Matching *vs* Differencing when Estimating Treatment Effects with Panel Data: the Example of the Effect of Job Training Programs on Earnings*

Sylvain Chabé-Ferret†
Irstea, Umr 1273 Métafort, Aubière, France.

First version : 15 July 2010, this version: October 8, 2012.

Abstract

This paper compares matching and Difference-In-Difference matching (DID) when estimating the effect of a program on a dynamic outcome. I detail the sources of bias of each estimator in a model of entry into a Job Training Program (JTP) and earnings dynamics that I use as a working example. I show that there are plausible settings in which DID is consistent while matching on past outcomes is not. Unfortunately, the consistency of both estimators relies on conditions that are at odds with properties of earnings dynamics. Using calibration and Monte-Carlo simulations, I show that deviations from the most favorable conditions severely bias both estimators. The behavior of matching is nevertheless less erratic: its bias generally decreases when controlling for more past outcomes and it generally provides a lower bound on the true treatment effect. I finally point to previously unnoticed empirical results that confirm that DID does well, and generally better than matching on past outcomes, at replicating the results of an experimental benchmark.

Keywords: Matching - Difference in Difference - Evaluation of Job training Programs.

JEL codes: C21, C23.

*A previous version of this paper was circulated under the title *"To Control or Not to Control? Bias of Simple Matching vs Difference-In-Difference Matching in a Dynamic Framework"*. This paper is a heavily revised version of this previous work and replaces it. The first version of this paper has been written as I was visiting Cowles Foundation for the Research in Economics at Yale University. I thank Don Andrews and Philip Haile for their invitation and Cemagref for its financial support. I thank Joe Altonji, Martin Browning, Sidd Chib, Xavier d'Hautfoeuille, Jim Heckman, Stefan Hoderlein, Martin Huber, Fabian Lange, Michael Lechner, Yoonseok Lee, Thierry Magnac, Giovanni Mellace, Christoph Röthe, Julie Subervie, Petra Todd, Neese Yildiz, Ed Vytlacil and seminar participants at Yale, the University of Pennsylvania, CERDI, TSE, CREST, the University of Orléans, the University of Sankt Gallen and the University of Chicago for their valuable comments and suggestions on this research. I also thank the editor, one associate editor and four anonymous referees for their comments on an earlier version of this paper. All remaining errors are my own.

†Correspondence to: Sylvain Chabé-Ferret, Toulouse School of Economics, Lerna 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 compares two different ways of using pre-treatment outcomes when estimating the effect of a program with panel data: including pre-treatment outcomes in the set of control variables or using them in a Difference-In-Difference (DID) strategy.

In their review of recent developments in the econometrics of program evaluation, Imbens and Wooldridge (2009) seem to favor matching as a default option, but call for more substantive knowledge on that issue:

> *In the end, the two approaches make fundamentally different assumptions. One needs to choose between them based on substantive knowledge.* [...] *As a practical matter, the DID approach appears less attractive than the unconfoundedness based approach in the context of panel data. It is difficult to see how making treated and control units comparable on lagged outcomes will make the causal interpretation of their difference less credible, as suggested by the DID assumptions.*

I approach this problem by using a model of earnings and entry into a Job Training Program (JTP) as a working example, thereby bringing substantive prior knowledge from labor economics. I study the conditions on this model that ensure the consistency of either matching or DID. Models of wage or earnings dynamics indeed offer a good rationale for both approaches. They generally include a random intercept and/or a random trend that is unobserved to the econometrician and captures unobserved ability. These types of terms are easily dealt with by time-differencing, which calls for a DID or fixed-effects type of approach. At the same time, these processes include ARMA terms so that transitory shocks persist. A classical stylized fact in labor economics is that individuals entering a JTP experience a transitory decrease in earnings, a phenomenon known as Ashenfelter's dip (Ashenfelter, 1978). This creates time-varying selection bias which calls for matching on past outcomes to capture these transitory declines in earnings. There is thus a tension between these two sources of bias: the one due to selection on a permanent unobserved component and the other

due to transitory income shocks and Ashenfelter's dip. Whether or not matching or DID can get rid of one or both of these bias terms remains an open question. In this paper, I carefully decompose the bias of matching and DID and state sufficient conditions that drive these terms to zero.

My first result is that there are plausible settings in which DID is consistent while matching on past outcomes is not. This is obviously true when selection bias is only due to additive unobserved individual fixed effects. In that case, past outcomes are not a good proxy for unobserved fixed effects and matching is biased whereas differencing gets rid of the unobserved fixed effects and is consistent. I also show a less obvious result that DID is consistent and matching is biased in a model allowing for selection on transitory shocks, even in the absence of fixed effects. This result has more empirical content because it allows for an Ashenfelter's dip. This is thus a credible case where making treated and control units comparable on lagged outcomes makes the causal interpretation of their difference less credible than the DID assumptions. This result nuances the intuition expressed in Imbens and Wooldridge (2009)'s statement and thus sheds new light on the tension between matching and DID.

This result relies on the fact that DID, when applied symmetrically around the treatment date and under plausible conditions, can get rid of selection bias due to transitory shocks and is thus robust to Ashenfelter's dip. The intuition for the consistency of symmetric DID is simple: under certain conditions, Ashenfelter's dip forms and dissipates at the same pace, thereby generating a symmetric wedge around the treatment date. The conditions for ensuring this symmetry are somewhat stringent, but they can credibly hold simultaneously. The first condition requires that the agents have full information on the wages they would have earned had they not entered the program. The second condition requires that the expectation of earnings conditional on some increasing transformation of the net utility of entering the program is linear. This obviously holds when error terms are normally distributed, but also extends to the whole class of elliptical disturbances (Chu, 1973), including for instance the lognormmal, Student's t and Cauchy distributions. The third condition imposes that

the idiosyncratic component of the wage process is stationary. This holds if there is no random slope term, transitory shocks are mean-reverting (*i.e.* there is no random walk) and initial conditions are drawn in the long run stable distribution.

Under these conditions, matching is biased. Indeed, in order to ensure the consistency of matching, I need to assume away both the random slopes and intercepts and the moving average terms in the dynamics of earnings and I have to impose that agents cannot perfectly forecast their forgone earnings.

My second main result is that the consistency of both estimators unfortunately relies on conditions that are at odds with properties of earnings dynamics (see Meghir and Pistaferri (2011) for a survey). There is a heated debate in the literature on empirical income dynamics as to whether random slopes and intercepts are needed to account for income dynamics or not. The Heterogeneous Income Profile (HIP) model advocated by Guvenen (2007) includes these terms. The Restricted Income Profile (RIP) model does not (MaCurdy, 1982). Both estimators are not consistent under the HIP. Under the RIP, and apart from restrictions on agent's information set, the consistency of DID also requires stationarity while that of matching requires the absence of MA terms. Both of these conditions are at odds with empirical estimates of the earnings process (Meghir and Pistaferri, 2011). For example, MaCurdy (1982) estimates sizable MA terms and an increasing variance over the life cycle.

The third contribution of this paper is to explore the sensitivity of both estimators to deviations from the specific conditions under which they are consistent and to study their small sample properties. I conduct a calibration exercise and Monte-Carlo simulations using estimates of the wage process from the literature. I vary several dimensions of the model: HIP *vs* RIP, full information *vs* limited information and initial conditions. I also check the sensitivity of both estimators to the inclusion of additional pre-treatment outcomes in the set of control variables. Results from the simulations show that both estimators are severely biased when the model deviates from the optimal conditions. The results show that, although matching is generally not consistent under credible parameterizations, its behavior is less erratic: it generally

underestimates the true treatment effect and thus provides a useful lower bound and its bias generally decreases when controlling for additional pre-treatment outcomes. DID is consistent under credible conditions, but is more sensitive to deviations from these conditions. Moreover, the different bias terms generated by deviations from these conditions do not always go in the same direction: lower initial variance of the ARMA process yields to an underestimation of the treatment effect, as does the introduction of a random slope, whereas limited information yields to an overestimation of the true treatment effect. Finally, adding more control variables increases the bias of DID in the RIP, but decreases it in the HIP.

The fourth contribution of this paper is to revisit the results of studies comparing matching and DID to an experimental benchmark (Heckman, Ichimura, Smith, and Todd, 1996; Smith and Todd, 2005). I point to previously unnoticed results in both papers favoring the use of symmetric DID. In these applications, matching controlling for past wages exhibits higher bias than DID.

Overall the combined results in this paper seems to picture a view somewhat more favorable to DID than the one advocated by Imbens and Wooldridge (2009). At the very least, presenting the results of both approaches and the sensitivity of those to the inclusion of past outcomes as control variables is to be recommended when applying matching on panel data.

The results in this paper shed also light on the evaluation of JTP with repeated cross-sectional data. Because only DID can be applied in this case, the results in this paper help understand the likely sources of bias with this approach. Implementing DID symmetrically around the treatment date is recommended in order to capture Ashenfelter's dip. The results in this paper also shed light on the evaluation of programs with similar selection rules and outcome processes, as for example the effect of payments for environmental services on agricultural practices (Chabé-Ferret and Subervie, Forthcoming), statutory sick pay on health (Puhani and Sonderhof, 2010; Ziebarth and Karlsson, 2010), fair trade certification on product quality and income (Balineau, 2012) and the effect of enterprise zones on firms' location (Mayer, Mayneris,

and Py, 2012). However, the results results in this paper do not apply to duration outcomes as the length of unemployment spells for instance, except when unemployment duration is averaged over geographical area (Gobillon, Magnac, and Selod, 2010).

While the trade-off between matching and DID has never been studied *per se* in the literature, the choice of control variables when using matching has already been studied. Wooldridge (2005) shows that controlling for variables altered by the treatment generates selection bias. Heckman and Navarro-Lozano (2004) show, in a static selection model, that including an additional variable to the set of control variables may increase selection bias. More recently, several papers have shown that controlling for instrumental variables amplifies bias when it is already present (Bhattacharya and Vogt, 2007; Wooldridge, 2009; Pearl, 2010, 2011; Myers, Rassen, Gagne, Huybrechts, Schneeweiss, Rothman, Joffe, and Glynn, 2011). The consistency of symmetric DID with time varying selection bias is an extension of a similar result in Heckman (1978) to a more general selection rule and wage process. I also extend the Monte-Carlo results of Heckman, LaLonde, and Smith (1999) to the HIP model, varying both initial conditions and agents' information set.

This paper is structured as follows: section 2 presents the model and the estimators I consider; section 3 presents separate sufficient conditions ensuring the consistency of matching and DID; in sections 4 and 5, I report the results of a calibration exercise and Monte-Carlo simulations checking the sensitivity of these results to deviations from the sufficient conditions; section 6 points to previously unnoticed results in the literature comparing matching and DID to an experimental benchmark and concludes.

2 The setting: a model of the wage process and of entry into a job training program

In order to give economic content to the results in this paper, I study the canonical case of a Job Training Program (JTP), as in Heckman and Robb (1985). I model

individuals facing an exogenous stochastic wage process and deciding whether or not to enter a JTP that is available only for one period. As in Heckman and Navarro-Lozano (2004), I vary the information that agents have when deciding to enter the program. In the remaining of this section, I describe each component of the model in turn and I end with a description of the estimation strategies that I am comparing for the estimation of the average treatment effect on the treated.

Wage dynamics

I use a model of earnings dynamics that nests the two leading views in the literature on the nature of the labor income process (Meghir and Pistaferri, 2011): the so-called heterogeneous income profile (HIP), that allows for a random idiosyncratic trend in income (Guvenen, 2007, 2009), and the restricted income profile (RIP) that does not (MaCurdy, 1982). The log-wage process of individual i at time t in the absence of the treatment has the following form:

$$Y_{i,t}^0 = g(X_i, \delta_t) + \mu_i + \beta_i t + U_{it} \tag{1a}$$

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

$$v_{i,t} \text{ i.i.d. mean-zero shocks with finite variance } \sigma^2, \tag{1c}$$

$$v_{i,t} \perp\!\!\!\perp (X_i, \beta_i, \mu_i), \forall t, \tag{1d}$$

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

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

with μ_i and β_i fixed factors unobserved by the econometrician, X_i a set of observed variables,[1] δ_t an economy-wide shock and $U_{i,t}$ an ARMA(1,2) process.[2,3] Note that

[1] I abstract from the problem of time varying covariates other than past outcomes in this analysis. A previous version of the paper available on my webpage (`https://sites.google.com/site/sylvainchabeferret/research`) develops some results for that case.

[2] I assume that $v_{i,t}$ is i.i.d. mostly for convenience. The results in this paper could accommodate heteroskedasticity at the individual level, except for economy-wide variations in the variance of log-wages. Then the wage process would not be stationary and symmetric DID would not be consistent.

[3] I assume that the error term is an ARMA(1,2) mostly for convenience. The results in this paper generalize to an arbitrary ARMA(p,q).

this model allows for the common time shock to interact with observed characteristics. I note Σ_t the covariance matrix of $(U_{i,t}, v_{i,t}, v_{i,t-1})$ and $\Sigma_\infty = \lim_{t\to\infty} \Sigma_t$. Finally, $Y^1_{i,t}$ denotes the log-wage of agent i after she has received the treatment. I leave this process unspecified. Let $\alpha_{i,t} = Y^1_{i,t} - Y^0_{i,t}$ denote the individual level causal effect of the program on log wages.

Selection rule

As in Heckman and Robb (1985) and Heckman, LaLonde, and Smith (1999), agents are offered the possibility of entering a JTP at period k, and only at this period. I assume that agents consume all their income at each period.[4] Let $W^0_{i,t} = \exp(Y^0_{i,t})$ denote wages in levels, $T_{i,k}$ transfers received by the agents if they enter the program and $C_{i,k}$ direct costs of the program for the agents. Let G denote a time separable strictly concave utility function, $\frac{1}{1+r}$ the discount rate used by the agents and T the total number of periods of their working life. Maximization of expected discounted utility yields to the following program participation rule:

$$D^\iota_{i,k} = \mathbb{1}[\mathbb{E}[\sum_{j=1}^{T-k} \frac{G(W^1_{i,k+j}) - G(W^0_{i,k+j})}{(1+r)^j} + G(T_{i,k} - C_{i,k}) - G(W^0_{i,k})|\mathcal{I}^\iota_{i,k}] \geq 0]. \quad (2)$$

where $\mathcal{I}^\iota_{i,k}$, denotes the information set of agent i when she considers entering the program (more on this below).

In order to simplify the formulation, I assume that $G = \ln$ and $T \to \infty$ and that the effect of the treatment is constant over time ($\alpha_{i,t} = \alpha_i, \forall t$). This yields to the following simple participation rule:

$$D^\iota_{i,k} = \mathbb{1}[\underbrace{\frac{\alpha_i}{r} - c_i - \mathbb{E}[Y^0_{i,k}|\mathcal{I}^\iota_{i,k}]}_{D^{*\iota}_{i,k}} \geq 0], \quad (3)$$

with $c_i = -\ln(T_{i,k} - C_{i,k})$ and $D^{*\iota}_{i,k}$ the index measuring the value of entering the

[4]Note that I could have assumed perfect credit markets, as in Heckman, LaLonde, and Smith (1999). Under this assumption, agents would maximize discounted income. It seems more credible that agents participating in a JTP cannot borrow in order to finance it.

program. Selection into the program is driven by gains from program participation ($\frac{\alpha_i}{r}$), direct costs (c_i) and opportunity costs that takes the form of expected forgone earnings.[5] This simple selection rule can generate an Ashenfelter's dip as long as the agent's information set is correlated to transitory idiosyncratic shocks to past earnings. Note that I allow for α_i and c_i to be correlated to μ_i and β_i, so that even if the econometrician observes $\mathbb{E}[Y^0_{i,k}|\mathcal{I}^\iota_{i,k}]$, there still is selection on unobservables. Because agents can only enter the program at period k, we have:

$$D^\iota_{i,t} = \begin{cases} 0 & \text{if } t < k \\ \\ D^\iota_{i,k} & \text{if } t \geq k \end{cases}, \tag{4}$$

along with the usual switching model governing observed outcomes:

$$Y_{i,t} = D^\iota_{i,t} Y^1_{i,t} + (1 - D^\iota_{i,t}) Y^0_{i,t}, \text{ if } t \neq k, \tag{5}$$

$$Y_{i,k} = (1 - D^\iota_{i,k}) Y^0_{i,k} \tag{6}$$

For simplicity, I omit the dependence of $Y_{i,t}$ on ι. Note that wages for the treated are unobserved at period k, when agents are actually enrolled in the program. We observe a zero wage, but it is not equal to the wage the agents would have earned had the program not existed. Because wages are censored at period k, we cannot use period-k wages as control variables.

Agents' information set

I vary the information that agents have on their forgone earnings when they consider entering the program. I consider four different informational contents:

[5] I assume that the only uncertainty the agent faces is with respect to forgone earnings. As suggested by a referee, the agents could also learn about their idiosyncratic gains α_i. Modelling how past earnings inform agents about their gains from program participation is a very nice area for further research but is beyond the scope of this paper.

(i) Agents know all the shocks up to period k:

$$\mathcal{I}_{ik}^{f} = \left\{ X_i, \alpha_i, c_i, \mu_i, \beta_i, \{\delta_j\}_{j=0}^{k}, \{v_{i,j}\}_{j=0}^{k} \right\}. \tag{7}$$

In that case, because I assume that agents know the parameters of the wage process, they can perfectly forecast their forgone earnings: $\mathbb{E}[Y_{ik}^{0}|\mathcal{I}_{ik}^{f}] = Y_{ik}^{0}$.

(ii) Agents does not know the last idiosyncratic shock to their earnings:[6]

$$\mathcal{I}_{ik}^{l} = \left\{ X_i, \alpha_i, \beta_i, c_i, \mu_i, \{\delta_j\}_{j=0}^{k}, \{v_{i,j}\}_{j=0}^{k-1} \right\}. \tag{8}$$

Limited information can arise because agents have to decide whether or not to enter the program in period k at the end of period $k-1$, not knowing the last innovation to their earnings. Note that in that case they forecast their forgone earnings with the information at hand: $\mathbb{E}[Y_{i,k}^{0}|\mathcal{I}_{i,k}^{l}] = g^{0}(X_i, \delta_k) + \mu_i + \rho U_{i,k-1} + m_1 v_{i,k-1} + m_2 v_{i,k-2}$.

(iii) Agents only know time and individual fixed effects:

$$\mathcal{I}_{ik}^{c} = \left\{ X_i, \alpha_i, c_i, \mu_i, \beta_i, \{\delta_j\}_{j=0}^{k} \right\}. \tag{9}$$

With this very coarse information set, there is no Ashenfelter's dip.

(iv) Agents observe their own earnings up to period $k-1$, the overall shocks to the economy and they have initial information about their unobserved intercept and trend:

$$\mathcal{I}_{ik}^{b} = \left\{ X_i, \alpha_i^k, \beta_i^k, c_i, \{\delta_j\}_{j=0}^{k}, \{Y_{i,j}\}_{j=0}^{k-1} \right\}, \tag{10}$$

where $\mu_i = \mu_i^k + \mu_i^u$ and $\beta_i = \beta_i^k + \beta_i^u$. Agents update their prior on the distribution of the unobserved variables at every period after observing their realized wage. The prior has covariance matrix $\mathbf{P}_{1|0}$. I closely follow Guvenen (2007)'s bayesian

[6]Note that I assume that agent know the shock to the overall economy δ_k. This is only for comparability with the full information case: I just vary information components one at a time.

learning model (see appendix C for a detailed description).

Parameters and estimators

The causal effect of interest is the average treatment effect on the treated τ periods after the treatment on log-wages:[7] $ATT_{\tau,\iota} = \mathbb{E}[Y^1_{i,k+\tau} - Y^0_{i,k+\tau} | D^\iota_{i,k+\tau} = 1]$, where $\tau > 0$. In order to save notation, I note $\Delta^{Y_i}_{\tau,\tau'} = Y_{i,k+\tau} - Y_{i,k-\tau'}$, with $\tau, \tau' > 0$. I compare the properties of two estimators, $D\hat{I}D_{\tau,\tau'}$ and $\hat{M}_{\tau,\tau'}$ such that, for any $(\tau, \tau') > 0$:[8]

$$\text{plim} D\hat{I}D_{\tau,\tau',\iota} = \mathbb{E}[\Delta^{Y_i}_{\tau,\tau'} - \mathbb{E}[\Delta^{Y_i}_{\tau,\tau'} | X_i, D^\iota_{i,k+\tau} = 0] | D^\iota_{i,k+\tau} = 1], \tag{11}$$

$$\text{plim}\hat{M}_{\tau,\tau',\iota} = \mathbb{E}[Y_{i,k+\tau} - \mathbb{E}[Y_{i,k+\tau} | X_i, Y_{i,k-\tau'}, D^\iota_{i,k+\tau} = 0] | D^\iota_{i,k+\tau} = 1]. \tag{12}$$

I also consider an unfeasible version of the matching estimator where I allow for conditioning on the censored $Y^0_{i,k}$. By convention, I write, $\forall \tau > 0$:

$$\text{plim}\hat{M}_{\tau,0,\iota} = \mathbb{E}[Y_{i,k+\tau} - \mathbb{E}[Y_{i,k+\tau} | X_i, Y^0_{i,k}, D^\iota_{i,k+\tau} = 0] | D^\iota_{i,k+\tau} = 1]. \tag{13}$$

In this paper, I study the asymptotic bias of these two estimators of ATT_τ:

$$B^{DID}_{\tau,\tau',\iota} = \text{plim} D\hat{I}D_{\tau,\tau',\iota} - ATT_{\tau,\iota} \tag{14}$$

$$B^{M}_{\tau,\tau',\iota} = \text{plim}\hat{M}_{\tau,\tau',\iota} - ATT_{\tau,\iota}. \tag{15}$$

In order to save space, I will note, for any two random variables T_i and Z_i, the average conditional difference between treated and untreated individuals:

$$\mathbb{CD}_\iota(Z_i | T_i) = \mathbb{E}[Z_i | T_i, D^\iota_{i,k} = 1] - \mathbb{E}[Z_i | T_i, D^\iota_{i,k} = 0] \tag{16}$$

[7] Note that it is possible to recover the effect on wages in levels by taking the exponential and multiplying by the average earnings of the treated at period $k + \tau$ in levels.

[8] $\sqrt{N}$-consistent estimators of these quantities can be built using results in the literature (Heckman, Ichimura, and Todd, 1998; Hahn, 1998; Hirano, Imbens, and Ridder, 2003; Abadie, 2005).

In this section, I will repeatedly use the following results:

$$B^M_{\tau,\tau',\iota} = \mathbb{E}[Y^0_{i,k+\tau} - \mathbb{E}[Y^0_{i,k+\tau}|X_i, Y_{i,k-\tau'}, D^\iota_{i,k+\tau} = 0]|D^\iota_{i,k+\tau} = 1], \tag{17}$$

$$= \mathbb{E}[\mathbb{CD}_\iota(Y^0_{i,k+\tau}|X_i, Y_{i,k-\tau'})|D^\iota_{i,k+\tau} = 1]. \tag{18}$$

$$B^{DID}_{\tau,\tau',\iota} = \mathbb{E}[\Delta^{Y^0_i}_{\tau,\tau'} - \mathbb{E}[\Delta^{Y^0_i}_{\tau,\tau'}|X_i, D^\iota_{i,k+\tau} = 0]|D^\iota_{i,k+\tau} = 1], \tag{19}$$

$$= \mathbb{E}[\mathbb{CD}_\iota(\Delta^{Y^0_i}_{\tau,\tau'}|X_i)|D^\iota_{i,k+\tau} = 1]. \tag{20}$$

In section 3, I derive sufficient conditions on the economic model that ensure consistency of these estimators. In section 4, I derive closed form formulas for these bias terms when all the error terms are normally distributed and I simulate them using MaCurdy (1982)'s estimates of the wage process. In section 5, I use Monte-Carlo simulations in order to grasp the small sample properties of both estimators. I also test the sensitivity of my results to the HIP process.

3 Conditions ensuring the consistency of matching and DID

In this section, I first study in more detail the asymptotic bias of the two estimators, and I derive conditions for their consistency. I derive two sets of conditions for $\hat{M}_{\tau,\tau'}$ to be consistent. I derive sufficient conditions for $D\hat{I}D_{\tau,\tau'}$ and symmetric DID ($D\hat{I}D_{\tau,\tau}$) to be consistent. Under these two sets of restrictions, $\hat{M}_{\tau,\tau'}$ is biased.

Consistency of matching estimators

In this section, I derive a two sets of sufficient conditions for matching estimators using only one pre-treatment outcome as a control variable to be consistent.

Proposition 1 (Infeasible matching) *If* $(\mu_i, \beta_i) \perp\!\!\!\perp (c_i, \alpha_i)|X_i$, *then* $B^M_{\tau,0,f} = 0$, $\forall \tau > 0$.

Proof: See in appendix A. ■

Proposition 1 shows that, under full information and if there is no selection on unobservables, matching on forgone earnings is consistent. Unfortunately, forgone earnings are generally unobserved, rendering this estimator infeasible. The following proposition states one set of conditions for a feasible matching estimator with $\tau' > 0$ to be consistent:

Proposition 2 (Feasible matching estimator) *If $F_{\mu,\beta}$ is degenerate and if $m_1 = m_2 = 0$, then $B^M_{\tau,1,l} = 0$, $\forall \tau > 0$.*

Proof: See in appendix A. ■

Proposition 2 rests on the following decomposition of the bias of matching conditional on $(X_i, Y^0_{i,k-1})$ that is an intermediate output of the proof of proposition 2 (it is valid for $\tau \geq 2$):

$$\mathbb{CD}_\iota(Y^0_{i,k+\tau}|X_i, Y_{i,k-1}) = (1-\rho^{\tau+1})\mathbb{CD}_\iota(\mu_i|X_i, Y_{i,k-1}) \tag{21a}$$

$$+ (k+\tau-\rho^{\tau+1}(k-1))\mathbb{CD}_\iota(\beta_i|X_i, Y_{i,k-1}) \tag{21b}$$

$$+ \rho^{\tau-1}(\rho m_1 + m_2)\mathbb{CD}_\iota(v_{i,k-1}|X_i, Y_{i,k-1}) \tag{21c}$$

$$+ \rho^{\tau} m_2 \mathbb{CD}_\iota(v_{i,k-2}|X_i, Y_{i,k-1}) \tag{21d}$$

$$+ \rho^{\tau-2}(\rho^2 + \rho m_1 + m_2)\mathbb{CD}_\iota(v_{i,k}|X_i, Y_{i,k-1}). \tag{21e}$$

The terms (21a) and (21b) are due to selection on the random intercept and random trend. They cancel out when when $F_{\mu,\beta}$ is degenerate. The terms (21c) and (21d) are due to selection on recent shocks through the moving average terms: the econometrician does not observe the outcome of these corrections, but they are correlated with expected forgone earnings. These terms cancel when $m_1 = m_2 = 0$. The last term (21e) is due to the last shock to earnings that the agent observes under full information but that the econometrician cannot observe. This term obviously cancels out under limited information.

Proposition 2 shows that in order to ensure that the feasible matching estimator is consistent, we need the following conditions: limited information, so that the last

shock to the wages is unknown to the agents when they decide to enter the program; no fixed effects and no moving average terms. These conditions impose selection on the observables $(X_i, Y^0_{i,k-1})$. Under these restrictions, we indeed have: $D^{l*}_{i,k} = \frac{\alpha_i}{r} - c_i - g(X_i, \delta_k) + \rho g(X_i, \delta_{k-1}) - \mu(1-\rho) - Y^0_{i,k-1}$. As $(\alpha_i, c_i) \perp\!\!\!\perp U_{i,k+\tau} | (X_i, Y^0_{i,k-1})$, we have $D^l_{i,k} \perp\!\!\!\perp Y^0_{i,k+\tau} | (X_i, Y^0_{i,k-1})$.

One could argue that conditioning on additional pre-treatment outcomes could ensure the consistency of matching under less stringent conditions. First, note that the bias term under full information is due to the last shock to earnings being unobserved by the econometrician. This is going to remain whatever the amount of pre-treatment data we observe. Consistency of matching thus at the very least requires limited information. Second, one could condition on a sufficient statistic for μ_i, for example by using the average of pre-treatment wages over T periods. But unfortunately, this proxy would converge to μ_i only as T becomes large. Under finite T, this estimator is biased. This problem is akin to an incidental parameters problem in classical panel data estimation. Third, it seems that we could get rid of the bias due to the moving average terms by controlling for two or three pre-treatment periods instead of one. Fourth, when the agent observes only her own wages up to period $k-1$ ($\iota = b$), and if her initial prior on (α_i, β_i) is not informative (*i.e.* if F_{μ^k, β^k} is degenerate), then matching controlling for all past wages ($\{Y_{i,j}\}_{j=0}^{k-1}$) is consistent. Note however that dimensionality of the control set is very large. I explore how these fixes perform in the Monte-Carlo simulations presented in section 5.

Consistency of DID estimators

In this section, I state two sets of sufficient conditions for DID matching on X_i to be unbiased. Under these conditions, matching on $(X_i, Y_{i,k-\tau'})$ is biased. The second set of conditions allows for an Ashenfelter's dip.

Proposition 3 (DID without dip) *Under coarse information and if F_β is degenerate, the DID matching estimator is consistent: $B^{DID}_{\tau,\tau',c} = 0$, $\forall \tau, \tau' > 0$. Under the same*

set of conditions, matching is biased: $B^{M}_{\tau,\tau',c} \neq 0$, $\forall \tau, \tau' > 0$.

PROOF: See in appendix A. ■

Proposition 3 thus shows that, in the RIP model, if agents do not know the transitory shocks and there is selection on an unobservable fixed effect, DID is unbiased whereas matching is. Conditioning on pre-treatment outcomes thus generates bias in this case: increments are independent of the treatment conditional on X_i, but they are not independent of treatment conditional on $(X_i, Y_{i,k-\tau'})$. This is because forcing treated and untreated individuals to have similar pre-treatment outcomes generates correlation between pre-treatment transitory shocks and the treatment: individuals having large values of their fixed effects and receiving the same pre-treatment wages as the treated have experienced a series of negative transitory shocks. The influence of these shocks is going to progressively fade away, thereby generating bias. Note that I have assumed away the random trend to obtain this result. With a random trend, consistency could be restored by using the matching version of the triple difference estimator of Heckman and Hotz (1989).

The following proposition deals with the more credible case of selection both on permanent and transitory unobserved income shocks.

Proposition 4 (DID with dip) *Under full information, if* F_β *is degenerate,* $|\rho| < 1$, $\Sigma_0 = \Sigma_\infty$ *(or* $k \to \infty$*) and* $\mathbb{E}[U_{i,t}|D^{f*}_{i,k}, X_i]$ *is linear in* $D^{f*}_{i,k}$, $\forall t$, *the symmetric DID matching estimator is consistent:* $B^{DID}_{\tau,\tau,f} = 0$, $\forall \tau > 0$. *Matching is biased under the same set of conditions:* $B^{M}_{\tau,\tau',f} \neq 0$, $\forall \tau, \tau' > 0$.

PROOF: See in appendix A. ■

Under the conditions in proposition 4, Ashenfelter's dip, and thus selection bias, is symmetric around the treatment date and as a consequence, symmetric DID is consistent. This is an extension of a result of Heckman (1978) to a more complex selection rule and to the matching version of DID. This result follows from the stationarity of the $U_{i,t}$ process and the linearity of the expectation of $U_{i,t}$ conditional on $D^{*\iota}_{i,k}$ that together imply that the dip forms and dissipates at the same pace.

This result rests on the following decomposition of the bias of DID under the assumption of linear conditional expectation and for $\iota \in \{f, l\}$ and $(\tau, \tau') > 0$ (see the proof of proposition 4):

$$B^{DID}_{\tau,\tau',\iota} = (\tau + \tau')\mathbb{E}[\mathbb{CD}_\iota(\beta_i|X_i)|D^\iota_{i,k} = 1] \tag{22a}$$

$$+ \left(\rho^{\tau'}\text{Var}(U_{i,k-\tau'}) - \rho^{\tau}\text{Var}(U_{i,k})\right) A^\iota_k \tag{22b}$$

$$+ \mathbb{1}[\iota = l](\rho^{\tau-2})(\rho^2 + m_1\rho + m_2)\sigma^2 A^\iota_k, \tag{22c}$$

Part (22a) of the bias term is due to selection on the random slopes. It disappears when F_β is degenerate. The second term is due to selection on transitory shocks. When $\tau = \tau'$ (*i.e.* when DID is applied symmetrically), this term cancels out if the variance of $U_{i,t}$ is constant over time. This is the case when the ARMA process is at its long run equilibrium.[9] Finally, the last term is due to limited information: when agents do not know the last shock to their earnings, the covariance between $U_{i,k+\tau}$ and the treatment index is not proportional to the variance of $U_{i,k}$. This term cancels out under full information.

Matching is biased in this model for three reasons: first, the selection on unobserved fixed effects (equation (21a)), second, the selection on the unobserved shock to earnings at period k (equation (21e)) and finally the relative composition of recent *vs* past shocks due to the moving average terms (equations (21c) and (21d)).

Discussion

The general take away message of this section is that it is possible to find restrictions that ensure that either matching or DID are consistent. A nice feature of DID is that

[9]Note that if $U_{i,t}$ follows a random walk (*i.e.* $\rho = 1$), the long run variance is infinite and the process has no long run equilibrium. The variance of $U_{i,t}$ increases with t and as a consequence, the bias term (22b) is negative and DID underestimates the true treatment effect. Note that when $\rho = 1$, it is possible to use a modified version of an insight of Heckman (1978) to build a consistent DID estimator with the additional assumption that F_μ is degenerate. In a random walk, the variance is proportional to time. If $F_{\mu,\beta}$ is degenerate and under full information, the bias term (22b) is equal to the difference in variances at periods $k - \tau'$ and k. Properly rescaling $Y_{i,k-\tau'}$ by a factor $\frac{k-3}{k-\tau'-3}$ cancels this bias term. This assumes that the order of the MA component is known.

it can get rid of selection bias due both to unobserved fixed effects and time-varying idiosyncratic shocks when applied symmetrically around the treatment date. I have not been able to find instances where matching is consistent without assuming away selection on the fixed effects.

I have stated two sets of restrictions under which DID is consistent whereas matching is biased. Thus, contrary to what Imbens and Wooldridge (2009)'s citation seems to suggest, it is possible to find instances where the DID assumptions are more credible than the unconfoundedness based approach. Furthermore, this result still holds even when assuming away selection on fixed effects and allowing for selection on time-varying idiosyncratic shocks. Symmetric DID is consistent in that case, under a set of restrictions that I make precise. Under the same set of restrictions, matching is biased.

In order to assess the empirical relevance of these results, I compare the restrictions that ensure the consistency of each estimator with estimates taken from the literature on earnings dynamics. There is a heated debate in the literature on empirical income dynamics as to whether random slopes and intercepts are needed to account for income dynamics or not. The Heterogeneous Income Profile (HIP) model advocated by Guvenen (2007) includes these terms. The Restricted Income Profile (RIP) model does not (MaCurdy, 1982). The theoretical advantage of symmetric DID (being able to solve both selection on a fixed effect and time varying selection bias) does thus not import in practice. Under the HIP, the random trend term biases both symmetric DID and matching.

Under the RIP, both estimators can be consistent. Apart from specific restrictions on agent's information set, the consistency of symmetric DID also requires stationarity and linearity of conditional expectations while that of matching requires the absence of MA terms. It is a general result from the literature estimating earnings processes that MA terms are needed to account for their dynamics and that the variance of earnings increases over time (MaCurdy, 1982; Meghir and Pistaferri, 2011). As a consequence, both estimators are likely to be biased in applications. Note that symmetric DID could be unbiased when the variance of the $U_{i,t}$ process stabilizes (if it ever does). In general,

authors using the RIP model impose the existence of permanent shocks following a random walk ($\rho = 1$). MaCurdy (1982)'s estimates for ρ oscillate between 0.975 and 0.995. This is very close to a random walk (MaCurdy (1982) does not reject the existence of a random walk) but these coefficients still characterize a stationary process. The main problem for symmetric DID is that these coefficients imply that reaching the long run stable distribution can take time. Finally, the linearity of the conditional expectation may seem rather restrictive. It is obviously true for jointly normally distributed variables, but is also a property of the much larger family of elliptical disturbances (Chu, 1973), that include Student's t and the Cauchy distributions for example.

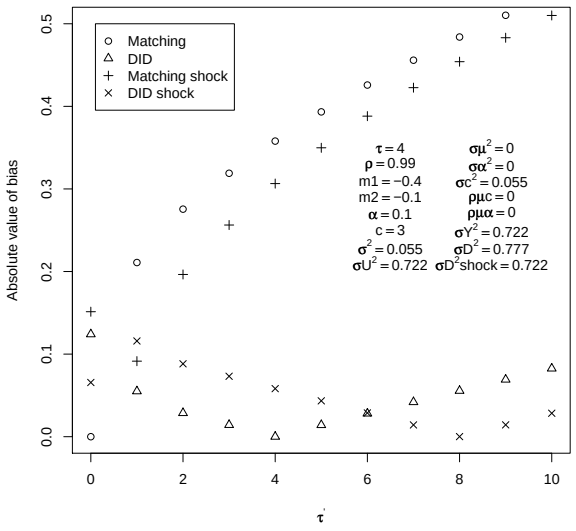
Overall, the conditions for the consistency of both estimators seem to be at odds with stylized properties of earnings processes. But these biases may be small in real applications. In what follows, I assess the relative severity of these biases using a calibration exercise and Monte-Carlo simulations.

4 Calibration results: size of the bias of matching and DID in the RIP model

The aim of this section is to gain more insight in the sources of bias of both estimators in the RIP model and to assess the likely importance of the bias of matching under limited information, where only MA terms are at play. I derive closed form formulas for the asymptotic bias terms of matching and DID in the model laid out in section 2 under the RIP, with error terms at their long run equilibrium and in the case of normally distributed error terms. I then calibrate these formulas with MaCurdy (1982)'s estimates of the wage process and compare the size of the bias of these two estimators as a function of the period at which we control for and the agent's information set. The derivation of the bias terms under normally distributed disturbances can be found in appendix B.

Figure 1 presents the absolute values of the bias terms of DID matching (equation 19) and simple matching (equation 17) with the values of the parameters estimated by MaCurdy (1982) under two information sets: full and limited. This figure illustrates the results from previous section and allows to compute how the asymptotic bias varies when moving away from the conditions ensuring consistency.

Figure 1 – Absolute value of bias of matching and DID for $\tau = 4$ under the RIP and with a stationary ARMA process

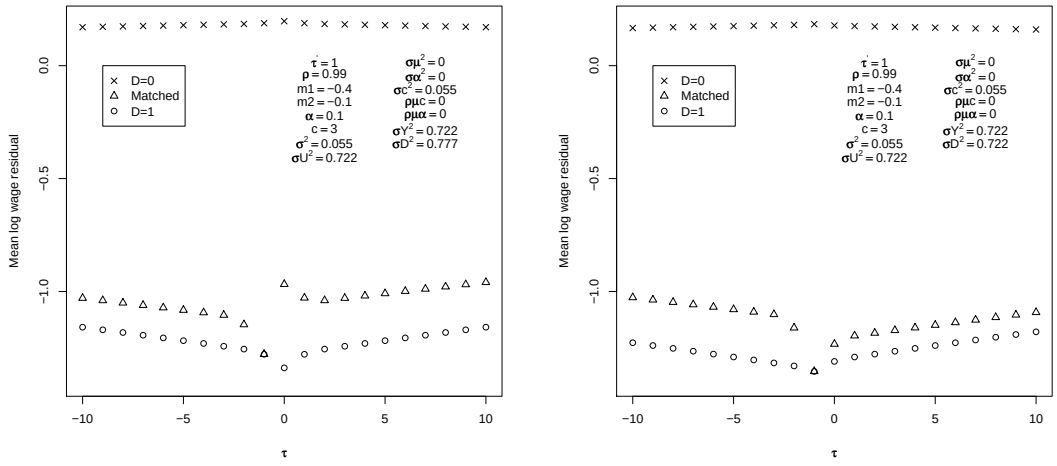


Note: the value of the bias for DID (resp. simple) matching are the absolute values of the terms in equation (19) (resp. (17)) under full information ($\iota = f$) computed in appendix B and calibrated with MaCurdy (1982)'s estimates of the log-wage process. τ indicates the number of period after assignment to treatment. τ' measures the period at which matching is performed. (shock) corresponds to the value of the corresponding terms under limited information $\iota = l$ (*i.e.* when $v_{i,k}$ is not observed when the individual decides to participate).

Note first that, under full information, the infeasible matching estimator that controls for the unobserved opportunity cost of the treatment ($Y^0_{i,k}$) is consistent. This is because MaCurdy (1982) estimates that the variance of the fixed effect μ_i is null, thereby complying with the conditions of proposition 1. Moving from this infeasible estimator to the next best feasible alternative, matching on $Y_{i,k-1}$, generates bias. With MaCurdy (1982)'s estimates of the wage process, this bias is equal to 200% of the assumed treatment effect (0.1). There is thus a very large loss when moving from the infeasible matching estimator to the feasible one.

Figure 2(a) provides an illustration of what drives the bias in matching under full information. This figure represents the average potential outcomes in the absence of the treatment ($Y^0_{i,t}$) for the treated, the untreated and the matched untreated, *i.e.* the untreated agents that have the same potential outcomes at period $k-1$ as the treated.[10]

Figure 2 – Evolution of average potential outcomes for the treated, the untreated and the matched untreated under the RIP and with a stationary ARMA process



(a) Full information (b) Incomplete information

Note: this figure plots the average potential outcomes in the absence of the treatment net of economy-wide shocks for three groups: the treated (○), the untreated (×) and the matched untreated (△), *i.e.* untreated that have the same potential outcomes at period $k-1$ as the treated. Because the variance of μ is null in these simulations, these curves respectively stand for $\mathbb{E}[U_{i,k+\tau}|D^\iota_{i,k}=1]$, $\mathbb{E}[U_{i,k+\tau}|D^\iota_{i,k}=0]$ and $\mathbb{E}[\mathbb{E}[U_{i,k+\tau}|D^\iota_{i,k}=0,Y_{i,k-1}]D^\iota_{i,k}=1]$. The outcomes are simulated from the formulas derived in appendix B calibrated with MaCurdy (1982)'s estimates of the log-wage process. τ indicates the number of period after assignment to treatment. τ' measures the period at which matching is performed. ρ is the autocorrelation parameter. (shock) corresponds to the value of the corresponding terms under limited information $\iota = l$ (*i.e.* when $v_{i,k}$ is not observed when the individual decides to participate).

The difference between the treated and the untreated measures overall selection bias before matching. The difference between the treated and the matched untreated measures the bias of the matching estimator. The difference in the difference between the treated and the untreated at period $k+\tau$ and at period $k-\tau'$ measures the bias of the DID estimator applied at both periods.

[10] I choose to present these averages net of the economy-wide shocks $g(X_i,\delta_t)$, in order to make the graph more easy to read. Because the variance of μ is null in these simulations, these curves respectively stand for $\mathbb{E}[U_{i,k+\tau}|D^\iota_{i,k}=1]$, $\mathbb{E}[U_{i,k+\tau}|D^\iota_{i,k}=0]$ and $\mathbb{E}[\mathbb{E}[U_{i,k+\tau}|D^\iota_{i,k}=0,Y_{i,k-1}]D^\iota_{i,k}=1]$.

We can see that the treated experience a transitory decrease in their earnings when they enter the treatment: this is Ashenfelter's dip. It is due to a series of negative transitory shocks that drives the opportunity cost of entering the treatment down. Matching on $Y_{i,k-1}$ imperfectly mimics this process. It selects untreated individuals that also experience a series of negative transitory shocks up to period $k-1$. But, under full information, these individuals do not participate because they receive good news on their earnings at period k (*e.g.* they expect to receive a promotion or a bonus) which drives their opportunity cost of entering the treatment up and drives them out of the treatment. At the same time, this last unobserved shock persists and creates a wedge between the average wages of the treated and that of their matched counterparts. This yields to negative selection bias and would underestimate the true effect of the treatment.

The second lesson from figure 1 is that DID applied symmetrically around the treatment date (here at $\tau' = 4 = \tau$) is consistent under full information. This is an instance of the more general result of proposition 4. Figure 2(a) illustrates why this estimator is unbiased: Ashenfelter's dip is symmetric, and thus selection bias forms and dissipates at the same pace around the treatment date.

The third lesson from figure 1 is that DID is inconsistent under limited information, as was expected after the results in the previous section. DID is biased because Ashenfelter's dip is no longer symmetric: agents do not know the last shock to their earnings and thus the covariance between wages and the selection index is larger before the treatment date. The lower point of the dip is at $k-1$, and thus selection bias is not symmetric around the treatment date, nor is it around $k-1$ for that matter. Note that it seems that DID is consistent under limited information for $\tau' = 8 = 2\tau$. This result is an artifact from improper scaling: the true bias term is non null, albeit very close to zero. Indeed, because selection bias is larger in pre-treatment periods, we have to go further away in time to find a period where pre-treatment selection bias is similar to the post-treatment one. The fact that this happens at $\tau' = 2\tau$ is an artifact

from the parameterization, and does not seem to be a general result.[11] Note that the bias term is positive in that case: DID overestimates the true treatment effect, because it overestimates the size of the selection bias.

The fourth lesson from figure 1 is that matching is also inconsistent under limited information, as was expected after the results in the previous section. Since there are no fixed effects under MaCurdy (1982)'s parameterization, the bias term of matching under limited information is due to the moving average terms. Figure 2(b) illustrates this result: before period $k-1$, the matched untreated have higher earnings than the treated. They thus experience a sharper decrease in their earnings just before period $k-1$, that makes them comparable to the treated. These very shocks are nevertheless corrected by the negative moving average terms (not all the innovation passes through to the next period). This correction yields to slightly higher wages for the matched untreated at period k that persist thereafter, generating negative selection bias and an underestimation of the true treatment effect.

The last lesson drawn from figure 1 is the relative size of the asymptotic bias terms of matching and DID matching under limited information. Matching at period $k-1$ yields a sizable bias term of 95% of the treatment effect. Symmetric DID matching generates a lower bias of 65% of the treatment effect. It thus seems that symmetric DID matching is to be preferred even under limited information.

These results nevertheless require the RIP, that the initial conditions are set at their long run values and that we use only one period of lagged outcomes as control variables. In the next section, I perform Monte-Carlo simulations to assess the sensitivity of these results to deviations from these assumptions.

[11] In fact, using equation (22), one can show that the bias of DID cancels when $\tau' = \tau + 2$, under limited information and when both MA terms are null. In the presence of MA terms, this result does not hold.

5 Simulation results: sensitivity of matching and DID to deviations from optimal conditions

In this section, I present results from Monte-Carlo simulations assessing the small sample performances of matching and DID and the sensitivity of both estimators to deviations from the conditions ensuring their consistency. I simulate both the HIP and the RIP versions of earnings dynamics varying initial conditions, agents' information set and the number of pre-treatment outcomes used as control variables.

Setting

The wage process is simulated according to equation (1). Time goes from $t = 1$ to $t = 40$, in order to model the working lifetime of an individual. The set of control variables X_i includes years of education E_i and experience $A_{i,t}$. The function g has two distinct additive parts: returns to schooling and experience. The latter is modeled as in Browning, Ejrnaes, and Alvarez (2010): $8.83 + 0.56A_{i,t} - 0.057A_{i,t}^2$, with $A_{i,t}$ age in decades ($A_{i,t} = (18+t)/10$). The former includes time varying returns to education in order to capture individual specific responses to economy-wide shocks: $\delta_t E_i$, where E_i follows a lognormal distribution with parameters 2.3 and 0.2, which yields to 10.17 years of education on average. I model education as a continuous variable in order to avoid matching on discrete covariates. $\delta_t = \delta + r_t d$, with $\delta = 0.08$, $d = 0.02$ and r_t follows a uniform distribution on $[0, 1]$. For the RIP process, I impose $\beta_i = 0$, $\forall i$ and use MaCurdy (1982)'s estimates of the wage process (see appendix C for the detailed parameterization). Note that MaCurdy (1982) finds that $\mu_i = 0$, $\forall i$. For the HIP process, I use the parameters estimated by Guvenen (2007). I consider two different types of initial conditions: either the first shocks are drawn from the long run distribution ($\Sigma_0 = \Sigma_\infty$) or there is only one shock $v_{i,0}$ with variance σ, $U_{i,0} = v_{i,0}$ and $v_{i,-1} = 0$, $\forall i$. All the disturbances are normally distributed.

For the selection equation, I use equation (3). I add a linear term $\beta_x E_i$ to generate

selection on education. With the RIP (resp. HIP) model, I allow for both full information and limited information (resp. bayesian updating). Bayesian updating closely follows Guvenen (2007)'s Kalman filter approach and is described in appendix C. At each period, the agent uses her observed log-wages up to the previous period to update her prior about $(\mu_i, \beta_i, U_{i,t})$. This enables her in turn to forecast her forgone log wage. I use the same starting values for the prior distribution of $(\mu_i, \beta_i, U_{i,0})$ as in Guvenen (2007): the agent knows nothing of μ_i ($\mu_i^k = 0$, $\forall i$) and knows β_i^k. The variance of β_i^k is a fraction $(1 - \lambda)$ of the variance of β_i. I use Guvenen (2007)'s preferred estimates for λ: 0.6. I also vary the treatment date by making the program available at periods 5, 10, 20 or 30.

I use the Local Linear Regression (LLR) Matching on the propensity score estimator proposed by Heckman, Ichimura, and Todd (1998). I follow closely Smith and Todd (2005) for the implementation. I first estimate a linear probit and use predicted values as the propensity score. I exclude observations not on the common support. I also trim the data in order to avoid well known problems with LLR at low densities in small samples (Frölich, 2004). Because matching on past outcomes drastically reduces the variance of the propensity score, I have to use a large trimming level (0.4). I use a biweight kernel. There is no agreed upon method to select the optimal bandwith for matching. After experimenting with the data, I set the bandwidth at 0.15. I experiment with four different sets of control variables: E_i alone, or E_i supplemented with $\{Y_{i,k-j}\}_{j=1}^{\tau}$, with $\tau \in \{1, 2, 3\}$. DID-matching is always implemented symmetrically around treatment period k and outcomes are measured at period $k + 4$. For the Monte-Carlo simulations, I draw 500 samples with 1000 individuals each.

Results

I investigate the respective performances of matching and DID when conditions for their consistency are relaxed. I examine in turn the role of initial conditions,[12] in-

[12]Note that with an AR term very close to one, altering initial conditions is very similar to allowing for a a random walk.

formation and finally the importance of the income process (HIP *vs* RIP). I interpret the consequences of relaxing these assumptions thanks to the decomposition of the bias terms of both estimators in equations (21) and (22). I also report results on the sensitivity of these estimators to the number of lagged pre-treatment outcomes used as control variables.

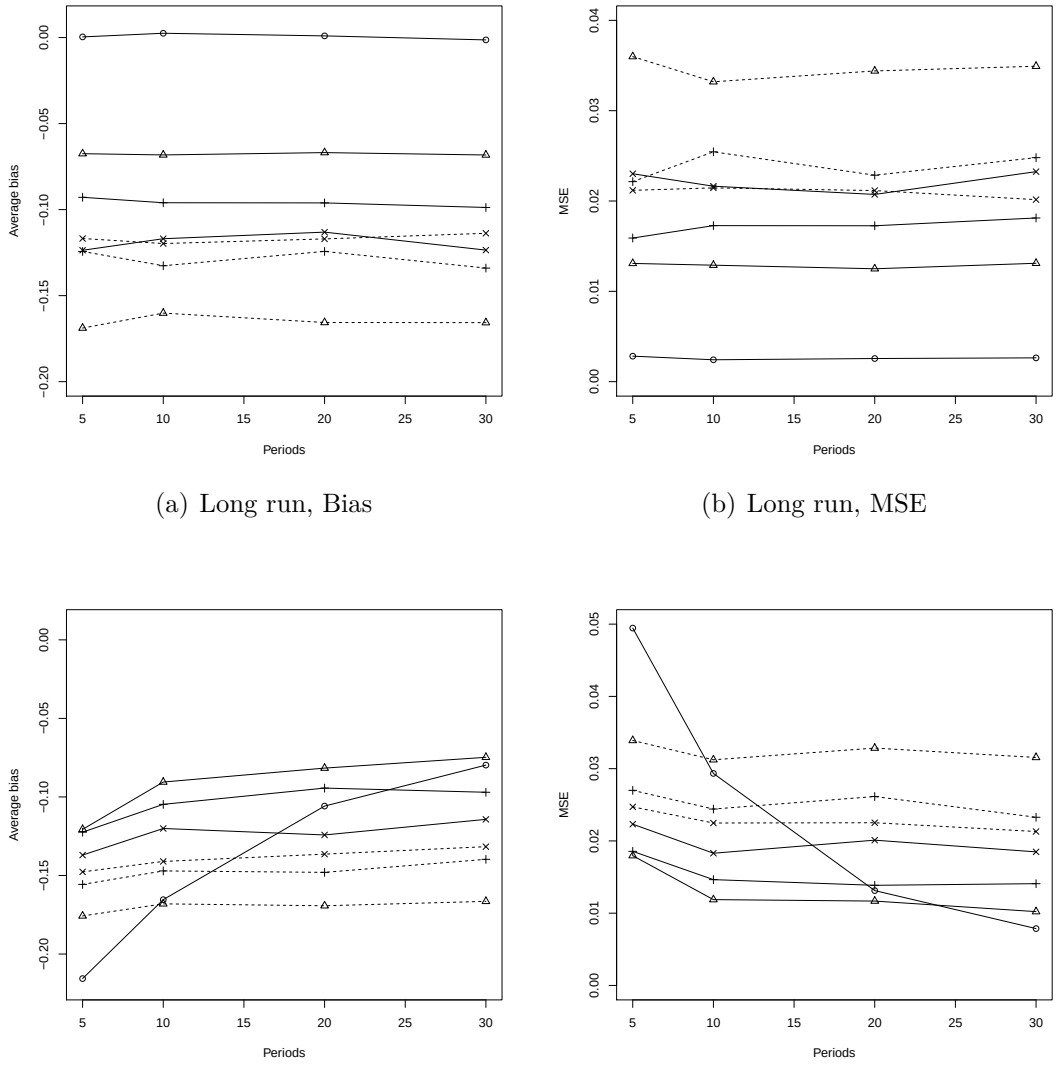
Initial conditions

The bias of matching should not be altered dramatically by changing the initial conditions of the ARMA process. Indeed, only the most recent shocks play a role in the bias term. This is not the case for DID. Moving from equilibrium long run initial conditions to off-equilibrium short run ones prevents term (22b) to cancel out: the variance of $U_{i,t}$ now varies with t. In the Monte-Carlo simulations, I start with an initial shock whose variance is lower than the long run variance $\sigma^2_{U_\infty}$. Because A^ι_k is always positive,[13] term (22b) is negative and DID underestimates the true treatment effect. This is because the pre-treatment average difference in outcomes between participants and non participants is lower in absolute value than the post-treatment difference.

The comparison of panels (a) and (c) in figures 3, 4, 5 and 6 confirms this analysis: matching is not sensitive to changes in initial conditions while DID estimates are always lower under short run initial conditions. DID is severely biased under short run initial conditions, but because the variance of $U_{i,t}$ increases with time, this bias decreases with experience. Under the RIP, at period 30, symmetric DID has the lowest absolute mean bias and MSE of all the estimators, as figures 3 and 4 show. Because it takes almost 100 periods for the $U_{i,t}$ process to reach the long run stable distribution, symmetric DID is nevertheless still biased after 30 periods.

[13]The average difference between $D^{\iota*}_{i,k}$ and its mean conditional on X_i is indeed positive for the treated and negative for the untreated, so that their difference is always positive.

Figure 3 – Mean bias and MSE of matching and symmetric DID-matching in the RIP under full information by type of initial condition



(a) Long run, Bias

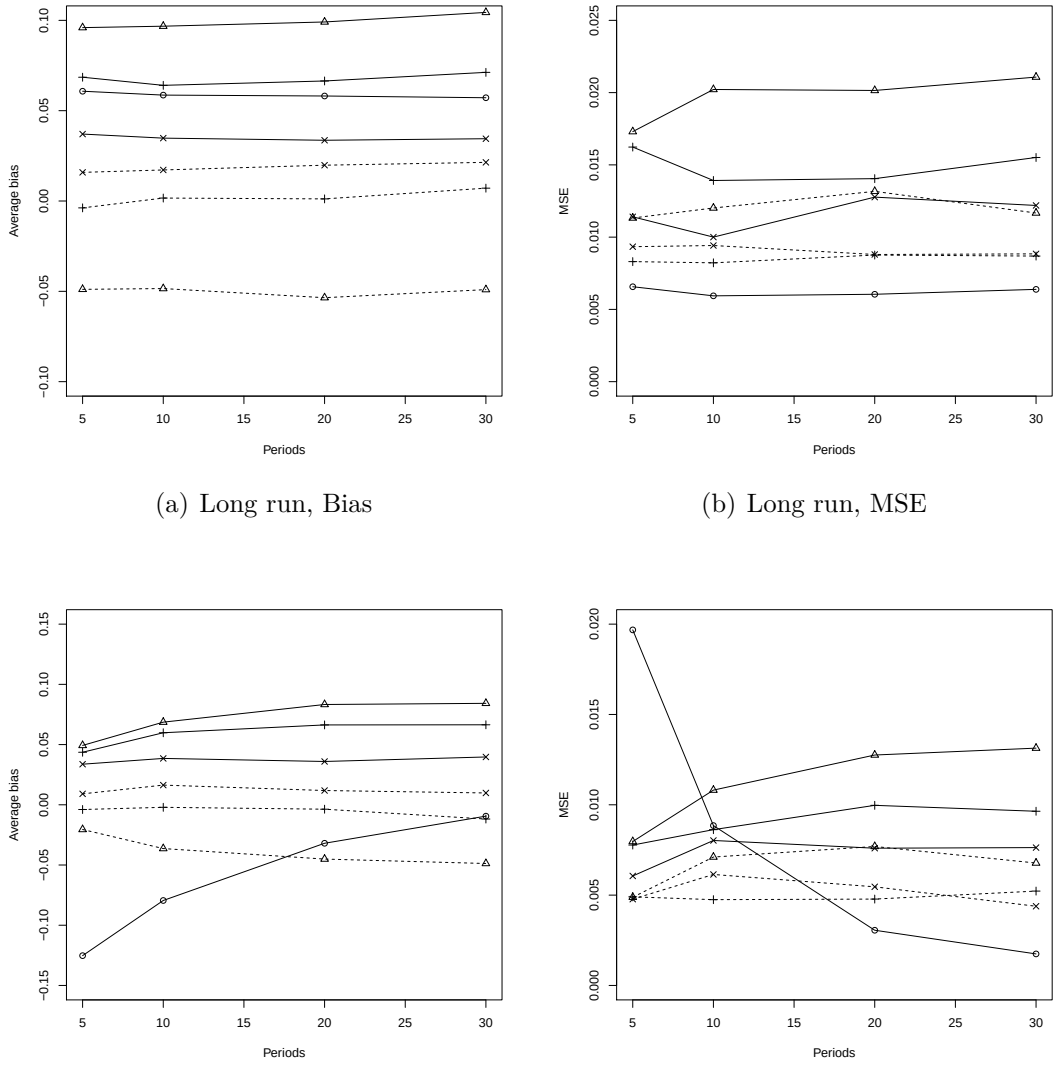
(b) Long run, MSE

(c) Short run, Bias

(d) Short run, MSE

Legend: ∘, △, + and × respectively stand for 0, 1, 2 and 3 periods of lagged outcomes as control variables. Solid lines are for symmetric DID and dotted lines for matching.
Note: mean bias and mean squared error (MSE) are calculated thanks to 500 Monte-Carlo replications. Each sample contains 1000 individuals with roughly 100 to 200 participants. The parameterization of the wage process uses MaCurdy (1982)'s estimates. "Long run" (resp. "short run") stands for the initial conditions of the ARMA process being drawn in the long run stable distribution (resp. in the distribution of the idiosyncratic shock $v_{i,t}$). The bias is estimated using local linear regression matching on the propensity score with a biweight kernel. The bandwidth is set to 0.15 and the trimming level is set to 0.4. The model and its parameterization are detailed in appendix C.

Figure 4 – Mean bias and MSE of matching and symmetric DID-matching in the RIP under limited information by type of initial condition



(a) Long run, Bias

(b) Long run, MSE

(c) Short run, Bias

(d) Short run, MSE

Legend: ∘, △, + and × respectively stand for 0, 1, 2 and 3 periods of lagged outcomes as control variables. Solid lines are for symmetric DID and dotted lines for matching.
Note: mean bias and mean squared error (MSE) are calculated thanks to 500 Monte-Carlo replications. Each sample contains 1000 individuals with roughly 100 to 200 participants. The parameterization of the wage process uses MaCurdy (1982)'s estimates. "Long run" (resp. "short run") stands for the initial conditions of the ARMA process being drawn in the long run stable distribution (resp. in the distribution of the idiosyncratic shock $v_{i,t}$). The bias is estimated using local linear regression matching on the propensity score with a biweight kernel. The bandwidth is set to 0.15 and the trimming level is set to 0.4. The model and its parameterization are detailed in appendix C.

Information

Moving from limited information (or bayesian updating) to full information adds the term (21e) to the bias of the matching estimator. This term is negative (the last shock to earnings is lower for participants). As a consequence, matching must underestimate the true treatment effect under full information. As for DID, moving from full information to limited information adds the bias term (22c). This term is positive and is due to the fact that when agents do not know the period k shock to their earnings, the covariance between future income and treatment utility is weaker than that between past income and treatment utility. As a consequence, DID must overestimate the true treatment effect under limited information.

This is what we observe when comparing figures 3 and 4 and figures 5 and 6: the matching and DID estimates always decrease when moving from limited to full information.

HIP *vs* RIP

Moving form the RIP to the HIP adds the random intercept and slope terms (21a) and (21b) to the bias of matching and the term (22a) to the bias of DID. We suspect that they are all negative, as individuals with high productivity level and/or growth will tend to select out of the treatment as their unobserved opportunity cost will be higher. This is true in the long run, after roughly 20 years, when random slopes start kicking in (see figures figures 5 and 6). The bias due to the random intercept (21a) should kick in immediately, at least under full information, but it does not seem to make a difference in the simulations during the first periods. It may be because the variance of μ_i is small. One obvious way to get around the bias due to the random trend in the HIP would be to implement a symmetric version of Heckman and Hotz (1989) triple difference estimator. Another solution would be to condition on averages of past log-wages and past variations in log-wages in a matching procedure.[14] I experiment with

[14] I thank an anonymous referee for suggesting this approach.

both approaches in Monte-Carlo results not shown here. The former approach works very well under bayesian updating, with a remaining downward bias oscillating between -0.01 and -0.04. The latter approach works extremely well under full information, but only at later periods (when $k \geq 20$).

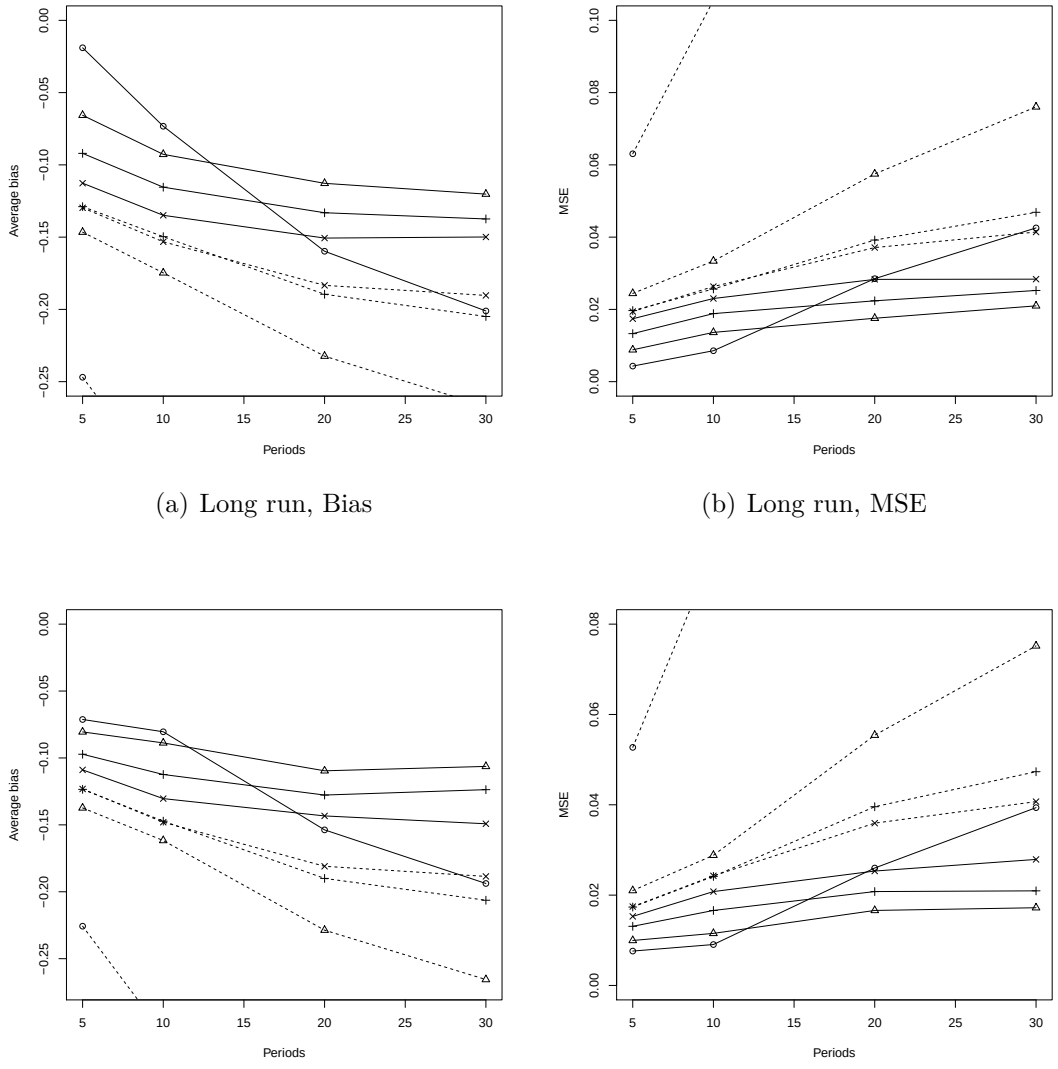
Increasing the number of pre-treatment outcomes used as control variables

Finally, adding more pre-treatment outcomes to the set of control variables almost always decreases the bias of matching estimators. First, under the RIP and limited information, the MA terms still generate a sizable bias when controlling for only one pre-treatment outcome: this bias is on average -0.05 with long run initial conditions (figure 4(a)), confirming the results of section 4. Adding $Y_{i,k-2}$ as a control variable almost completely cancels the bias. At the same time, adding also $Y_{i,k-3}$ does a little worse. The bias of matching always decreases when adding more pre-treatment outcomes as control variables in all the remaining experiments. The same is not true for DID. Under the RIP, adding more control variables may or may not decrease the bias of DID. Under the HIP, adding more control variables to the DID estimator generally decreases its mean bias.

DID *vs* matching

To sum up the results of this section, although matching is generally not consistent under credible parameterizations, it is less sensitive to deviations from the most favorable setting. It is not sensitive to initial conditions, generally underestimates the true treatment and thus provides a useful lower bound. Note that matching provides a lower bound in the HIP because I have assumed that the direct costs of entering the JTP (c_i) are independent of the random slope and intercept μ_i and β_i. Allowing for costs to be decreasing with both terms could reverse this effect. Finally, when more pre-treatment outcomes are added to the set of control variables, the bias of matching generally decreases in absolute value. Note that the RIP and with under limited information, controlling for additional pre-treatment outcomes almost completely cancels

Figure 5 – Mean bias and MSE of matching and symmetric DID-matching in the HIP under full information by type of initial condition



(a) Long run, Bias

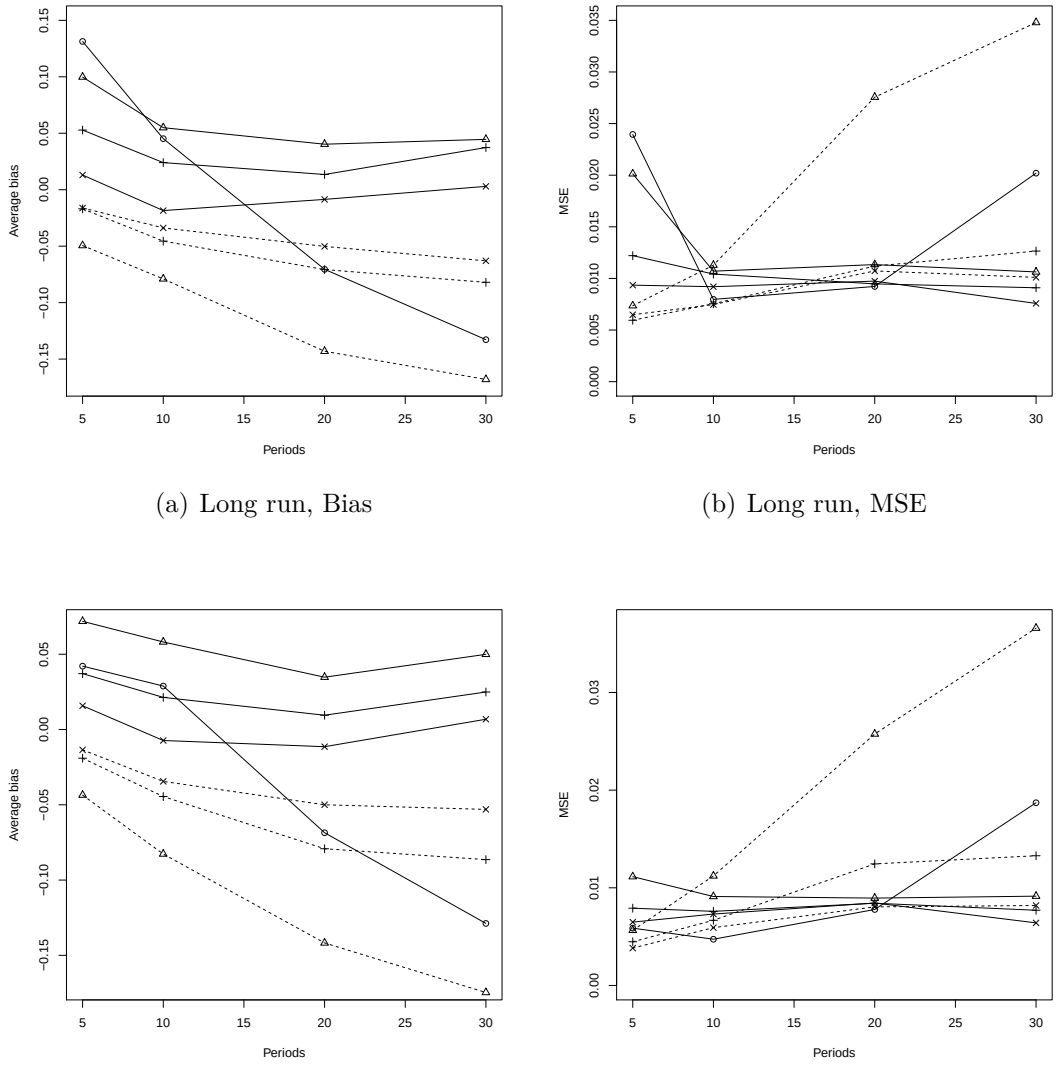
(b) Long run, MSE

(c) Short run, Bias

(d) Short run, MSE

Legend: ∘, △, + and × respectively stand for 0, 1, 2 and 3 periods of lagged outcomes as control variables. Solid lines are for symmetric DID and dotted lines for matching.
Note: mean bias and mean squared error (MSE) are calculated thanks to 500 Monte-Carlo replications. Each sample contains 1000 individuals with roughly 100 to 200 participants. The parameterization of the wage process uses Guvenen (2007)'s estimates. "Long run" (resp. "short run") stands for the initial conditions of the ARMA process being drawn in the long run stable distribution (resp. in the distribution of the idiosyncratic shock $v_{i,t}$). The bias is estimated using local linear regression matching on the propensity score with a biweight kernel. The bandwidth is set to 0.15 and the trimming level is set to 0.4. The model and its parameterization are detailed in appendix C.

Figure 6 – Mean bias and MSE of matching and symmetric DID-matching in the HIP under bayesian updating by type of initial condition



(a) Long run, Bias

(b) Long run, MSE

(c) Short run, Bias

(d) Short run, MSE

Legend: ∘, △, + and × respectively stand for 0, 1, 2 and 3 periods of lagged outcomes as control variables. Solid lines are for symmetric DID and dotted lines for matching.
Note: mean bias and mean squared error (MSE) are calculated thanks to 500 Monte-Carlo replications. Each sample contains 1000 individuals with roughly 100 to 200 participants. The parameterization of the wage process uses Guvenen (2007)'s estimates. "Long run" (resp. "short run") stands for the initial conditions of the ARMA process being drawn in the long run stable distribution (resp. in the distribution of the idiosyncratic shock $v_{i,t}$). The bias is estimated using local linear regression matching on the propensity score with a biweight kernel. The bandwidth is set to 0.15 and the trimming level is set to 0.4. The model and its parameterization are detailed in appendix C.

the remaining bias due to the MA terms.

DID is consistent under plausible conditions, but is much more sensitive to deviations from these conditions. Moreover, the bias generated by deviations from these conditions does not always go in the same direction: lower initial variance of the ARMA process yields to an underestimation of the treatment effect, as do the introduction of a random slope, whereas limited information yields to an overestimation of the true treatment effect. Results from the simulations do not suggest that the size of either the overestimation or the underestimation are innocuous. Finally, adding more control variables may increase bias in absolute value, but it can also decrease it.

Ideally, one would like to test for some of these conditions so as to discard some possible sources of bias. For example, if it was possible to choose either the HIP or the RIP and to know the distribution of the initial conditions, the only remaining source of uncertainty would be about agents' information set. Under the RIP and with long run initial conditions, DID overestimates the true treatment effect. The combination of matching and DID would then yield bounds on the true treatment effect. Under the HIP, controlling for more periods would decrease absolute bias, but it would not be possible to know whether DID overestimates (limited information) or underestimates (full information) the true effect.

Unfortunately, it is very hard to differentiate the HIP from the RIP (Guvenen, 2009). Indeed, rejecting the common trend assumption on pre-treatment data, as suggested by Heckman and Hotz (1989), can be a sign of the HIP or of an Ashenfelter's dip. As Guvenen (2009) shows, autocovariances of wage innovations are equal to the sum of the variance of the random trend and the effect of persistent shocks (when $\rho < 1$). In the long run, this second term eventually vanishes. But tests using this insight lack power with finite autocovariances (Guvenen, 2009). Hryshko (2012) argues that using all the autocovariances at once identifies the variance of the random trend (and that it is zero). This is an important area for further research.

6 Discussion

Only results from applied work comparing both matching and DID to an experimental benchmark can provide evidence on whether the consistency of DID is overturned in practice by its sensitivity to deviations from the conditions ensuring its consistency. Indeed, previously unnoticed results in Heckman, Ichimura, Smith, and Todd (1998) and Smith and Todd (2005) provide evidence that DID is stable and most often the least biased estimator in empirical applications.[15] Heckman, Ichimura, Smith, and Todd (1998, p.1062) compare the relative ability of different sets of control variables to reproduce the experimental results of the evaluation of the Job Training Partnership Act (JTPA) thanks to matching and DID matching. When using a crude control set not including wages at the date of enrollment, the average bias of the symmetric DID estimator is of 73 % of the treatment effect, lower than that obtained with matching on past wages at enrollment and labor market transitions (382 % of the treatment effect). Note however that the inclusion of labor market transitions, a topic not discussed in this paper, improves the performance of matching noticeably (58% to 88% of the treatment effect). Note that matching still does not outperform symmetric DID in that case. Smith and Todd (2005) use two sets of control variables when estimating the bias of propensity score matching and DID propensity score matching in the National Support for Work experimental study: the first set (they name it the Lalonde set) does not contain past income while the second set (the Dehejia and Wahba (DW) set) does contain past income. When they apply DID matching with the first set of controls, the bias is of respectively -2 %, 22 % and -16 % of the treatment effect when using the most efficient matching estimators (respectively nearest neighbour matching with one neighbour restricted to the common support, local linear matching with a small

[15]Note nevertheless that these papers apply matching and DID matching to earnings in levels, not in logarithms, as I study in this paper. This does not change the symmetry property of Ashenfelter's dip, because taking the exponential of both the selection index and earnings preserves the ellipticity of their joint distribution. This is confirmed in Monte-Carlo simulations not presented here. The main problem comes from the assumption on parallel trend: it is not fulfilled anymore because trends are now exponential. This is mitigated in the studies under investigation as the time scope is rather small.

bandwith (1.0) and local linear regression adjusted matching with the same bandwith, see their table 6 p.340). When using matching controlling for past outcomes, the bias is larger: respectively -95 %, -156 % and -159 % with the same estimators (see their table 5, p.336). In both of these works, matching generally underestimates the true treatment effect and adding past outcomes in the set of control variables when using matching decreases selection bias, as results in this paper predict.

Overall, the results presented in this paper tend to mitigate Imbens and Wooldridge (2009)'s statement that matching is more credible than DID. In the context of JTP, credible conditions for the consistency of symmetric DID matching can be stated whereas the conditions I have been able to state in order to ensure the consistency of matching contradict important properties of empirical earnings processes. Matching is nevertheless more stable: it is less sensitive to deviations from the conditions ensuring its consistency, generally yields a lower bound and its bias decreases with the number of pre-treatment outcomes included in the set of control variables.

In view of the results in this paper, what is an empirical researcher to do when evaluating a JTP with panel data? First, it seems recommended to apply DID symmetrically around the treatment date. Second, comparing the results of symmetric DID with those of matching would give a sense of how sensitive the results are to the sources of bias delineated in this paper.

Devising procedures for testing whether the conditions ensuring the consistency of symmetric DID are met is an interesting avenue for further research. This would involve telling apart HIP from RIP (a very difficult undertaking from income data alone (Guvenen, 2009; Hryshko, 2012)); inferring agents' information set when entering the treatment, using insights from Cunha, Heckman, and Navarro-Lozano (2005); and testing for initial conditions.

References

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

Arnold, B., R. Beaver, R. Groeneveld, and W. Meeker (1993): "The Non-truncated Marginal of a Truncated Bivariate Normal Distribution," *Psychometrika*, 58(3), 471–488.

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

Balineau, G. (2012): "Disentangling the Effects of Fair Trade on the Quality of Malian Cotton," FERDI Working Paper 39.

Bhattacharya, J., and W. B. Vogt (2007): "Do Instrumental Variables Belong in Propensity Scores?," NBER Working Paper 343.

Browning, M., M. Ejrnaes, and J. Alvarez (2010): "Modelling Income Processes with Lots of Heterogeneity," *Review of Economic Studies*, 77(4), 1353–1381.

Chabé-Ferret, S., and J. Subervie (Forthcoming): "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*.

Chu, K.-C. (1973): "Estimation and Decision for Linear Systems with Elliptical Random Processes ," *IEEE Transactions on Automatic Control*, 18(5), 199–505.

Cunha, F., J. J. Heckman, and S. Navarro-Lozano (2005): "Separating Uncertainty from Heterogeneity in Life Cycle Earnings, the 2004 Hicks Lecture," *Oxford Economic Papers*, 57(2), 191–261.

Dawid, A. P. (1979): "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1), 1–31.

Frölich, M. (2004): "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," *Review of Economics and Statistics*, 86(1), 77–90.

Gobillon, L., T. Magnac, and H. Selod (2010): "Do Unemployed Workers Benefit from Enterprise Zones? The French Experience," IDEI Working Paper 645.

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.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315–331.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1996): "Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method," *Proceedings of the National Academy of Sciences*, 93(23), 13416–13420.

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

Heckman, J. J., and V. J. Hotz (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: the Case of Manpower Training," *Journal of the American Statistical Association*, 84(408), 862–874.

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

Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *The Review of Economic Studies*, 65(2), 261–294.

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.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

Hryshko, D. (2012): "Labor Income Profiles are Not Heterogeneous: Evidence from Income Growth Rates," *Quantitative Economics*, 3(2), 177–209.

Imbens, G. W., and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86.

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.

Mayer, T., F. Mayneris, and L. Py (2012): "The Impact of Urban Enterprise Zones on Establishments' Location Decisions: Evidence from French ZFUs," CEPR Discussion Paper 9074.

Meghir, C., and L. Pistaferri (2011): "Earnings, Consumption and Life Cycle

Choices," in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, vol. 4, Part B, chap. 9, pp. 773 – 854. Elsevier.

Myers, J. A., J. A. Rassen, J. J. Gagne, K. F. Huybrechts, S. Schneeweiss, K. J. Rothman, M. M. Joffe, and R. J. Glynn (2011): "Effects of Adjusting for Instrumental Variables on Bias and Precision of Effect Estimates," *American Journal of Epidemiology*, 174(11), 1213–1222.

Pearl, J. (2010): "On a Class of Bias-Amplifying Variables that Endanger Effect Estimates," in *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, ed. by P. Grunwald, and P. Sprites, pp. 417–424. AUAI Press, Corvallis, Oregon.

——— (2011): "Invited Commentary: Understanding Bias Amplification," *American Journal of Epidemiology*, 174(11), 1223–1227.

Puhani, P. A., and K. Sonderhof (2010): "The Effects of a Sick Pay Reform on Absence and on Health-related Outcomes," *Journal of Health Economics*, 29(2), 285 – 302.

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.

Wooldridge, J. M. (2005): "Violating Ignorability of Treatment by Controlling for Too Many Factors," *Econometric Theory*, 21(05), 1026–1028.

——— (2009): "Should Instrumental Variables be Used as Matching Variables?," Unpublished.

Ziebarth, N. R., and M. Karlsson (2010): "A Natural Experiment on Sick Pay Cuts, Sickness Absence, and Labor Costs," *Journal of Public Economics*, 94(11-12), 1108 – 1122.

A Proof of the propositions in section 3

Proof of proposition 1

From equation (18), $\mathbb{CD}_\iota(Y^0_{i,k+\tau}|X_i, Y_{i,k-\tau'}) = 0$ is a sufficient condition for $B^M_{\tau,\tau',\iota}$ to be null. We have:

$$\mathbb{CD}_\iota(Y^0_{i,k+\tau}|X_i, Y^0_{i,k}) = \mathbb{CD}_\iota(\mu_i + \beta_i(k+\tau) + U_{i,k+\tau}|X_i, Y^0_{i,k}) \tag{23}$$

Under full information, this will be null if:

$$\begin{aligned}&\mathbb{E}\left[\mu_i + \beta_i(k+\tau) + U_{i,k+\tau}|X_i = x, Y^0_{i,k} = y, \mathbb{1}[\frac{\alpha_i}{r} - c_i - y \geq 0]\right] \\ &= \mathbb{E}[\mu_i + \beta_i(k+\tau) + U_{i,k+\tau}|X_i = x, Y^0_{i,k} = y, \mathbb{1}[\frac{\alpha_i}{r} - c_i - y < 0]]\end{aligned} \tag{24}$$

In our model, we have assumed that $(U_{i,0}, \{v_{i,j}\}_{j=-1}^k) \perp\!\!\!\perp (\alpha_i, \beta_i, c_i, \mu_i, X_i)$, so that $(U_{i,0}, \{v_{i,j}\}_{j=-1}^k) \perp\!\!\!\perp (\alpha_i, c_i)|(\mu_i, \beta_i, X_i)$. If we assume that $(\mu_i, \beta_i) \perp\!\!\!\perp (c_i, \alpha_i)|X_i$, then we have that $(\mu_i, \beta_i, U_{i,0}, \{v_{i,j}\}_{j=-1}^k) \perp\!\!\!\perp (\alpha_i, c_i)|X_i$, using lemma 4.3 in Dawid (1979). Because $\mu_i + \beta_i(k+\tau) + U_{i,k+\tau}$ and $Y^0_{i,k}$ are a function of $(\mu_i, \beta_i, U_{i,0}, \{v_{i,j}\}_{j=-1}^k)$ conditional on X_i, we have $(\mu_i + \beta_i(k+\tau) + U_{i,k+\tau}) \perp\!\!\!\perp (\alpha_i, c_i)|(X_i, Y^0_{i,k})$, using lemma 4.2 in Dawid (1979), which proves equation (24) and completes the proof.

Proof of proposition 2

Because U_{it} follows an ARMA(1,2), we have, for $\tau \geq 2$ (the proof for $\tau = 1$ is similar and thus omitted):

$$\begin{aligned}U_{i,k+\tau} = \rho^{\tau+1}U_{i,k-1} + \rho^\tau m_2 v_{i,k-2} + \rho^{\tau-1}(\rho m_1 + m_2)v_{i,k-1} \\ + (\rho^2 + \rho m_1 + m_2)\sum_{j=0}^{\tau-2}\rho^{\tau-2-j}v_{i,k+j} + (\rho + m_1)v_{i,k+\tau-1} + v_{i,k+\tau}\end{aligned} \tag{25}$$

Substituting for $U_{i,k-1}$ and acknowledging that all shocks posterior to period k are orthogonal to the conditioning set, we have, for $\tau \geq 2$:

$$\mathbb{CD}_\iota(Y^0_{i,k+\tau}|X_i, Y_{i,k-1}) = (1-\rho^{\tau+1})\mathbb{CD}_\iota(\mu_i|X_i, Y_{i,k-1}) \tag{26a}$$

$$+ (k+\tau-\rho^{\tau+1}(k-1))\mathbb{CD}_\iota(\beta_i|X_i, Y_{i,k-1}) \tag{26b}$$

$$+ \rho^{\tau-1}(\rho m_1 + m_2)\mathbb{CD}_\iota(v_{i,k-1}|X_i, Y_{i,k-1}) \tag{26c}$$

$$+ \rho^{\tau} m_2 \mathbb{CD}_\iota(v_{i,k-2}|X_i, Y_{i,k-1}) \tag{26d}$$

$$+ \rho^{\tau-2}(\rho^2 + \rho m_1 + m_2)\mathbb{CD}_\iota(v_{i,k}|X_i, Y_{i,k-1}). \tag{26e}$$

Note that parts (26a) and (26b) are equal to zero when $F_{\mu,\beta}$ is degenerate and that parts (26c) and (26d) are zero when $m_1 = m_2 = 0$. Finally, let's write the expected foregone wage in terms of the conditioning variable under limited information:

$$\mathbb{E}[Y^0_{i,k}|\mathcal{I}^l_{i,k}] = g(X_i, \delta_k) + \mu_i + \rho U_{i,k-1} + m_1 v_{i,k-1} + m_2 v_{i,k-2} \tag{27}$$

$$= g(X_i, \delta_k) - \rho g(X_i, \delta_{k-1}) + \mu_i(1-\rho) + Y^0_{i,k-1} + m_1 v_{i,k-1} + m_2 v_{i,k-2} \tag{28}$$

This result comes from $v_{i,k}$ being mean-zero and not contained in the limited information set and by substituting for $U_{i,k-1}$. We see that the conditioning set in (26e) is not correlated with v_{ik}, so that this bias term is also zero. Using the law of iterated expectation proves the result.

Proof of proposition 3

$B^{DID}_{\tau,\tau',\iota} = 0$ if $\mathbb{CD}_\iota(\Delta^{Y^0_i}_{\tau,\tau'}|X_i) = 0$. If F_β is degenerate, it is easy to show that this is equivalent to $\mathbb{CD}_\iota(\Delta^{U_i}_{\tau,\tau'}|X_i) = 0$. Under coarse information and F_β degenerate, we have: $D^{c*}_{i,k} = \frac{\alpha_i}{r} - c_i - g(X_i, \delta_k) - \mu_i$, so that:

$$\mathbb{E}\left[\Delta^{U_i}_{\tau,\tau'}|X_i, \mathbb{1}[D^{c*}_{i,k} \geq 0]\right] = \mathbb{E}[\Delta^{U_i}_{\tau,\tau'}|X_i, \mathbb{1}[D^{c*}_{i,k} < 0]], \tag{29}$$

because $\{v_{i,j}\}_{j=0}^{k+\tau} \perp\!\!\!\perp (\alpha_i, c_i, \mu_i, X_i)$, by assumption. This completes the proof of the consistency of DID matching. Matching is inconsistent even though (26e) is equal to zero. The first bias term (26a) is non null: since the agent knows μ_i and self-selects on it, $\mathbb{CD}_c(\mu_i|X_i, Y_{i,k-1}) \neq 0$. Moreover, (26c) and (26d) are also non null under these conditions. This is because conditioning on $Y_{i,k-1}$ makes them correlated with $D_{i,k}^c$ through μ_i. These terms do not cancel out in general, which proves the result.

Proof of proposition 4

We have:

$$B_{\tau,\tau',\iota}^{DID} = \mathbb{E}[\mathbb{CD}_\iota(\Delta_{\tau,\tau'}^{Y_i^0}|X_i)|D_{i,k}^\iota = 1] \tag{30}$$

$$= \mathbb{E}[\mathbb{CD}_\iota(\beta_i(\tau+\tau') + \Delta_{\tau,\tau'}^{U_i}|X_i)|D_{i,k}^\iota = 1]. \tag{31}$$

Because $\mathbb{E}[U_{i,t}|D_{i,k}^{\iota*}, X_i]$ is linear, we can write:

$$\mathbb{E}[\Delta_{\tau,\tau'}^{U_i}|X_i, D_{i,k}^\iota = 1] = \mathbb{E}[\mathbb{E}[\Delta_{\tau,\tau'}^{U_i}|X_i, D_{i,k}^{*\iota}]|X_i, D_{i,k}^\iota = 1] \tag{32}$$

$$= \mathbb{E}[\mathbb{E}[\Delta_{\tau,\tau'}^{U_i}|X_i] + \frac{\mathrm{Cov}(\Delta_{\tau,\tau'}^{U_i}, D_{i,k}^{*\iota}|X_i)}{\mathrm{Var}(D_{i,k}^{*\iota}|X_i)}(D_{i,k}^{*\iota} - \mathbb{E}[D_{i,k}^{*\iota}|X_i])|X_i, D_{i,k}^\iota = 1] \tag{33}$$

When $\iota \in \{f, l\}$, we have:

$$\mathrm{Cov}(\Delta_{\tau,\tau'}^{U_i}, D_{i,k}^{*\iota}|X_i) = \mathrm{Cov}(\Delta_{\tau,\tau'}^{U_i}, \frac{\alpha_i}{r} - c_i - Y_{i,k}^0 + \mathbb{1}[\iota = l]v_{i,k}|X_i) \tag{34}$$

$$= -\mathrm{Cov}(\Delta_{\tau,\tau'}^{U_i}, U_{i,k}) + \mathbb{1}[\iota = l]\mathrm{Cov}(\Delta_{\tau,\tau'}^{U_i}, v_{i,k}). \tag{35}$$

The second equality follows from $\{v_{i,j}\}_{j=-1}^{k+\tau} \perp\!\!\!\perp (\alpha_i, c_i, \mu_i, X_i)$, which also implies that $\mathbb{E}[\Delta_{\tau,\tau'}^{U_i}|X_i] = 0$. Now, we have:

$$\mathrm{Cov}(\Delta_{\tau,\tau'}^{U_i}, D_{i,k}^{*\iota}|X_i) = \mathrm{Cov}(U_{i,k-\tau'}, U_{i,k}) - \mathrm{Cov}(U_{i,k+\tau}, U_{i,k})$$
$$+ \mathbb{1}[\iota = l](\rho^{\tau-2})(\rho^2 + m_1\rho + m_2)\mathrm{Var}(v_{i,k}). \tag{36}$$

This follows from the ARMA(1,2) process: only $U_{i,k+\tau}$ is correlated to $v_{i,k}$.

Note first that, for $\tau \geq 2$:

$$\text{Cov}(U_{i,k-\tau}, U_{i,k}) = \text{Cov}(U_{i,k-\tau}, \rho^{\tau-2}(\rho^2 U_{i,k-\tau} + \rho m_2 v_{i,k-\tau-1} + (\rho m_1 + m_2)v_{i,k-\tau})) \tag{37}$$

$$= \rho^{\tau}\text{Var}(U_{i,k-\tau}) + \rho^{\tau-1}m_2(\rho + m_1)\text{Var}(v_{i,k-\tau-1}) + \rho^{\tau-2}(\rho m_1 + m_2)\text{Var}(v_{i,k-\tau}). \tag{38}$$

Because $v_{i,t}$ is an i.i.d. process, $\text{Var}(v_{i,t}) = \sigma^2$, $\forall t$. Using the fact that $\{U_{i,t}\}_{t=0}^{\infty} \perp\!\!\!\perp X_i$, we thus have, for $\tau \geq 2$:

$$B^{DID}_{\tau,\tau',\iota} = (\tau + \tau')\mathbb{E}[\mathbb{CD}_{\iota}(\beta_i|X_i)|D^{\iota}_{i,k} = 1] \tag{39a}$$

$$+ \left(\rho^{\tau'}\text{Var}(U_{i,k-\tau'}) - \rho^{\tau}\text{Var}(U_{i,k})\right) A^{\iota}_k \tag{39b}$$

$$+ \mathbb{1}[\iota = l](\rho^{\tau-2})(\rho^2 + m_1\rho + m_2)\sigma^2 A^{\iota}_k, \tag{39c}$$

with:

$$A^{\iota}_k = \mathbb{E}[\mathbb{CD}_{\iota}(\frac{D^{*\iota}_{i,k} - \mathbb{E}[D^{*\iota}_{i,k}|X_i]}{\text{Var}(D^{*\iota}_{i,k}|X_i)}|X_i)|D^{\iota}_{i,k} = 1] \tag{40}$$

We can write, for $k > 2$:

$$U_{i,k} = \rho^k U_{i,0} + \rho^{k-1}m_2 v_{i,-1} + \rho^{k-2}(\rho m_1 + m_2)v_{i,0} + (\rho^2 + \rho m_1 + m_2)\sum_{j=0}^{k-3}\rho^j v_{i,k-j-2} + (\rho + m_1)v_{i,k-1} + v_{i,k}. \tag{41}$$

As a consequence:

$$\text{Var}(U_{i,k}) = \left(1 + (\rho + m_1)^2 + (\rho^2 + \rho m_1 + m_2)^2\frac{1 - \rho^{2(k-2)}}{1 - \rho^2}\right)\sigma^2 + \rho^{2k}\text{Var}(U_{i,0}) + \rho^{2(k-1)}m_2\text{Var}(v_{i,-1}) + \rho^{2(k-2)}(\rho m_1 + m_2)\text{Var}(v_{i,0}) + 2\rho^{2k-1}m_2\text{Cov}(U_{i,0}, v_{i,-1}) + 2\rho^{2k-2}(\rho m_1 + m_2)\text{Cov}(U_{i,0}, v_{i,0}). \tag{42}$$

First, note that when $|\rho| < 1$, we have:

$$\sigma^2_{U_\infty} = \lim_{k\to\infty} \text{Var}(U_{i,k}) = \left(1 + (\rho + m_1)^2 + (\rho^2 + \rho m_1 + m_2)^2 \frac{1}{1-\rho^2}\right)\sigma^2, \tag{43}$$

If we replace Σ_0 with Σ_∞ in equation (42), we also have that $\text{Var}(U_{i,k}) = \sigma^2_{U_\infty}$, $\forall k$. Indeed we have:

$$\begin{aligned}
&\text{Var}(\rho^k U_{i,0} + \rho^{k-1} m_2 v_{i,-1} + \rho^{k-2}(\rho m_1 + m_2) v_{i,0}) \\
&\quad = \rho^{2(k-2)}\Bigg(\rho^4 \text{Var} U_{i,0} + \rho^2 m_2^2 \text{Var}(v_{i,-1}) + (\rho m_1 + m_2)^2 \text{Var}(v_{i,0}) \\
&\qquad + 2\rho^3 m_2 \text{Cov}(U_{i,0}, v_{i,-1}) + 2\rho^2(\rho m_1 + m_2)\text{Cov}(U_{i,0}, v_{i,0})\Bigg) \qquad (44) \\
&\quad = \rho^{2(k-2)}\sigma^2\Bigg(\frac{\rho^4}{1-\rho^2}(\rho^2 + \rho m_1 + m_2) + \rho^4 + \rho^4(m_1 + \rho)^2 + \rho^2 m_2^2 \\
&\qquad + (\rho m_1 + m_2)^2 + 2\rho^3 m_2(m_1 + \rho) + 2\rho^2(\rho m_1 + m_2)\Bigg) \qquad (45) \\
&\quad = \rho^{2(k-2)}\sigma^2(\rho^2 + \rho m_1 + m_2)\left(\frac{\rho^4}{1-\rho^2} + 1 + \rho^2\right) \qquad (46) \\
&\quad = \rho^{2(k-2)}\sigma^2(\rho^2 + \rho m_1 + m_2)\left(\frac{1}{1-\rho^2}\right). \qquad (47)
\end{aligned}$$

Replacing the two last lines of equation (42) by the right hand side of equation (47) yields the result.

We thus have $\text{Var}(U_{i,k}) = \sigma^2_{U_\infty}$, $|\rho| < 1$ and $k \to \infty$ (or $\forall k$ when $\Sigma_0 = \Sigma_\infty$). Using equation (39), we can see that this, together with F_β degenerate, implies that $B^{DID}_{\tau,\tau,f} = 0$.

In order to prove that $B^M_{\tau,\tau',f} \neq 0$, it is enough to find a sub-model that follows the restriction of the proposition and in which matching is biased. The model in section 4 fulfills these conditions.

B Derivation of bias terms in the labor example with normal MA terms

Not controlling for past outcomes

In this section, I derive closed form expressions for the asymptotic bias terms of section 3 under the assumption that the i.i.d MA terms are normal with variance σ^2. I also assume that the process generating the outcomes began sufficiently far in the past so that I can abstract from the dependence on t by considering that the MA terms are a sum of an infinite number of shocks. I moreover posit that α_i, c_i, μ_i is normally distributed with variance $\sigma^2_\alpha, \sigma^2_c, \sigma^2_\mu$ and corresponding covariances. To obtain the asymptotic bias terms, I study the joint distribution of normal variables conditional on $X_i = x$. Keeping the conditioning on $X_i = x$ implicit, the bias term of DID is equal to (see equation 20):

$$B^{DID}_{\tau,\tau',\iota} = \mathbb{CD}_\iota(\Delta^{Y^0_i}_{\tau,\tau'}) \tag{48}$$

$$= \frac{\text{Cov}(Y^0_{i,k+\tau}, D^{*\iota}_{i,k}) - \text{Cov}(Y^0_{i,k-\tau'}, D^{*\iota}_{i,k})}{\sigma^2_{D^{*\iota}}}(\mathbb{CD}_\iota(D^{*\iota}_{i,k})) \tag{49}$$

After some calculations, we can show that, $\forall \tau \in \mathbb{Z}$:

$$\begin{aligned}\text{Cov}(Y^0_{i,k+\tau}, D^{*\iota}_{ik}) = \frac{\sigma_{\mu,\alpha}}{r} - (1-\rho^{|\tau|})\sigma^2_\mu - \sigma_{\mu,c} - \rho^{|\tau|}\sigma^2_Y \\ - \rho^{|\tau|-2}\sigma^2\Bigg(\mathbb{1}[\tau \neq 0](\rho m_2(m_1+\rho) + \mathbb{1}[\iota = f \text{ or } \tau < 0]\rho m_1) \\ + \mathbb{1}[\tau \neq 1 \text{ and } \tau \neq 0]\mathbb{1}[\iota = f \text{ or } \tau < 0]m_2 - \mathbb{1}[\iota = l \text{ and } \tau \geq 0]\rho^2\Bigg),\end{aligned} \tag{50}$$

and:

$$\mathbb{CD}_\iota(D^{*\iota}_{i,k}) = \frac{1}{\sigma_{D^{*\iota}}}\left(\frac{\phi(A_x)}{1-\Phi(A_x)} + \frac{\phi(A_x)}{\Phi(A_x)}\right), \tag{51}$$

with:

$$\sigma_Y^2 = \sigma_{U_\infty}^2 + \sigma_\mu^2 \tag{52}$$

$$\sigma_{U_\infty}^2 = \sigma^2 \left(1 + (m_1 + \rho)^2 + \frac{\left(\rho^2 + \rho m_1 + m_2\right)^2}{1 - \rho^2}\right) \tag{53}$$

$$\sigma_{D^{*\iota}}^2 = \sigma_U^2 - \mathbb{1}[\iota = l]\sigma^2 + \sigma_\mu^2 + \sigma_c^2 + \frac{\sigma_\alpha^2}{r^2} - 2(\frac{\sigma_{c,\alpha}}{r} + \frac{\sigma_{\mu,\alpha}}{r} - \sigma_{\mu,c}) \tag{54}$$

$$A_x = \frac{g(x, \delta_k) - \frac{\bar{\alpha}}{r} + \bar{c} + \bar{\mu}}{\sigma_{D^{*\iota}}}. \tag{55}$$

Controlling for past outcomes

To derive the bias term of matching on past outcomes, I use the fact that it can be rewritten in the following way:

$$B^m(\tau, \tau', \iota, y) = \mathbb{E}[Y_{i,k+\tau}^0 | D^\iota = 1] - \mathbb{E}[\mathbb{E}[Y_{i,k+\tau}^0 | D_{ik}^\iota = 0, Y_{i,k-\tau'}^0] | D^\iota = 1]. \tag{56}$$

The average outcome for the treated can be obtained by results in the previous section. The main difficulty is to form the second part of the term on the right hand side of equation (56): the mean outcome of the matched non-participants. To form this quantity, first note that, because these variables are jointly normally distributed, their conditional expectation is linear, so that, , $\forall(\tau, \tau') \in \mathbb{Z}^2$:

$$\begin{aligned}
&\mathbb{E}\left[Y_{i,k+\tau}^0 | D_{ik}^{*\iota}, Y_{i,k-\tau'}^0\right] \\
&\quad = \mathbb{E}[Y_{i,k+\tau}^0] + \beta_{\tau, D^{*\iota}} \left(D_{ik}^{*\iota} - \mathbb{E}[D_{ik}^{*\iota}]\right) + \beta_{\tau,\tau'} \left(Y_{i,k-\tau'}^0 - \mathbb{E}[Y_{i,k-\tau'}^0]\right),
\end{aligned} \tag{57}$$

with:

$$\beta_{\tau,D^{*\iota}} = \frac{\text{Cov}(Y^0_{i,k+\tau}, D^{*\iota}_{ik})\sigma^2_Y - \text{Cov}(Y^0_{i,k-\tau'}, D^{*\iota}_{ik})\sigma_{Y^0_{k+\tau},Y^0_{k-\tau'}}}{\sigma^2_{D^{*\iota}}\sigma^2_Y - \text{Cov}(Y^0_{i,k-\tau'}, D^{*\iota}_{ik})^2}, \tag{58}$$

$$\beta_{\tau,\tau'} = \frac{\sigma_{Y^0_{k+\tau},Y^0_{k-\tau'}}\sigma^2_{D^{*\iota}} - \text{Cov}(Y^0_{i,k+\tau}, D^{*\iota}_{ik})\text{Cov}(Y^0_{i,k-\tau'}, D^{*\iota}_{ik})}{\sigma^2_{D^{*\iota}}\sigma^2_Y - \text{Cov}(Y^0_{i,k-\tau'}, D^{*\iota}_{ik})^2}, \tag{59}$$

$$\sigma_{Y_{k+\tau},Y_{k-\tau'}} = \sigma_{U_{k+\tau},U_{k-\tau'}} + (1-\rho^{|\tau+\tau'|})\sigma^2_\mu, \tag{60}$$

$$\begin{aligned}\sigma_{U_{k+\tau},U_{k-\tau'}} = \rho^{|\tau+\tau'|}\sigma^2_U + \rho^{|\tau+\tau'|-2}\sigma^2\Bigg(&\mathbb{1}[|\tau+\tau'|>0]\rho(m_2(m_1+\rho)+m_1)\\ &+\mathbb{1}[|\tau+\tau'|>1]m_2\Bigg).\end{aligned} \tag{61}$$

From this, we again use the law of iterated expectation to derive the conditional expectation of non-participants' outcomes:

$$\begin{aligned}\mathbb{E}[Y^0_{i,k+\tau}|D^\iota_{ik}=0, Y^0_{i,k-\tau'}] &= \mathbb{E}[\mathbb{E}[Y^0_{i,k+\tau}|D^{*\iota}_{ik}, Y^0_{i,k-\tau'}]|D^{*\iota}_{ik}<0, Y^0_{i,k-\tau'}] &(62)\\ &= \mathbb{E}[Y^0_{i,k+\tau}] + \gamma_{\tau,\tau'}\left(Y^0_{i,k-\tau'} - \mathbb{E}[Y^0_{i,k-\tau'}]\right)\\ &\quad + \gamma_{\tau,D^{*\iota}}\frac{\phi(A_{xy})}{\Phi(A_{xy})}, &(63)\end{aligned}$$

with:

$$\gamma_{\tau,\tau'} = \beta_{\tau,D^{*\iota}}\frac{\text{Cov}(Y^0_{i,k-\tau'}, D^{*\iota}_{ik})}{\sigma^2_Y} + \beta_{\tau,\tau'}, \tag{64}$$

$$\gamma_{\tau,D^{*\iota}} = \beta_{\tau,D^{*\iota}}\sqrt{\sigma^2_{D^{*\iota}} - \frac{\text{Cov}(Y^0_{i,k-\tau'}, D^{*\iota}_{ik})^2}{\sigma^2_Y}}, \tag{65}$$

$$A_{xy} = \frac{\bar{c}+\bar{\mu}-\frac{\bar{\alpha}}{r}+g(x,\delta_k)+(y-g(x,\delta_{k-\tau'})-\bar{\mu})\frac{\text{Cov}(Y^0_{i,k-\tau'},D^{*\iota}_{ik})}{\sigma^2_Y}}{\sqrt{\sigma^2_{D^{*\iota}}-\frac{\text{Cov}(Y^0_{i,k-\tau'},D^{*\iota}_{ik})^2}{\sigma^2_Y}}}. \tag{66}$$

In order to obtain bias terms that are comparable to those calculated for DID matching, we have to integrate B^{m1}_{xy} and B^{m2}_{xy} with respect to the distribution $F_{Y^0_{i,k-\tau'}|D^\iota_{i,k}=1}(y)$, which has the following density (Arnold, Beaver, Groeneveld, and Meeker, 1993):

$$f_{Y^0_{i,k-\tau'}|D^\iota_{i,k}=1}(y) = \frac{1}{\sigma_Y}\phi\left(\frac{y-g(x,\delta_{k-\tau'})}{\sigma_Y}\right)\frac{1-\Phi(A_{xy})}{1-\Phi(A_x)}. \tag{67}$$

After integrating out $Y_{i,k-\tau'}|D_{ik}=1$, we have:

$$\mathbb{E}[\mathbb{E}[Y^0_{i,k+\tau}|D^\iota_{ik}=0,Y^0_{i,k-\tau'}]|D^\iota=1]=\mathbb{E}[Y^0_{i,k+\tau}]-\gamma_{\tau,\tau'}\frac{\mathrm{Cov}(Y^0_{i,k-\tau'},D^{*\iota}_{ik})}{\sigma_{D^{*\iota}}}\frac{\phi(A_x)}{1-\Phi(A_x)}$$
$$+\gamma_{\tau,D^{*\iota}}\int_{-\infty}^{+\infty}\frac{1}{\sigma_Y}\frac{\phi(A_{xy})}{\Phi(A_{xy})}\phi\left(\frac{y-g(x,\delta_{k-\tau'})-\bar{\mu}}{\sigma_Y}\right)\frac{1-\Phi(A_{xy})}{1-\Phi(A_x)}dy. \quad (68)$$

There is no closed form expression for the last integral. I use 32-point Gauss-Hermite quadrature to compute this integral numerically.

C Parameterizations of the Monte-Carlo simulations

The g function and the selection equation take the following from:

$$g(X_i,\delta_t)=\alpha_a+\beta_a A_{i,t}+\gamma_a A^2_{i,t}+(\delta+r_t d)E_i \quad (69)$$
$$D^{*\iota}_{i,k}=\alpha_x+\beta_x E_{i,t}+\frac{\alpha_i}{r}-c_i-\mathbb{E}[Y^0_{i,k}|\mathcal{I}^\iota_{i,k}]. \quad (70)$$

Bayesian updating in the HIP model follows Guvenen (2007). The state and equations have the following form:

$$\underbrace{\begin{bmatrix}\mu_i\\ \beta_i\\ U_{i,t+1}\end{bmatrix}}_{\mathbf{S}_{i,t+1}}=\underbrace{\begin{bmatrix}1&0&0\\0&1&0\\0&0&\rho\end{bmatrix}}_{\mathbf{F}}\underbrace{\begin{bmatrix}\mu_i\\ \beta_i\\ U_{i,t}\end{bmatrix}}_{\mathbf{S}_{i,t}}+\underbrace{\begin{bmatrix}0\\0\\v_{i,t+1}\end{bmatrix}}_{\mathbf{v}_{i,t+1}} \quad (71)$$

$$y^0_{i,t}=\underbrace{\begin{bmatrix}1&t&1\end{bmatrix}}_{\mathbf{H}'_t}\underbrace{\begin{bmatrix}\mu_i\\ \beta_i\\ U_{i,t}\end{bmatrix}}_{\mathbf{S}_{i,t}}, \quad (72)$$

where $y^0_{i,t}=Y^0_{i,t}-g(X_i,\delta_t)$.

As all variables are normally distributed, the prior belief over $(\mu_i, \beta_i, U_{i,0})$ is a multivariate normal distribution with mean $\hat{\mathbf{S}}_{i,1|0} \equiv (0, \beta_i^k, 0)$ and covariance matrix:

$$\mathbf{P}_{1|0} = \begin{bmatrix} \sigma_\alpha^2 & \sqrt{1-\lambda}\sigma_{\alpha,\beta} & 0 \\ \sqrt{1-\lambda}\sigma_{\alpha,\beta} & \sqrt{1-\lambda}\sigma_{\alpha,\beta} & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}. \tag{73}$$

After observing t periods of outcomes, the individual's posterior for $(\mu_i, \beta_i, U_{i,t})$ is a normal distribution with mean $\hat{\mathbf{S}}_{i,t|t}$ and covariance matrix $\mathbf{P}_{t|t}$. From this, the individual can form one period ahead forecasts of these variables. They will also be normally distributed with mean $\hat{\mathbf{S}}_{i,t+1|t}$ and covariance matrix $\mathbf{P}_{t+1|t}$. The evolution of these matrices induced by optimal learning is:

$$\hat{\mathbf{S}}_{i,t|t} = \hat{\mathbf{S}}_{i,t|t-1} + \mathbf{P}_{t|t-1}\mathbf{H}_t \left[\mathbf{H}_t'\mathbf{P}_{t|t-1}\mathbf{H}_t\right]^{-1} \times \left(y_{i,t} - \mathbf{H}_t'\hat{\mathbf{S}}_{i,t|t-1}\right) \tag{74}$$

$$\hat{\mathbf{S}}_{i,t+1|t} = \mathbf{F}\hat{\mathbf{S}}_{i,t|t} \tag{75}$$

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}_t \left[\mathbf{H}_t'\mathbf{P}_{t|t-1}\mathbf{H}_t\right]^{-1} \times \mathbf{H}_t'\mathbf{P}_{t|t-1} \tag{76}$$

$$\mathbf{P}_{t+1|t} = \mathbf{F}\mathbf{P}_{t|t}\mathbf{F}' + \mathbf{Q}, \tag{77}$$

with $\mathbf{Q}$ the covariance matrix of $\mathbf{v}_{i,t+1}$.

Conditional on individual's beliefs at period t, log wages is normally distributed with mean $\mathbf{H}_t'\hat{\mathbf{S}}_{i,t+1|t} + g(X_i, \delta_{t+1})$. These expected foregone wages are then fed in the selection equation.

Table 1 – 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
α_a	8.83	8.83	8.83	8.83
β_a	0.56	0.56	0.56	0.56
γ_a	-0.057	-0.057	-0.057	-0.057
α_x	0	0.5	0.5	0.6
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
$\sigma_{U_0}^2$	σ^2	$\sigma_{U_\infty}^2$	σ^2	$\sigma_{U_\infty}^2$