures these time varying confounders, the argument goes, while DID differences out the permanent fixed confounders, if there are any. The empirical argument is based on the fact that combining DID with conditioning on pre-treatment outcomes has been found to reproduce the results of Randomized Controlled Trials (RCTs) well, at least when evaluating the effect of JTPs on earnings (Heckman, Ichimura, Smith, and Todd, 1998; Smith and Todd, 2005; Mueser, Troske, and Gorislavsky, 2007).

[1] In this paper, I only consider the case where conditioning and differencing use pre-treatment outcomes observed at different dates. When conditioning and differencing use pre-treatment outcomes observed at the same date, the DID estimator converges to a simple matching estimator. The question then simplifies to whether we should use matching or DID, which is examined by Chabé-Ferret (2015).

[2] See *e.g.* Heckman, Ichimura, Smith, and Todd (1998); Smith and Todd (2005); Mueser, Troske, and Gorislavsky (2007)

[3] See *e.g.* Galiani, Gertler, and Schargrodsky (2005); Pufahl and Weiss (2009); Fowlie, Holland, and Mansur (2012); Chabé-Ferret and Subervie (2013)

[4] See Abadie (2005) for a statement of this informal theoretical argument.

[5] See Heckman, LaLonde, and Smith (1999) for a survey of the evidence on this phenomenon.

Despite the increasing use of DID combined with conditioning on pre-treatment outcomes in applied work, there exists no assessment of the soundness of the theoretical and empirical arguments on which it rests. In this paper, I make a careful assessment of this approach, both in theory and in practice. First, I build a simple model that exhibits selection on both permanent and transitory confounders. I derive necessary and sufficient conditions on the parameters of this model for the combination of DID with conditioning on pre-treatment outcomes to be consistent. Second, I run simulations of a model of earnings dynamics and self-selection into a JTP in order to assess the size of the bias in a realistic application. Third, I revisit experimental estimates of the performance of DID conditioning on pre-treatment outcomes in order to compare them with the predictions from my model.

In my simulations, I focus on the example of the effect of JTPs on earnings for several reasons. First, JTPs are crucial components of the modern welfare state, especially in a context in which innovations and trade disrupt entire sectors in developed countries and require the retooling of millions of workers. Second, earnings are the main outcome that a JTP seeks to influence, especially by increasing the human capital of workers. Third, earnings dynamics are described extensively by some well-known processes whose parameters have been estimated in labor economics. Fourth, both observational methods and RCTs have been and still are extensively used to evaluate JTPs. Fifth, empirical results comparing the bias of observational methods to an experimental benchmark are available for the effect of JTPs on earnings and their results can be contrasted with the predictions of the model.

The main result of this paper is that both the theoretical and the empirical arguments in favor of combining DID with conditioning on pre-treatment outcomes are incorrect. My main theoretical result shows that there is no configuration in which the combination of DID with conditioning on pre-treatment outcomes is consistent in a model that exhibits selection on both permanent and transitory confounders. The only configurations where the combination of DID with conditioning on pre-treatment outcomes is consistent is when there is either no selection on a fixed confounder – and matching conditioning on pre-treatment outcomes is consistent – or no selection on transitory ones – and DID without conditioning on pre-treatment outcomes is consistent. The intuition for this result is that conditioning on pre-treatment outcomes generates time varying selection bias while the validity of DID is predicated upon the assumption that selection bias is constant over time.

Worse, I find two cases where combining DID with conditioning on pre-treatment outcomes can generate bias for an otherwise consistent DID estimator. These are, to my knowledge, the first concrete examples of the fallacy of alignment, a term coined by Heckman and Navarro-Lozano (2004) to describe situations where conditioning on observed covariates might actually increase the bias of an estimator. The first instance of the fallacy of alignment in my results appears when selection is only due to permanent confounders and transitory shocks are persistent. In that case, DID without conditioning on pre-treatment outcomes is consistent but DID conditioning on pre-treatment outcomes is not. The second theoretical instance of the fallacy of alignment appears when selection is due to both permanent and transitory confounders. In that case, under conditions made precise in Chabé-Ferret (2015),[6] selection bias is symmetric around the treatment date and DID applied symmetrically around the treatment date without conditioning on pre-treatment outcomes is consistent. Under these conditions, conditioning on pre-treatment outcomes generates bias for the DID estimator. The intuition for this result is that conditioning on pre-treatment outcomes breaks the symmetry of selection bias around the treatment date.

Although interesting, these results are mainly theoretical. It is possible that the bias of DID combined with conditioning on pre-treatment outcomes is small in actual applications and that this approach, although theoretically inconsistent, is approximately valid. This might explain why it has been found to reproduce the results of RCTs very well in the case of JTPs. I use a model of self-selection and earnings dynamics calibrated with realistic parameter values taken from the literature to gauge how the the bias of DID varies as we condition on pre-treatment outcomes. There are two main results from the simulations.

[6] These conditions are that the agents have full information on transitory shocks when they select into the treatment, that transitory shocks are stationary and that the conditional expectation of transitory shocks conditional on the net utility of entering the program is linear.

First, when the conditions for selection bias to be symmetric are fulfilled, the bias of DID conditioning on one pre-treatment outcome is sizable: it is in absolute value of the order of magnitude of the treatment effects of JTPs usually found with RCTs. The bias of combining DID with conditioning on one observation of pre-treatment outcomes is thus large enough to mask the effects of most JTPs. Second, even when selection bias is not symmetric around the treatment date, the bias of DID conditioning on one pre-treatment outcome is still generally larger in absolute value than the bias of DID not conditioning on pre-treatment outcomes and applied symmetrically around the treatment date.

Both the theoretical and simulation results suggest that combining DID with conditioning on one observation of pre-treatment outcomes might not be the reason why DID matching performs so well at reproducing the results of RCTs. Both results suggest that it is rather DID applied symmetrically around the treatment date not conditioning on pre-treatment outcomes that might be the reason for the good performance of DID matching against RCTs. In order to check these predictions, I take a closer look at the original studies comparing DID matching with RCTs: Heckman, Ichimura, Smith, and Todd (1998) and Smith and Todd (2005).[7] I separate the estimates depending on whether they condition or not on pre-treatment outcomes and whether or not they are applied symmetrically around the treatment date. In both papers, I find that it is the application of DID symmetrically around the treatment date rather than the combination of DID with conditioning on pre-treatment outcomes that performs well, in agreement with both my theoretical and simulation results.

Taken together, these results suggest that we should not combine DID with conditioning on pre-treatment outcomes. Indeed, not only do the theoretical and empirical arguments in favor of this approach not hold, but there are theoretical and empirical arguments suggesting that conditioning on pre-treatment outcomes might increase the bias of DID.

What to do then when the PTA does not hold and what we thought was a silver bullet – combining DID with conditioning on pre-treatment outcomes – actually does not work?

[7] I do not include Mueser, Troske, and Gorislavsky (2007) in this analysis since this paper compares the observational estimates to experimental estimates on a different out-of-state population, thereby suggesting that differences between both approaches might stem from differences in populations.

Results in this paper suggest two possible approaches. First, DID applied symmetrically around the treatment date performs well in simulations and when compared with RCTs when estimating the effect of JTPs on earnings. Second, matching on several observations of pre-treatment outcomes performs well in simulations, but evidence on its empirical performance is lacking.

The approach of using a model of outcome dynamics and selection in a program to study the properties of observational estimators is rooted in an ancient literature. Ashenfelter (1978) formalizes the evaluation problem as a combination of selection on a fixed effect and on transitory shocks. Heckman (1978), Heckman and Robb (1985) and Ashenfelter and Card (1985) combine the selection equation with the outcome dynamics equation and introduce DID applied symmetrically around the treatment date. Bassi (1984) acknowledges that combining differencing with conditioning on pre-treatment changes in outcomes suffers from Nickell (1981)'s problem: pre-treatment changes in outcomes are correlated with transitory shocks. LaLonde (1986) tests whether observational estimators, among them DID combined with conditioning on pre-treatment outcomes, reproduce the results of an RCT. Heckman, Ichimura, and Todd (1997), Heckman, Ichimura, Smith, and Todd (1998) introduce the DID matching estimator. Abadie (2005) develops the informal theoretical argument in favor of combining DID with conditioning on pre-treatment outcomes and introduces a new DID matching estimator. Heckman and Navarro-Lozano (2004) show that the bias of matching might increase when conditional on additional covariates and coin the term "fallacy of alignment." Chabé-Ferret (2015) studies the bias of matching conditioning on pre-treatment outcome and of DID not conditioning on pre-treatment outcomes but does not study the combination of DID with conditioning on pre-treatment outcomes. He derives sufficient conditions for the estimators to be consistent in a model similar to the one in this paper, but does not derive necessary and sufficient conditions as in this paper.

This paper is structured as follows: Section 2 presents the theoretical results on when DID conditioning on one observation of pre-treatment outcomes is consistent; Section 3 presents the results of simulations of a model of earnings dynamics and selection in a JTP calibrated with realistic parameter values; Section 4 summarizes evidence from comparisons of observational methods to an experimental benchmark. Section 5 concludes.

2 Theoretical results

In this section, I formally derive theoretical results for the consistency of DID conditioning on one observation of pre-treatment outcomes in a simple model exhibiting selection both on a fixed effect and on transitory shocks. I also derive instances where conditioning on pre-treatment outcomes generates bias for an otherwise consistent DID estimator. I first present the model and then the main theoretical results.

Setting

The outcomes in the absence of the treatment depend on time and individual fixed effects and on transitory shocks (Equation (1a)). Transitory shocks are persistent: they follow an AR(1) process with $|\rho| < 1$ (Equation (1b)).

$$Y_{i,t}^0 = \delta_t + \mu_i + U_{it} \tag{1a}$$

$$\text{with } U_{i,t} = \rho U_{i,t-1} + v_{i,t} \tag{1b}$$

$$D_{i,k} = \mathbb{1}[t \geq k]\mathbb{1}[\underbrace{\theta_i + \gamma Y_{i,k-1}}_{D_{i,k}^*} \geq 0]. \tag{1c}$$

Treatment is offered at period k. Selection into the program depends on an individual fixed effect θ_i and on outcomes at date $k-1$ (Equation (1c)). The two critical parameters for selection are γ and $\rho_{\theta,\mu}$ (the correlation of the fixed effect μ_i with the unobserved shifter of participation θ_i). When $\gamma = 0$, selection is due to the fixed effect only. When $\rho_{\theta,\mu} = 0$, selection is on the observed pre-treatment outcome $Y_{i,k-1}$ only.

I make the following assumptions: $\sigma^2 > 0$, $\sigma_{U_0}^2 > 0$, $\sigma_\mu^2 > 0$, $\sigma_\theta^2 > 0$. $v_{i,t}$ are *i.i.d.* mean-zero shocks with finite variance σ^2 and $U_{i,0}$ is a mean-zero shock with variance $\sigma_{U_0}^2$. $v_{i,t} \perp\!\!\!\perp (\mu_i, \theta_i), \forall t$ and $U_{i,0} \perp\!\!\!\perp (\mu_i, \theta_i, v_{i,t}), \forall t$. I finally assume that the conditional expectations $\mathbb{E}[Y_{i,k+\tau}^0 | D_{i,k}^*]$ and $\mathbb{E}[Y_{i,k+\tau}^0 | D_{i,k}^*, Y_{k-1}]$ are linear. This assumption simplifies a lot the analysis of the biases. It holds for example when all the error terms are normal.

Although admittedly very simple, the model described by equation (1) has several virtues. First, it encapsulates in the simplest possible setting the problem that combining DID with conditioning on pre-treatment outcomes is trying to solve: selection on a fixed effect and on transitory shocks. Second, this model also accounts for various types of realistic selection processes: namely self-selection in a JTP and a cutoff eligibility rule. Assuming no idiosyncratic trend, no MA terms and limited information, setting $\gamma = -\rho$ and $\theta_i = \frac{\alpha_i}{r} - c_i$, the model of earnings dynamics and entry into a JTP presented in Section 3 simplifies to the model described by Equation (1). As argued in Chabé-Ferret (2015), a program allocated when a running variable falls below some eligibility threshold can also be described by Equation (1c). In that case, $\gamma = -1$ and θ_i accounts for measurement error in the variable determining eligibility.

I study the asymptotic bias of three estimators of the average effect of the treatment on the treated (ATT) on outcomes observed τ periods after the treatment date: matching (M), DID and DID matching conditioning on one observation of pre-treatment outcomes (DIDM).

$$B(M_{k,\tau,1}) = \mathbb{E}[\mathbb{E}[Y^0_{i,k+\tau}|D_{i,k} = 1, Y_{i,k-1}] - \mathbb{E}[Y^0_{i,k+\tau}|D_{i,k} = 0, Y_{i,k-1}]|D_{i,k} = 1] \quad \text{(2a)}$$

$$B(DID_{k,\tau,\tau'}) = \mathbb{E}[Y^0_{i,k+\tau} - Y^0_{i,k-\tau'}|D_{i,k} = 1] - \mathbb{E}[Y^0_{i,k+\tau} - Y^0_{i,k-\tau'}|D_{i,k} = 0] \quad \text{(2b)}$$

$$\begin{aligned} B(DIDM_{k,\tau,1,\tau'}) = \mathbb{E}\Big[&\mathbb{E}[Y^0_{i,k+\tau} - Y^0_{i,k-\tau'}|D_{i,k} = 1, Y_{i,k-1}] \\ &- \mathbb{E}[Y^0_{i,k+\tau} - Y^0_{i,k-\tau'}|D_{i,k} = 0, Y_{i,k-1}]|D_{i,k} = 1\Big]. \quad \text{(2c)} \end{aligned}$$

The matching estimator compares the expected outcomes of the treated τ periods after the treatment to those of the untreated conditional on $Y_{i,k-1}$. $Y_{i,k-1}$ is the last pre-treatment outcome observed before the treatment is taken and intuitively the one containing the most relevant information for selection. The bias of matching is the expected difference in potential outcomes in the absence of the treatment between the treated and the untreated groups, conditional on $Y_{i,k-1}$, integrated over the distribution of $Y_{i,k-1}$ conditional on $D_{i,k} = 1$ (Equation (2a)).

The DID estimator compares the change in outcomes over time in the treated group to the change in outcomes over time in the untreated group. The change over time is computed by comparing outcomes τ periods after the treatment to outcomes τ' periods before the treatment. The bias of DID is equal to the difference in the change over time in potential outcomes in the absence of the treatment between the treated and the untreated groups (Equation (2b)).

The DIDM estimator compares the change in outcomes over time in the treated group to the change in outcomes over time in the untreated group conditional on $Y_{i,k-1}$. The bias of DIDM is equal to the difference in the change over time in potential outcomes in the absence of the treatment between the treated and the untreated groups conditional on $Y_{i,k-1}$, integrated over the distribution of $Y_{i,k-1}$ conditional on $D_{i,k} = 1$ (Equation (2c)).

Basic results

In this section, I derive necessary and sufficient conditions for M, DID and DIDM to be consistent in the model defined by Equation (1). As I want to state general results on the model parameters for each of the estimators to be consistent, I have to define the sets of periods k, τ and τ' for which I want the biases of the various estimators to cancel. The usual practice is to use these estimators without restricting their validity to any particular subset of the possible treatment dates (k) or lag between treatment and observation of outcomes (τ). Thus, I will define consistency in this model as requiring that the estimators are valid for all $k > 0$ and for all $\tau \geq 0$. Similarly, for DID and DIDM, I define consistency as the fact that the bias of the estimator is zero regardless of the pre-treatment period $k - \tau'$ used to construct the estimator, with $\tau' > 1$.[8]

Theorem 1 is the main result of this section. It shows that the intuitive story that combining DID with conditioning on pre-treatment outcomes combines their strengths – DID differencing out the fixed effect and conditioning on pre-treatment outcomes capturing the transitory shocks – is wrong. Indeed, Theorem 1 shows that there is no configuration of the parameter space such that the combination of DID with conditioning on pre-treatment

[8] I do not include the case where $\tau' = 1$ since $B(DIDM_{k,\tau,1,1}) = B(M_{k,\tau,1})$ by construction.

outcomes is consistent, apart from when either $\rho_{\theta,\mu} = 0$ (and there is no selection on the fixed effect) or $\rho = 0$ (and there is no selection bias due to transitory shocks). Theorem 1 also means that combining DID with conditioning on one observation of pre-treatment outcomes does not add any identifying power to simple matching and simple DID: it is consistent only if one of them also is. Worse, Theorem 1 also implies that there are instances ($\gamma = 0$ and $\rho \neq 0$) when DID not conditioning on pre-treatment outcomes is consistent while DID conditioning on one observation of pre-treatment outcomes is not. This is an instance of the fallacy of alignment, where conditioning on observed covariates actually increases the bias of an otherwise consistent estimator.

Theorem 1 (Consistency of M, DID and DIDM) *The following three statements hold in the model defined in Equation (1):*

(i) $\forall k > 0,\ \forall \tau \geq 0,\ B(M_{k,\tau,1}) = 0 \Leftrightarrow \rho_{\theta,\mu} = 0$

(ii) $\forall k > 0,\ \forall \tau \geq 0,\ \forall \tau' > 1,\ B(DID_{k,\tau,\tau'}) = 0 \Leftrightarrow \gamma = 0 \text{ or } \rho = 0$

(iii) $\forall k > 0,\ \forall \tau \geq 0,\ \forall \tau' > 1,\ B(DIDM_{k,\tau,1,\tau'}) = 0 \Leftrightarrow \rho_{\theta,\mu} = 0 \text{ or } \rho = 0$

PROOF: see Section A in the Appendix. ■

Part (i) of Theorem 1 shows that matching is consistent if and only if there is no selection on the fixed effect. Part (ii) of Theorem 1 shows that DID not conditioning on pre-treatment outcomes is consistent if and only if there is either no selection on transitory shocks or transitory shocks are not persistent. Part (iii) of Theorem 1 shows that combining DID with conditioning on pre-treatment outcomes is consistent if and only if either there is no selection on the fixed effect or temporary shocks are not persistent.

The intuition for Theorem 1 is that conditioning on pre-treatment outcomes generates time varying selection bias while the validity of DID is predicated upon the assumption that selection bias is constant over time. The only settings in which conditioning on pre-treatment outcomes generates selection bias that is constant over time is either when selection bias after conditioning is zero, and thus simple conditioning is also consistent or when temporary shocks are not auto-correlated, in which case simple DID is also consistent.

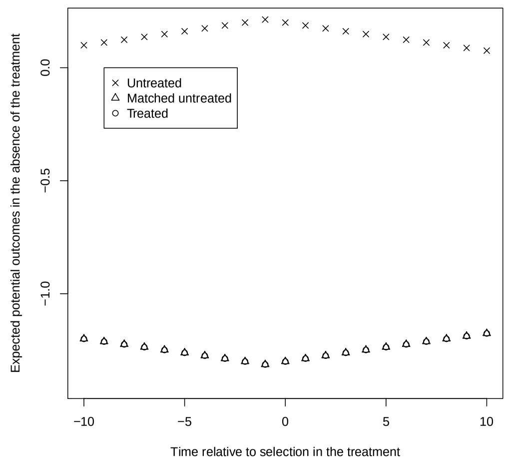
Figure 1 illustrates the results of Theorem 1. It shows the results of simulations of the model described in equation 1 using the formulae derived in Chabé-Ferret (2015)'s Appendix B. Figure 1 shows the expected value of the outcomes in the absence of the treatment around the treatment date for the treated (circles), the untreated (crosses) and the matched untreated (triangles), *i.e.* the untreated with the same distribution of $Y_{i,k-1}$ as the treated. The difference between treated and untreated measures selection bias. The difference between treated and matched untreated measures the bias of matching. The difference between the bias of matching before and after the treatment date measures the bias of DID matching.

In Figure 1(a), selection is on transitory shocks only ($\rho_{\theta,\mu} = 0$). As expected from part (i) of Theorem 1, simple matching is consistent since treated and matched untreated are aligned at every period after the treatment date: the matched untreated perfectly proxy for the counterfactual outcomes of the treated. DID conditioning on one observation of pre-treatment outcomes is consistent when matching is consistent since pre-treatment outcomes are also aligned before period $k-1$, despite the fact that they have not been explicitly conditioned on. This is because $Y_{i,k-1}$ is a sufficient statistics for selection in that case.

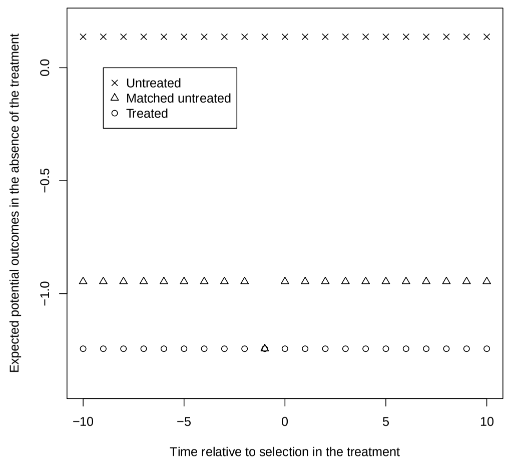
In Figure 1(b), selection is on the fixed effect only ($\gamma = 0$) and transitory shocks are not persistent ($\rho = 0$). As expected from part (ii) of Theorem 1, DID not conditioning on pre-treatment outcomes is consistent as selection bias is constant over time. DID conditioning on one observation of pre-treatment outcomes is also consistent, because the bias of matching is constant over time, except at period $k-1$ where it is zero by construction.

Figure 1(c) illustrates the main result of Theorem 1: contrary to the intuitive idea that they combine their strengths, combining DID with conditioning on pre-treatment outcomes does not get rid of selection bias when it is due to both a fixed effect and transitory shocks. Indeed, Figure 1(c) selection is both on the fixed effect and on transitory shocks. As expected from Theorem 1, simple matching is biased, which is apparent because treated and matched untreated are not perfectly aligned after the treatment date. DID not

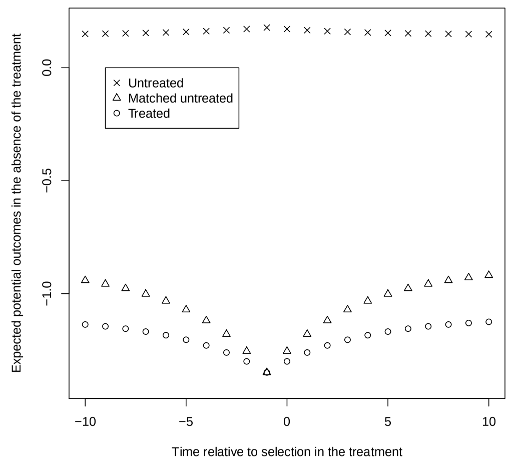
Figure 1 – Expected outcomes in the absence of the treatment in the model of Equation (1) under various parametric restrictions



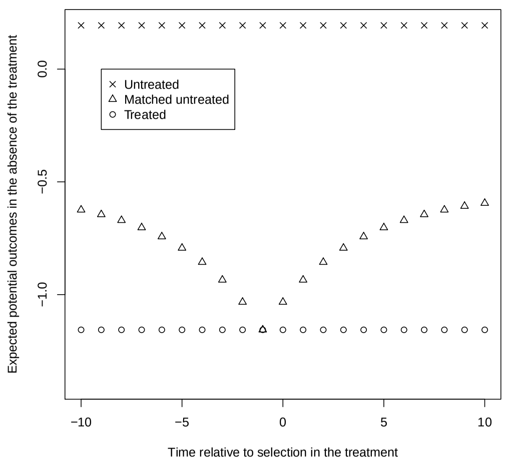
(a) $\rho_{\theta,\mu} = 0$



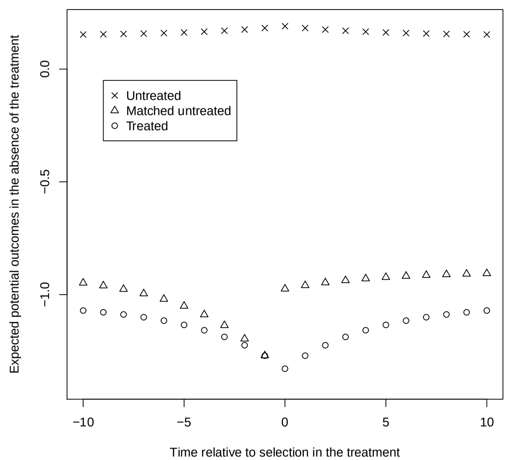
(b) $\gamma = 0$ and $\rho = 0$



(c) $\rho_{\theta,\mu} \neq 0$ and $\rho \neq 0$ and $\gamma \neq 0$



(d) $\gamma = 0$ and $\rho \neq 0$



(e) Full information

Note: this figure plots the average potential outcomes in the absence of the treatment for three groups: treated ($\circ$) ($\mathbb{E}[Y^0_{i,k+\tau}|D_{i,k} = 1]$), untreated ($\times$) ($\mathbb{E}[Y^0_{i,k+\tau}|D_{i,k} = 0]$) and matched untreated ($\triangle$), *i.e.* untreated that have the same potential outcomes at period $k-1$ as the treated ($\mathbb{E}[\mathbb{E}[Y^0_{i,k+\tau}|D_{i,k} = 0, Y_{i,k-1}]|D_{i,k} = 1]$). These are results of simulations of the model presented in Chabé-Ferret (2015)'s Appendix B.

conditioning on pre-treatment outcomes is also biased since selection bias varies over time. DID conditioning on one observation of pre-treatment outcomes is also biased because the difference between treated and matched untreated varies over time.

Figure 1(d) illustrates one instance of the fallacy of alignment. In Figure 1(d), selection is on the fixed effect only ($\gamma = 0$) and thus DID is consistent. Selection bias is constant over time in that case and the difference between treated and untreated at any pre-treatment date is a consistent proxy for post-treatment selection bias. Unlike in Figure 1(b), though, transitory shocks are persistent ($\rho \neq 0$), the bias of simple matching is not constant over time, and thus DID conditioning on one observation of pre-treatment outcomes is inconsistent.

Results under full information

In the selection model under full information studied by Chabé-Ferret (2015), DID applied symmetrically around the treatment date is consistent while DID conditioning on pre-treatment outcomes is not, as Theorem 2 shows. In this model, agents select into the program based on their outcomes at period k. It is the case for example if agents anticipate their future earnings shocks (bonus decrease, layoff, etc) and decide to enter a JTP as a consequence. In the selection model under full information, the selection Equation (1c) is replaced by:

$$D^f_{i,k} = \mathbb{1}[t \geq k]\mathbb{1}[\underbrace{\theta^f_i + \gamma^f Y^0_{i,k}}_{D^{*f}_{i,k}} \geq 0]. \tag{3}$$

Theorem 2 *In the model under full information (where Equation (3) substitutes for Equation (1c)), the following two statements hold when $\sigma^2_{U_0} = \frac{\sigma^2}{1-\rho^2}$:*

(i) $\forall k > 0$, $\forall \tau > 0$, $B(DID_{k,\tau,\tau}) = 0$,

(ii) $\forall k > 0$, $\forall \tau > 0$, $B(DIDM_{k,\tau,1,\tau}) \neq 0$ *except if* $a\sigma_\mu = -\gamma^f \rho(\sigma^2_\mu + \frac{\sigma^2}{1-\rho^2})$.

Proof: see Section A in the Appendix. ■

Theorem 2 shows that when the outcome process is stationary,[9] DID applied symmetrically around the treatment date not conditioning on pre-treatment outcomes is consistent while combining DID with conditioning on one observation of pre-treatment outcomes is not. The result in Theorem 2 is important since it shows that there exists an estimator that is consistent even when there is selection both on a fixed effect and on transitory shocks. Figure 1(e) illustrates this case: selection bias forms and dissipates at the same pace and is thus symmetric around the treatment date k. Conditioning on $Y_{i,k-1}$ breaks the symmetry of the dip and renders DID inconsistent. On Figure 1(e), the bias of matching decreases as it gets closer to $k-1$, increases sharply at date k because of the last shock before selection ($v_{i,k}$), and decreases thereafter. DID conditioning on $Y_{i,k}$ and applied symmetrically around the treatment date would be consistent, but it is infeasible since the potential outcomes of the participants are unobserved.

3 Simulation results

In this section, I use simulations a model of earnings dynamics and selection into a JTP calibrated with realistic parameter values in order to gauge the likely size of the bias of DID conditioning on pre-treatment outcomes and how it compares with DID and matching in real applications.

The model used in the simulations combines an equation for earnings dynamics and a selection equation (see Chabé-Ferret (2015) for a detailed discussion of this model). The process for the log-earnings of individual i at time t in the absence of the treatment ($Y^0_{i,t}$)

[9] Because $\sigma^2_{U_0}$ is equal to its long run value of $\frac{\sigma^2}{1-\rho^2}$.

has the following form:

$$Y^0_{i,t} = a + b\frac{18+t}{10} + c\left(\frac{18+t}{10}\right)^2 + (\delta + r_t d)E_i + \mu_i + \beta_i t + U_{it} \tag{4a}$$

$$\text{with } U_{i,t} = \rho U_{i,t-1} + m_1 v_{i,t-1} + m_2 v_{i,t-2} + v_{i,t} \tag{4b}$$

$$v_{i,t} \text{ i.i.d. mean-zero shocks with finite variance } \sigma^2, \tag{4c}$$

$$v_{i,t} \perp\!\!\!\perp (E_i, \beta_i, \mu_i), \forall t, \tag{4d}$$

$$(U_{i,0}, v_{i,0}, v_{i,-1}) \text{ mean-zero shocks with covariance matrix } \Sigma_0, \tag{4e}$$

$$(U_{i,0}, v_{i,0}, v_{i,-1}) \perp\!\!\!\perp (E_i, \beta_i, \mu_i, v_{i,t}), \forall t. \tag{4f}$$

The main parameters used to calibrate the model are presented in Table 1, while the full list of parameter values is presented in Table 2 in Appendix B. This model encapsulates

Table 1 – Parameters values used for the earnings process

	RIP (MaCurdy, 1982)	HIP (Guvenen, 2007, 2009)
ρ	0.99	0.821
m_1	-0.4	0
m_2	-0.1	0
σ^2	0.055	0.055
σ^2_μ	0	0.022
σ^2_β	0	0.00038
$\sigma_{\mu,\beta}$	0	-0.002

Note: σ^2_μ (resp. σ^2_β) is the variance of μ_i (resp. β_i). $\sigma_{\mu,\beta}$ is the covariance between μ_i and β_i. The values of the parameters of the yearly earnings process come from MaCurdy (1982) and Guvenen (2007, 2009). The only exception is the estimate of σ^2 in the HIP: for simplicity, it is set to the same value as the one estimated by MaCurdy (1982). Guvenen (2007, 2009) estimates the HIP model with a measurement error term on top of the AR(1) component. The sum of the variances of these two shocks is of the same order of magnitude as σ^2 estimated by MaCurdy (1982).

the two leading views on the process of earnings dynamics: the Restricted Income Profile (RIP) and the Heterogeneous Income Profile (HIP).

The equation for modeling the net utility of entering a JTP at period k has the following shape:

$$D^{*\iota}_{i,k} = \frac{\alpha_i}{r} - c_{i,k} - \mathbb{E}[Y^0_{i,k}|\mathcal{I}^\iota_{i,k}], \tag{5}$$

$$c_{i,k} = c_i + \beta_x E_i - \left(a + b\frac{18+k}{10} + c\left(\frac{18+k}{10}\right)^2 + (\delta + r_k d)E_i \right) \tag{6}$$

$D^{*\iota}_{i,k}$ has three components. The first part is the discounted gains from entering the program. The second part is the direct cost of participation to the program $c_{i,k}$. The third part is the opportunity cost of entering the program (expected foregone earnings). $c_{i,k}$ has itself two parts. The first part is composed of administrative costs which depend partly on education ($c_i + \beta_x E_i$). The second part is income support from the government that is equal to experience and education rated average earnings. $\mathcal{I}^\iota_{i,k}$ denotes agents' information set when computing the expected foregone earnings. I consider three distinct assumptions on $\mathcal{I}^\iota_{i,k}$:

Full information: $\mathcal{I}^f_{i,k} = \left\{X_i, \alpha_i, c_i, \mu_i, \beta_i, \{\delta_j\}^k_{j=1}, \{v_{i,j}\}^k_{j=1}\right\}$. Agents know all the shocks up to period k and can perfectly forecast their foregone earnings ($\mathbb{E}[Y^0_{i,k}|\mathcal{I}^f_{i,k}] = Y^0_{i,k}$).

Limited information: $\mathcal{I}^l_{i,k} = \left\{X_i, \alpha_i, c_i, \mu_i, \beta_i, \{\delta_j\}^k_{j=1}, \{v_{i,j}\}^{k-1}_{j=1}\right\}$. Agents do not know the last idiosyncratic shock to their earnings.[10] Limited information can arise because agents have to decide whether or not to enter the program at the end of period $k-1$ before observing the change to their earnings that occurs at period k. Their expected foregone earnings are $\mathbb{E}[Y^0_{i,k}|\mathcal{I}^l_{i,k}] = Y^0_{i,k} - v_{i,k}$.

Bayesian updating: $\mathcal{I}^b_{i,k} = \left\{E_i, \alpha_i, c_i, \mu^o_i, \beta^o_i, \{\delta_j\}^k_{j=1}, \{Y_{i,j}\}^{k-1}_{j=1}\right\}$. In this setup, the idiosyncratic intercept and slope terms are the sum of two components: $\mu_i = \mu^o_i + \mu^u_i$ and $\beta_i = \beta^o_i + \beta^u_i$. Agents observe $\{\mu^o_i, \beta^o_i\}$ at period 0 but have no information on $\{\mu^u_i, \beta^u_i, U_{i,0}\}$. They thus start with a prior on $\{\mu_i, \beta_i, U_{i,0}\}$ centered at $\{\mu^o_i, \beta^o_i, 0\}$. They then observe $Y_{i,1}$, E_i and δ_1 and use Kalman filtering to form a posterior on $\{\mu_i, \beta_i, U_{i,1}\}$. Expected foregone earnings at period k ($\mathbb{E}[Y^0_{i,k}|\mathcal{I}^b_{i,k}]$) are formed us-

[10]Note that agents know the shock to the overall economy δ_k. This is for comparability with the full information case.

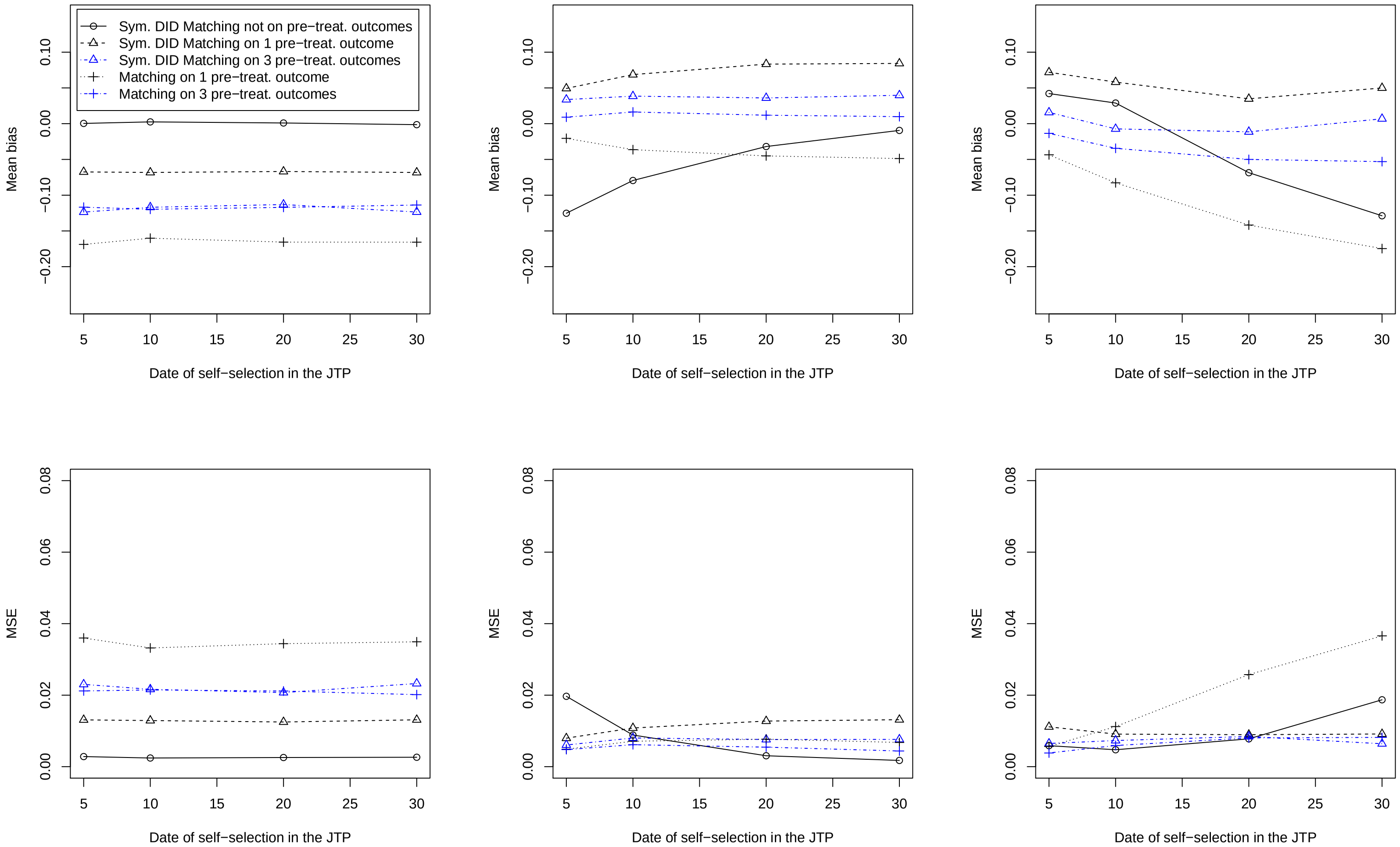
ing the posterior distribution of $\{\mu_i, \beta_i, U_{i,k}\}$ given $\mathcal{I}^b_{i,k}$ (see Chabé-Ferret (2015)'s Appendix C for a complete description).

Figure 2 presents the results of three simulations that are representative of the overall behavior of the estimators under various configurations of the model presented just above. Additional simulation results can be found in Appendix C.

The main results out of the simulations are the following. First, DID applied symmetrically around the treatment date not conditioning on pre-treatment outcomes is generally less biased than DID combined with conditioning on one observation of pre-treatment outcomes. DID applied symmetrically around the treatment date not conditioning on pre-treatment outcomes is also generally less biased than matching on one observation of pre-treatment outcomes. Second, DID applied symmetrically around the treatment date not conditioning on pre-treatment outcomes is generally more biased than by DID conditioning on three observations of pre-treatment outcomes. Third, DID and matching are close to each other when conditioning on three observations of pre-treatment outcomes.

Figure 2(a) shows that DID applied symmetrically around the treatment date not conditioning on pre-treatment outcomes is consistent when selection bias is symmetric around the treatment date, as expected from Theorem 2. Figure 2(a) also shows that the bias of combining DID with conditioning on one observation of pre-treatment outcomes is roughly equal to -0.06, which is in absolute value in the ballpark of the estimates of the causal effects of JTPs on (log-)earnings. The bias of combining DID with conditioning on one observation of pre-treatment outcomes is thus in this case large enough to mask the effects of most JTPs. Two additional results are noteworthy on Figure 2(a). The bias of matching on a single observation of pre-treatment outcomes is much larger than that of DID with the same conditioning set, so that differencing does bring something in that configuration. This is because the bias of matching before the treatment date is of the same sign as and of half the size of the bias of matching after the treatment date (see Figure 1(e)). Also, conditioning on additional observations of pre-treatment outcomes makes DID and matching more similar.

Figure 2 – Mean bias and MSE of matching and DID in simulations of a model of earnings dynamics and selection in a JTP



(a) Symmetric selection bias (b) Asymmetric selection bias – no trend (c) Asymmetric selection – trend

Note: the bias of both estimators is estimated $\tau = 4$ periods after the JTP. DID is always applied symmetrically around the treatment date. Matching and DID matching on pre-treatment outcomes is performed using $Y_{i,k-1}$ or $\left\{Y_{i,k-1}, Y_{i,k-2}, Y_{i,k-3}\right\}$ as conditioning variables, for $k \in \{5, 10, 20, 30\}$. The mean bias and the mean squared error (MSE) are calculated thanks to 500 Monte-Carlo replications. Each sample contains 1000 individuals with roughly 100 to 200 participants. The bias is estimated using LLR Matching on the propensity score with a biweight kernel. The bandwidth is set to .15 and the trimming level is set to .4. The "symmetric selection bias" results are obtained simulating the RIP model with full information and $\sigma^2_{U_0} = \sigma^2_{U_\infty} = \left(1 + (\rho + m_1)^2 + (\rho^2 + \rho m_1 + m_2)^2 \frac{1}{1-\rho^2}\right)\sigma^2$. The "asymmetric selection bias – no trend" results are obtained simulating the RIP model with limited information and $\sigma^2_{U_0} = \sigma^2$. The "asymmetric selection bias – trend" results are obtained simulating the HIP model with Bayesian updating and $\sigma^2_{U_0} = \sigma^2$. Details on the parametrization of the model can be found in Appendix B.

Figure 2(b) shows the results of simulations where selection bias is not symmetric around the treatment date. This is due to a combination of two phenomena: first, agents have limited information, in that they do not know the last shock to their earnings when deciding to enter the treatment; and second, the earnings process is not stationary, since it starts with a very small variance for the initial shock. An asymmetric selection bias generates bias for the DID estimator applied symmetrically around the treatment date not conditioning on pre-treatment outcomes. The size of the bias decreases in absolute value as individuals gain experience and the variance of the shocks increases. The DID estimator applied symmetrically around the treatment date not conditioning on pre-treatment outcomes does pretty well after the middle of the career, and actually better than the alternatives in terms of MSE. Figure 2(b) also shows that conditioning on additional observations of pre-treatment outcomes improves both DID and matching and makes them closer to each other and more stable along the life-cycle. Conditional on three observations of pre-treatment outcomes, DID and matching perform as well as DID applied symmetrically around the treatment date not conditioning on pre-treatment outcomes late in the life-cycle and better early in the life-cycle.

Figure 2(c) shows the results of a simulation of the HIP model with Bayesian learning and initial conditions for the $U_{i,t}$ process different from the long run ones, where selection bias is also asymmetric around the treatment date. We can see the same features as in the previous configuration: DID applied symmetrically around the treatment date not conditioning on pre-treatment outcomes dominates both matching and DID matching conditioning on one observation of pre-treatment outcomes, but is dominated by DID and matching conditioning on three observations of pre-treatment outcomes. Additionally, both DID matching and matching conditioning on three observations of pre-treatment outcomes are close to each other.

4 Revisiting experimental estimates

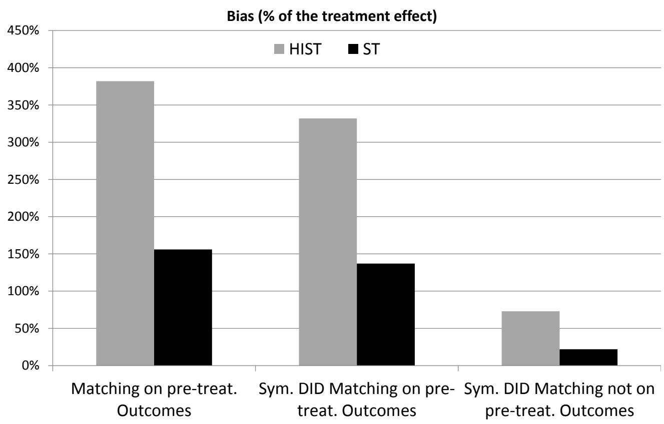
This section revisits the results of Heckman, Ichimura, Smith, and Todd (1998) and Smith and Todd (2005), that compare DID and matching to an experimental benchmark.[11] Following the approach initiated by LaLonde (1986), these studies compare experimental estimates of the effects of JTPs stemming from RCTs to observational estimates of the effects of the same program, using as much as possible the same data. These studies have found that DID matching is the method that reproduces best the results of RCTs. Revisiting the results of these studies in detail, I find support for the main prediction from Section 3: DID applied symmetrically around the treatment date not conditioning on pre-treatment outcomes performs better at reproducing the experimental results than DID conditioning on one observation of pre-treatment outcomes. Figure 3 summarizes the main results of these two studies.

Heckman, Ichimura, Smith, and Todd (1998) compare nonexperimental estimates of the effect of the Job Training Partnership Act (JTPA) obtained with matching and DID matching to the experimental benchmark, making use of the random allocation of the program. They implement DID symmetrically around the treatment date. They vary the set of control variables when assessing the performance of DID matching. Heckman, Ichimura, Smith, and Todd (1998)'s results suggest that conditioning on pre-treatment earnings increases the bias of DID. With a coarse set of predictors (only variables that are constant over time like age, schooling and marital status), the bias of DID applied symmetrically around the treatment date is equal to 73% of the experimental treatment effect. When conditioning on one observation of pre-treatment earnings (model PII), the bias of DID applied symmetrically around the treatment date increases and equals 332% of the treatment effect.

[11] Mueser, Troske, and Gorislavsky (2007) provide a similar analysis but their experimental benchmark stems from a different population and thus differences between the estimates might be due to differences in populations. Mueser, Troske, and Gorislavsky (2007) nevertheless provide two results that are in line with the results in this paper. First, combining DID and conditioning on pre-treatment outcomes is less biased than matching on pre-treatment outcomes, a result apparent in Figure 2(a). Second, DID applied symmetrically around the treatment date without conditioning on pre-treatment outcomes is in the ballpark of the experimental estimate.

Smith and Todd (2005) compare the ability of matching and DID matching to reproduce the results of the famous National Supported Work (NSW) experiment already analyzed by LaLonde (1986). They apply DID roughly symmetrically around the treatment date, since the outcomes are measured in 1975 and 1978, and the treatment is allocated between 1976 and 1977. They vary the set of control variables when assessing the performances of matching and DID matching. Smith and Todd (2005)'s results show again that conditioning on pre-treatment earnings increases the bias of DID. With a coarse set of controls not including pre-treatment outcomes, the bias of DID matching is of 22% of the experimental treatment effect, with the smaller (and most efficient) bandwidth. On the same sample, the bias of DID matching conditioning on pre-treatment outcomes is of -137%.

Figure 3 – Empirical estimates of the absolute value of the bias of matching and DID relative to RCTs for two JTPs



Note: the figure presents the bias of various observational estimators estimated relative to an experimental estimate obtained using randomly allocated JTPs. HIST stand for Heckman, Ichimura, Smith, and Todd (1998) and ST for Smith and Todd (2005). The results of the bias of Matching and DID Matching from HIST are from their Table XIII on p.1062. The coarse set of predictors does not condition on pre-treatment earnings while the set PII does. The results of the bias of Matching from ST are from their Table 5 p.336 and the bias of DID Matching is from their Table 6 p.340. The LaLonde set of predictors does not contain pre-treatment earnings while the DW set does. The sample is the full LaLonde sample. The Matching estimator used for the comparisons is the local linear Matching with a small bandwidth (1.0).

It is more difficult to find support for the prediction that conditioning on additional observations of pre-treatment outcomes reduces the bias of DID matching, and dominates DID not conditioning on pre-treatment outcomes. Indeed, to my knowledge, most available studies comparing RCTs to DID matching do not have several observations of pre-treatment outcomes or do not use them. Dehejia and Wahba (2002) find that matching on two observations of pre-treatment earnings has a negligible bias relative to an experimental estimate. Smith and Todd (2005) criticize this result on the grounds that the second observation of pre-treatment earnings is only available for a very specific subgroup and that selection bias on this specific subgroup is actually zero without conditioning on any covariate. Andersson, Holzer, Lane, Rosenblum, and Smith (2013) do not directly test the prediction but find that adding more observations than two years of pre-treatment earnings does not change the matching estimator much.

5 Conclusion

Taken together, the results presented in this paper cast doubt on the theoretical and empirical arguments in favor of combining DID with conditioning on pre-treatment outcomes when estimating the effect of an intervention. Indeed, the intuitive story that DID and conditioning on pre-treatment outcomes combine their strengths – DID differencing out the fixed effect and conditioning on pre-treatment outcomes capturing transitory shocks – is not valid theoretically. Worse, there are cases in which conditioning on pre-treatment outcomes actually generates bias for an otherwise consistent DID estimator, an instance of Heckman and Navarro-Lozano (2004)'s fallacy of alignment. It is especially the case when DID is applied symmetrically around the treatment date: under certain conditions, this estimator is consistent not conditioning on pre-treatment outcomes even when there is selection both on a fixed effect and on transitory shocks, while it is biased when conditioning on pre-treatment outcomes.

When estimating the effect of a JTP on earnings, the results presented in this paper suggest to use DID applied symmetrically around the treatment date not conditioning on pre-treatment outcomes when a few observations of pre-treatment outcomes are available and to use matching conditioning on all available pre-treatment earnings when there are at least three observations of pre-treatment outcomes. In simulations of a model of earnings and self-selection in a JTP, the bias of DID conditioning on one observation of pre-treatment outcomes is indeed sizable (of the order of magnitude of treatment effects usually estimated in the literature) and DID applied symmetrically around the treatment date performs better even when it is not consistent. Results of studies comparing observational estimators to an experimental benchmark confirm that the bias of DID conditioning on one observation of pre-treatment outcomes is actually higher than the bias of DID applied symmetrically around the treatment date not conditioning on pre-treatment outcomes. When conditioning on three observations of pre-treatment outcomes, simulations show that matching and DID matching are actually very close and dominate DID applied symmetrically around the treatment date not conditioning on pre-treatment outcomes. Unfortunately, to my knowledge, there is no empirical result testing that prediction.

The results in this paper also cast doubt on the practice of combining DID with conditioning on pre-treatment outcomes in applications other than the evaluation of the effect of JTPs on earnings. Indeed, this paper shows that there is no sound theoretical basis for combining DID with conditioning on pre-treatment outcomes. When few observations on pre-treatment outcomes are available, I would advise for using DID without conditioning on pre-treatment outcomes and, if possible, applied symmetrically around the treatment date. When several observations of pre-treatment outcomes are available, I would advise for conditioning on all of them non-parametrically, using matching. In order to strengthen these suggestions, we are in dire need of simulations and of experimental results gauging the bias of observational methods for interventions other than JTPs.

References

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

Andersson, F., H. J. Holzer, J. I. Lane, D. Rosenblum, and J. Smith (2013): "Does Federally-Funded Job Training Work? Nonexperimental Estimates of WIA Training Impacts Using Longitudinal Data on Workers and Firms," Working Paper 19446, National Bureau of Economic Research.

Ashenfelter, O. (1978): "Estimating the Effect of Training Programs on Earnings," *The Review of Economics and Statistics*, 60(1), 47–57.

Ashenfelter, O., and D. Card (1985): "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *The Review of Economic Statistics*, 67(4), 648–660.

Bassi, L. J. (1984): "Estimating the Effect of Training Programs with Non-Random Selection," *The Review of Economics and Statistics*, 66(1), 36–43.

Chabé-Ferret, S. (2015): "Analysis of the Bias of Matching and Difference-In-Difference under Alternative Earnings and Selection Processes," *Journal of Econometrics*, 185(1), 110–123.

Chabé-Ferret, S., and J. Subervie (2013): "How Much Green for the Buck? Estimating Additional and Windfall Effects of French Agro-Environmental Schemes by Difference-In-Difference Matching," *Journal of Environmental Economics and Management*, 65(1), 12–27.

Dehejia, R. H., and S. Wahba (2002): "Propensity Score-Matching Methods For Non-experimental Causal Studies," *The Review of Economics and Statistics*, 84(1), 151–161.

Fowlie, M., S. P. Holland, and E. T. Mansur (2012): "What Do Emissions Markets Deliver and to Whom? Evidence from Southern California's NOx Trading Program.," *American Economic Review*, 102(2), 965 – 993.

Galiani, S., P. Gertler, and E. Schargrodsky (2005): "Water for Life: The Impact of the Privatization of Water Services on Child Mortality," *Journal of Political Economy*, 113(1), 83–120.

Guvenen, F. (2007): "Learning Your Earning: Are Labor Income Shocks Really Very Persistent?," *American Economic Review*, 97(3), 687 – 712.

——— (2009): "An Empirical Investigation of Labor Income Processes," *Review of Economic Dynamics*, 12(1), 58–79.

Heckman, J. J. (1978): "Longitudinal Studies in Labor Economics: A Methodological Review," University of Chicago.

Heckman, J. J., H. Ichimura, J. A. Smith, and P. E. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1099.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, 64(4, Special Issue: Evaluation of Training and Other Social Programmes), 605–654.

Heckman, J. J., R. J. LaLonde, and J. A. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics*, ed. by O. C. Ashenfelter, and D. Card, vol. 3, chap. 31, pp. 1865–2097. Elsevier, North Holland.

Heckman, J. J., and S. Navarro-Lozano (2004): "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models," *The Review of Economics and Statistics*, 86(1), 30–57.

Heckman, J. J., and R. Robb (1985): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman, and B. Singer, pp. 156–245. Cambridge University Press, New-York.

LaLonde, R. J. (1986): "Evaluating the Econometric Evaluation of Training Programs with Experimental Data," *American Economic Review*, 76, 604–620.

MaCurdy, T. E. (1982): "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis," *Journal of Econometrics*, 18(1), 83–114.

Mueser, P. R., K. R. Troske, and A. Gorislavsky (2007): "Using State Administrative Data to Measure Program Performance," *The Review of Economics and Statistics*, 89(4), pp. 761–783.

Nickell, S. (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49(6), 1417–1426.

Pufahl, A., and C. R. Weiss (2009): "Evaluating the Effects of Farm Programmes: Results from Propensity Score Matching," *European Review of Agricultural Economics*, 36(1), 79–101.

Smith, J. A., and P. E. Todd (2005): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics*, 125(1-2), 305–353.

A Proofs

I use $\sigma_{A,B}$ to denote the covariance between random variables A_i and B_i and σ_A^2 for the variance of A_i.

Lemma 1 *$\forall k > 0$, $\forall \tau \geq 0$, $B(M_{k,\tau,1}) = 0 \Leftrightarrow num_{k,\tau} = 0$, with $num_{k,\tau} = \sigma_{Y_{k+\tau},D_k^*}\sigma^2_{Y_{k-1}} - \sigma_{Y_{k-1},D_k^*}\sigma_{Y_{k-1},Y_{k+\tau}}$.*

Proof: By linearity of conditional expectations:

$$\mathbb{E}[Y^0_{i,k+\tau}|D^*_{i,k}, Y_{k-1}] = \mathbb{E}[Y^0_{i,t}] + \theta_{Y^0_{k+\tau},D^*_k}\left(D^*_{i,k} - \mathbb{E}[D^*_{i,k}]\right) + \theta_{Y^0_{k+\tau},Y^0_{k-1}}\left(Y^0_{i,k-1} - \mathbb{E}[Y^0_{i,k-1}]\right),$$

with $\theta_{Y^0_{k+\tau},D^*_k} = \frac{\mathrm{num}_{k,\tau}}{\sigma^2_{D^*_k}\sigma^2_{Y_{k-1}} - \sigma^2_{Y_{k-1},D^*_k}}$. As a consequence,

$$B(M_{k,\tau,1}) = \theta_{Y^0_{k+\tau},D^*_k}\mathbb{E}[\mathbb{E}[D^*_{i,k}|D_{i,k} = 1, Y_{k-1}] - \mathbb{E}[D^*_{i,k}|D_{i,k} = 0, Y_{k-1}]|D_{i,k} = 1].$$

The result follows because $\mathbb{E}[D^*_{i,k}|D_{i,k} = 1, Y_{k-1}] - \mathbb{E}[D^*_{i,k}|D_{i,k} = 0, Y_{k-1}] > 0$ and $\sigma^2_{D^*_k}\sigma^2_{Y_{k-1}} - \sigma^2_{Y_{k-1},D^*_k} > 0$. ■

Lemma 2 *$\forall k > 0$, $\forall \tau \geq 0$, $num_{k,\tau} = \sigma^2_{U_{k-1}}\sigma_\mu a(1-\rho^{\tau+1})$, with $a = \rho_{\theta,\mu}\sigma_\theta$ and $b = a + \gamma\sigma_\mu$.*

Proof:

$$\begin{aligned}
\mathrm{num}_{k,\tau} &= \left[b\sigma_\mu + \gamma\rho^{\tau+1}\sigma^2_{U_{k-1}}\right]\left[\sigma^2_\mu + \sigma^2_{U_{k-1}}\right] - \left[b\sigma_\mu + \gamma\sigma^2_{U_{k-1}}\right]\left[\sigma^2_\mu + \rho^{\tau+1}\sigma^2_{U_{k-1}}\right] \\
&= \gamma\rho^{\tau+1}\sigma^4_{U_{k-1}} - \gamma\rho^{\tau+1}\sigma^4_{U_{k-1}} \\
&\quad + \sigma^2_{U_{k-1}}\left[\gamma\rho^{\tau+1}\sigma^2_\mu + b\sigma_\mu - b\sigma_\mu\rho^{\tau+1} - \gamma\sigma^2_\mu\right] \\
&\quad + \sigma^2_\mu\left[b\sigma_\mu - b\sigma_\mu\right] \\
&= \sigma^2_{U_{k-1}}\sigma_\mu(b - \gamma\sigma_\mu)(1 - \rho^{\tau+1}) \\
&= \sigma^2_{U_{k-1}}\sigma_\mu a(1 - \rho^{\tau+1}).
\end{aligned}$$

■

Lemma 3 *$\forall k > 0$, $\forall \tau \geq 0$, $B(M_{k,\tau,1}) = 0 \Leftrightarrow \rho_{\theta,\mu} = 0$.*

Proof: Using Lemma 1 and 2, we have $\rho_{\theta,\mu} = 0 \Rightarrow \mathrm{num}_{k,\tau} = 0$. The reciprocal follows from the fact that $\sigma_\mu > 0$, $\sigma_\theta > 0$, $\sigma^2_{U_{k-1}} > 0$, $\forall k > 0$ and $(1 - \rho^{\tau+1}) > 0$, $\forall \tau \geq 0$. Thus $\mathrm{num}_{k,\tau} = 0 \Rightarrow \rho_{\theta,\mu} = 0$. ■

Lemma 4 *$\forall k > 0$, $\forall \tau \geq 0$, $\forall \tau' > 1$, $B(DIDM_{k,\tau,1,\tau'}) = 0 \Leftrightarrow num_{k,\tau} - num_{k,-\tau'} = 0$.*

Proof: This stems from the proof of Lemma 1. ■

Lemma 5 *$num_{k,-\tau'} = \sigma_\mu a(\sigma^2_{U_{k-1}} - \rho^{\tau'-1}\sigma^2_{U_{k-\tau'}})$.*

Proof:

$$\begin{aligned}
\text{num}_{k,-\tau'} &= \left[b\sigma_\mu + \gamma\rho^{\tau'-1}\sigma^2_{U_{k-\tau'}}\right]\left[\sigma^2_\mu + \sigma^2_{U_{k-1}}\right] - \left[b\sigma_\mu + \gamma\sigma^2_{U_{k-1}}\right]\left[\sigma^2_\mu + \rho^{\tau'-1}\sigma^2_{U_{k-\tau'}}\right] \\
&= \gamma\rho^{\tau'-1}\sigma^4_{U_{k-\tau'}} - \gamma\rho^{\tau'-1}\sigma^4_{U_{k-\tau'}} \\
&\quad + \sigma^2_{U_{k-1}}\sigma_\mu\left[b-\gamma\sigma_\mu\right] - \rho^{\tau'-1}\sigma^2_{U_{k-\tau'}}\sigma_\mu\left[b-\gamma\sigma_\mu\right] \\
&\quad + \sigma^2_\mu\left[b\sigma_\mu - b\sigma_\mu\right] \\
&= \sigma_\mu(b-\gamma\sigma_\mu)(\sigma^2_{U_{k-1}} - \rho^{\tau'-1}\sigma^2_{U_{k-\tau'}}) \\
&= \sigma_\mu a(\sigma^2_{U_{k-1}} - \rho^{\tau'-1}\sigma^2_{U_{k-\tau'}}).
\end{aligned}$$

■

Lemma 6 $\sigma^2_{U_t} = \frac{1-\rho^{2t}}{1-\rho^2}\sigma^2 + \rho^{2t}\sigma^2_{U_0}$.

Proof: The result follows from Equation (38) in Chabé-Ferret (2015) with $m_1 = m_2 = 0$. ■

Lemma 7 $\forall k > 0, \forall\tau \geq 0, \forall\tau' > 1, B(DIDM_{k,\tau,1,\tau'}) = 0 \Leftrightarrow \rho_{\theta,\mu} = 0.$

Proof: Using Lemma 4, 5 and 6, we have that:

$$\begin{aligned}
\text{num}_{k,\tau} - \text{num}_{k,-\tau'} &= \sigma_\mu a(\rho^{\tau'-1}\sigma^2_{U_{k-\tau'}} - \rho^{\tau+1}\sigma^2_{U_{k-1}}) \\
&= \sigma_\mu a\left(\underbrace{(\rho^{\tau'-1} - \rho^{\tau+1})\frac{\sigma^2}{1-\rho^2}}_{B(\tau,\tau')} + \rho^{2(k-\tau')}\underbrace{\left(\sigma^2_{U_0} - \frac{\sigma^2}{1-\rho^2}\right)(\rho^{\tau'-1} - \rho^{\tau+1}\rho^{2(\tau'-1)})}_{C(\tau,\tau')}\right)
\end{aligned}$$

From this, we have that $\rho_{\theta,\mu} = 0$ or $\rho = 0 \Rightarrow B(DIDM_{k,\tau,1,\tau'}) = 0$.

Since $\sigma^2_\mu > 0$ and $\sigma^2_\theta > 0$, $\forall k > 0, \forall\tau \geq 0, \forall\tau' > 1, B(DIDM_{k,\tau,1,\tau'}) = 0 \Rightarrow$ either $\rho_{\theta,\mu} = 0$ or $\forall k > 0, \forall\tau \geq 0, \forall\tau' > 1, A(k,\tau,\tau') = B(\tau,\tau') + \rho^{2(k-\tau')}C(\tau,\tau') = 0$. To prove the final result, it remains to be shown that condition $A(k,\tau,\tau') = 0$ only implies that $\rho = 0$. Let's assume that $\rho \neq 0$ and show that this yields to a contradiction. Fix τ and τ' such that $\tau + 2 \neq \tau'$. $A(k,\tau,\tau')$ as a function of k has at most one real root as long as $B(\tau,\tau') \neq 0$ or $C(\tau,\tau') \neq 0$. So $A(k,\tau,\tau') = 0, \forall k > 0 \Rightarrow B(\tau,\tau') = 0$ and $C(\tau,\tau') = 0$. But $B(\tau,\tau') = 0 \Rightarrow \rho = 0$ or $\tau + 2 = \tau'$ or $\sigma^2 = 0$, a contradiction (since $\sigma^2 > 0$ by assumption). This proves the result. ■

Lemma 8 $\forall k > 0, \forall\tau \geq 0, \forall\tau' > 1, B(DID_{k,\tau,\tau'}) = 0 \Leftrightarrow Cov(Y^0_{i,k+\tau}, D^*_{i,k}) - Cov(Y^0_{i,k-\tau'}, D^*_{i,k}) = 0.$

PROOF:

$$\mathbb{E}[Y^0_{i,t}|D^*_{i,k}] = \mathbb{E}[Y^0_{i,t}] + \frac{\text{Cov}(Y^0_{i,k+\tau}, D^*_{i,k})}{\text{Var}(D^*_{i,k})}\left(D^*_{i,k} - \mathbb{E}[D^*_{i,k}]\right)$$

$$\mathbb{E}[Y^0_{i,t}|D_{i,k} = 1] = \mathbb{E}[Y^0_{i,t}] + \frac{\text{Cov}(Y^0_{i,k+\tau}, D^*_{i,k})}{\text{Var}(D^*_{i,k})}\left(\mathbb{E}[D^*_{i,k}|D^*_{i,k} \geq 0] - \mathbb{E}[D^*_{i,k}]\right)$$

$$\mathbb{E}[Y^0_{i,t}|D_{i,k} = 1] - \mathbb{E}[Y^0_{i,t}|D_{i,k} = 0] = \frac{\text{Cov}(Y^0_{i,k+\tau}, D^*_{i,k})}{\text{Var}(D^*_{i,k})}\left(\mathbb{E}[D^*_{i,k}|D^*_{i,k} \geq 0] - \mathbb{E}[D^*_{i,k}|D^*_{i,k} < 0]\right)$$

The result follows because $\mathbb{E}[D^*_{i,k}|D^*_{i,k} \geq 0] - \mathbb{E}[D^*_{i,k}|D^*_{i,k} < 0] > 0$. ■

Lemma 9 *$Cov(Y^0_{i,k+\tau}, D^*_{i,k}) - Cov(Y^0_{i,k-\tau'}, D^*_{i,k}) = -\gamma A(k, \tau, \tau')$.*

PROOF:

$$\begin{aligned}
\text{Cov}(Y^0_{i,t}, D^*_{i,k}) &= \text{Cov}(\mu_i + U_{i,t}, \theta_i + \gamma\mu_i + \gamma U_{i,k-1}) \\
&= b\sigma_\mu + \gamma\text{Cov}(U_{i,t}, U_{i,k-1}) \\
&= b\sigma_\mu + \gamma\rho^{|t-k+1|}\sigma^2_{U_{\min\{t,k-1\}}} \\
&= \gamma(\rho^{\tau+1}\sigma^2_{U_{k-1}} - \rho^{\tau'-1}\sigma^2_{U_{k-\tau'}}).
\end{aligned}$$

Using the definition of $A(k, \tau, \tau')$ completes the proof. ■

Lemma 10 *$\forall k > 0$, $\forall \tau \geq 0$, $\forall \tau' > 1$, $B(DID_{k,\tau,\tau'}) = 0 \Leftrightarrow \gamma = 0$ or $\rho = 0$.*

PROOF: From Lemma 8 and 9, we have $\gamma = 0$ or $\rho = 0 \Rightarrow B(DID_{k,\tau,\tau'}) = 0$. Moreover, $\forall k > 0$, $\forall \tau \geq 0$, $\forall \tau' > 0$, $B(DID_{k,\tau,\tau'}) = 0 \Rightarrow \gamma = 0$ or $\forall k > 0$, $\forall \tau \geq 0$, $\forall \tau' > 0$, $A(k, \tau, \tau') = 0$. The same reasoning as in the proof of Lemma 7 shows that the second condition on $A(k, \tau, \tau')$ implies that $\rho = 0$. This proves the result. ■

Proof of Theorem 1

PROOF: Lemma 3, 7 and 10 prove the result. ■

Proof of Theorem 2

PROOF: Using the same line of reasoning as the proof of Theorem 1, but modifying it accordingly, yields the following result: $\forall k > 0$, $\forall \tau \geq 0$, $\forall \tau' > 0$, $B(DID_{k,\tau,\tau'}) = 0 \Rightarrow -\gamma^f A^f(k, \tau, \tau')$, with:

$$A^f(k, \tau, \tau') = \left((\rho^{\tau'} - \rho^{\tau})\frac{\sigma^2}{1-\rho^2} + \rho^{2(k-\tau')}\left(\sigma^2_{U_0} - \frac{\sigma^2}{1-\rho^2}\right)(\rho^{\tau'} - \rho^{\tau}\rho^{2\tau'})\right).$$

This proves the consistency of Symmetric DID when $\sigma^2_{U_0} = \frac{\sigma^2}{1-\rho^2}$.

In order to derive the bias of DIDM, it is useful to rewrite $D^{*f}_{i,k}$ as a function of $D^{*}_{i,k}$:

$$
\begin{aligned}
D^{*f}_{i,k} &= \theta^f_i + \gamma^f Y^0_{i,k} \\
&= \underbrace{\theta^f_i + \gamma^f \mu_i (1-\rho)}_{\theta_i} + \underbrace{\gamma^f \rho}_{\gamma} Y^0_{i,k-1} + \gamma^f v_{i,k} \\
&= D^{*}_{i,k} + \gamma^f v_{i,k}.
\end{aligned}
$$

Following the line of the proof of Theorem 1, we have that $B(DIDM_{k,\tau,1,\tau'}) = 0 \Leftrightarrow a\sigma_\mu A(k,\tau,\tau') + \gamma^f \rho^\tau \sigma^2(\sigma^2_\mu + \sigma^2_{U_{k-1}})$. When $\sigma^2_{U_0} = \frac{\sigma^2}{1-\rho^2}$, we have $B(DIDM_{k,\tau,1,\tau}) = 0 \Leftrightarrow \rho^{\tau-1}\sigma^2\left(a\sigma_\mu + \gamma^f \rho(\sigma^2_\mu + \frac{\sigma^2}{1-\rho^2})\right)$. This proves the result. ■

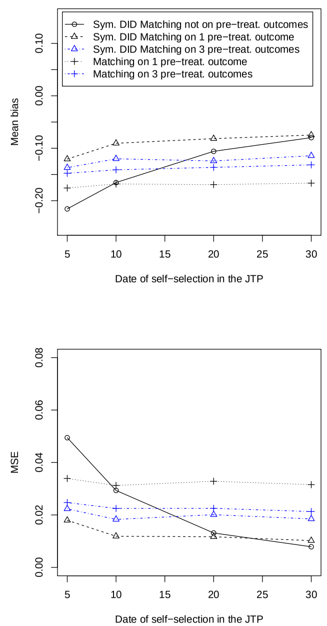
B Parameter values used in the simulations

Table 2 – Parameters used for the Monte-Carlo simulations

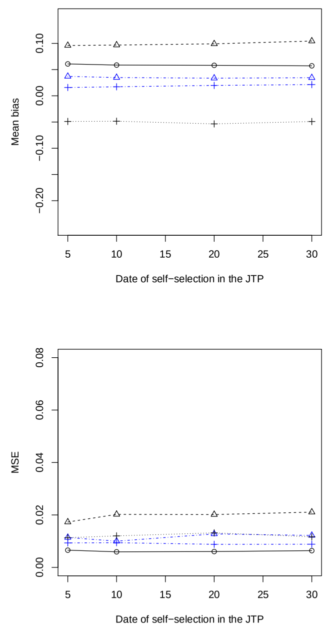
	RIP, long run	RIP, short run	HIP, long run	HIP, short run
Trimming level	0.4	0.4	0.4	0.4
Sample size	1000	1000	1000	1000
Number of periods	40	40	40	40
δ	0.08	0.08	0.08	0.08
d	0.02	0.02	0.02	0.02
a	8.83	8.83	8.83	8.83
b	0.56	0.56	0.56	0.56
c	-0.057	-0.057	-0.057	-0.057
β_x	-0.001	-0.001	-0.001	-0.001
ρ	0.99	0.99	0.821	0.821
m_1	-0.4	-0.4	0	0
m_2	-0.1	-0.1	0	0
$\bar{\alpha}$	0.1	0.1	0.1	0.1
$\bar{c}$	3	3	3	3
r	0.1	0.1	0.1	0.1
$\bar{\mu}$	0	0	0	0
$\bar{\beta}$	0	0	0	0
$\bar{x}$	2.3	2.3	2.3	2.3
σ_x^2	0.2	0.2	0.2	0.2
σ_μ^2	0	0	0.022	0.022
σ_β^2	0	0	0.00038	0.00038
σ^2	0.055	0.055	0.055	0.055
σ_c^2	0.05	0.05	0.05	0.05
σ_α^2	0	0	0	0
$\sigma_{\mu,\beta}$	0	0	-0.002	-0.002
$\rho_{\mu,c}$	0	0	0	0
$\rho_{\mu,x}$	0	0	0	0
$\rho_{\mu,\alpha}$	0	0	0	0
$\rho_{\beta,c}$	0	0	0	0
$\rho_{\beta,x}$	0	0	0	0
$\rho_{\beta,\alpha}$	0	0	0	0
$\rho_{c,x}$	0	0	0	0
λ	0	0	0.6	0.6
$\sigma_{U_0}^2$	$\sigma_{U_\infty}^2$	σ^2	$\sigma_{U_\infty}^2$	σ^2

C Additional simulation results

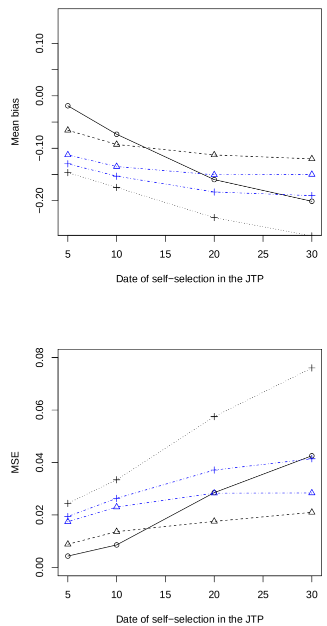
Figure 4 – Mean bias and MSE of matching and DID in additional simulations of a model of earnings dynamics and selection in a JTP



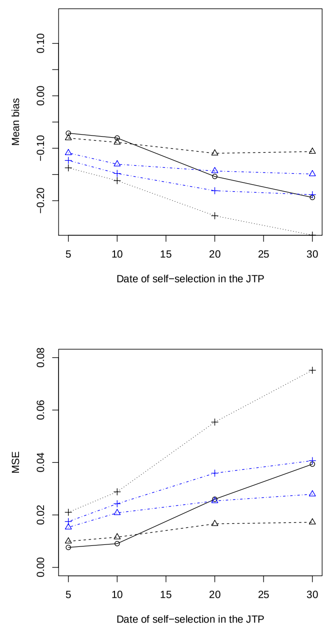
(a) RIP–full info–short run



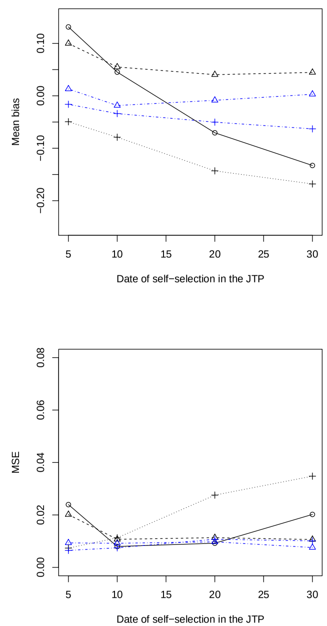
(b) RIP–limited info–long run



(c) HIP–full info–long run



(d) HIP–full info–short run



(e) HIP–Bayesian update–long run

Note: the bias of both estimators is estimated $\tau = 4$ periods after the JTP. Matching and DID Matching on pre-treatment outcomes is performed using $Y_{i,k-1}$ as a conditioning variable, for $k \in \{5, 10, 20, 30\}$. The mean bias and the mean squared error (MSE) are calculated thanks to 500 Monte-Carlo replications. Each sample contains 1000 individuals with roughly 100 to 200 participants. The bias is estimated using LLR Matching on the propensity score with a biweight kernel. The bandwidth is set to .15 and the trimming level is set to .4. Details on the parametrization of the model can be found in Appendix B.