To Control or Not to Control? Bias of Simple Matching *vs* Difference-In-Difference Matching in a Dynamic Framework

Sylvain Chabé-Ferret *

February 12, 2010

Abstract

In this paper, I examine whether the claim "the more control variables, the better" holds when matching in a dynamic context. I exhibit two situations where it is better not to control for past outcomes because DID matching is unbiased whereas matching on past outcomes is biased. I study the special case of evaluating a job training program, where both estimators are biased, borrowing a credible selection rule from Heckman, LaLonde, and Smith (1999) and relying on the parameters of the wage process estimated by MaCurdy (1982). I derive closed forms for the bias terms of the two estimators when the error terms are normally distributed. I show that DID matching performs better when used symmetrically around the period of enrollment, as implemented by Heckman, Ichimura, Smith, and Todd (1998). Matching is less biased than DID-matching when considering to use the first pre-enrollment period, but symmetric DID-matching is less biased than the less biased matching estimator.

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

JEL codes: C21, C23.

*CEMAGREF, UMR MÉTAFORT, 24 avenue des Landais, BP 50085, 63172 Aubière Cedex, France. Email: sylvain.chabe-ferret@cemagref.fr. This paper has been written as I was visiting the Cowles Foundation for the Research in Econmics at Yale University. I thank Don Andrews, Philip Haile and Ed Vytlacil for their invitation and Cemagref for its financial support. I thank Stefan Hoderlein, Yoonseok Lee, Julie Subervie, Neese Yildiz and Ed Vytlacil for useful discussions on this research.

1 Introduction

When considering the practical implementation of matching methods, it is widely believed that the more pre-treatment variables one can control for, the lower the bias of the estimated treatment effect.[1] Dehejia and Wahba (1999), in their re-anaylis of LaLonde (1986)'s critique, argue that controlling for past outcomes goes a long way in reducing the bias of propensity score matching.[2] Heckman, Ichimura, Smith, and Todd (1998) also show that when adding past wages and more importantly on past labor transitions, greatly decreases bias of propensity score matching estimators.

The claim that the more control variables the better has been challenged by Heckman and Navarro-Lozano (2004): these authors show that, in a static setting, controlling for more variables or for proxies may on the contrary increase bias. Whether it does so depends on the correlation between the added control variable and the remaining unobservable confounders. Wooldridge (2005) has also shown that controlling for post-treatment outcomes may lead to bias.

In this paper, I examine whether the claim that the more control variables, the better, also holds in a dynamic context. Time varying control variables combining persistent unobserved heterogeneity and true state dependence are indeed the most common case encountered in micro-economics. Past outcomes are a particular case of this type of control variables. Is it always better to control for past outcomes when we observe them? To answer this question is equivalent to comparing the bias of simple matching *vs* Difference-In-Difference (DID) matching.[3] To see why this is the case, note that matching on all pre-treatment covariates, including past outcomes, implies that pre-treatment mean differences between the treated and their matched counterparts converges to zero when the sample size gets large. When considering the use of past outcomes as control variables, we can thus restate the previous question as: is it better to control for past outcomes by including them in the set of control variables in a simple matching procedure, or is bias lower when not controlling for past outcomes and applying DID instead?

In this paper, I answer this question in two steps: in section 2, I introduce the problem in a general non structural framework. I exhibit two situations where it is better not to control for past outcomes because DID matching is unbiased whereas matching on past outcomes is biased: first if past outcomes do not directly determine current outcomes, but determine selection into the treatment; second, if past outcomes directly determine current outcomes, but do not directly determine selection into the treatment and outcomes follow a linear autoregressive process starting at a random point around its long run equilibrium. These are to my knowledge the first examples of controlling for too many variables in a dynamic context. In section 3, I examine the case where both current outcomes and selection depend on past outcomes. In this case, both estimators are biased. I study the special case of evaluating a job training program, borrowing a credible selection rule from Heckman, LaLonde, and Smith (1999) and relying on the parameters of the wage process estimated by MaCurdy (1982). I derive closed forms for the bias terms of the two estimators when the error terms are normally distributed. I show that DID matching performs better when used symmetrically around the period of enrollment, as implemented by Heckman, Ichimura, Smith, and Todd (1998). Matching is less biased than DID-matching when considering to use the first pre-enrollment period, but symmetric DID-matching is less biased than the less biased matching estimator. I have thus proved in an empirically relevant example that controlling for observed covariates may indeed enhance bias.

This paper is organized as follows: section 2 presents a non structural first pass at DID-matching while section 3 is an in-depth study of the evaluation of a job training program. Section 4 concludes.

[1] See for example the discussion at `http://www.stat.columbia.edu/~cook/movabletype/archives/2010/02/a_propensity_fo.html`.

[2] Smith and Todd (2005) argue that this is the special structure of Dehejia and Wahba (1999)'s sample that makes the most part in the bias reduction.

[3] DID matching has been introduced by Heckman, Ichimura, and Todd (1997). An alternative estimator based on the same identifying conditions has been proposed by Abadie (2005).

2 A Non Structural First-Pass at DID-matching

I start the analysis by building on Ashenfelter and Card (1985)'s simple setting. By progressively extending their simple framework, the main problems appear and a series of results can be stated. As in Ashenfelter and Card (1985), write the observed outcome as: $Y_{it} = \mu_i + d_t + \beta D_{it} + \epsilon_{it}$, where μ_i and d_t are unobserved respectively individual and time fixed effects, ϵ_{it} is an i.i.d. shock and β is a fixed parameter. Participation in the program whose average effect we would like to measure is indicated by D_{it}. Suppose we only observe two periods of data, k and $k-1$: in period $k-1$, non one receives treatment ($D_{ik-1} = 0$). In period k, treatment is allocated among the population according to the following rule: $D_{ik} = \mathbb{1}\left[\mu_i + u_{ik} \geq 0\right]$, where u_{ik} is an i.i.d. shock independent from all other variables in the model. In this simple setting, the difference-in-difference (DID) estimator recovers the treatment on the treated parameter: $\mathbb{E}[Y_{ik} - Y_{ik-1}|D_{ik} = 1] - \mathbb{E}[Y_{ik} - Y_{ik-1}|D_{ik} = 0] = \beta = \mathbb{E}\left[Y^1_{ik} - Y^0_{ik}|D_{ik} = 1\right]$, where Y^1_{ik} is the value of Y_{ik} if individual i receives treatment in period k and Y^0_{ik} is the value of Y_{ik} if individual i does not receives the treatment in period k.

The previous setting is highly restrictive in that we need to assume unconditional linear separability of both time and individual fixed effects from the treatment effect and that we make no use of potential observed control variables. A non linear version of the previous argument could take the following form: $Y_{it} = g(D_{it}, d_t, X_i, \epsilon_{it}) + \mu_i$ and $D_{ik} = \mathbb{1}\left[h(\mu_i, X_i, u_{ik}) \geq 0\right]$, with g and h unrestricted functions and X_i observed variables fixed through time and potentially correlated to μ_i but neither to u_{it} or ϵ_{it}. In this much less restrictive setting, treatment on the treated is identified by DID-matching. This is because the independent increments condition is fulfilled:

$$\mathbb{E}\left[Y^0_{ik} - Y^0_{ik-1}|D_{ik} = d, X_i\right] = \mathbb{E}\left[g(0, d_k, X_i, \epsilon_{ik}) - g(0, d_{k-1}, X_i, \epsilon_{ik-1})\right]$$

does not depend on d. Together with a support condition, this restriction implies that treatment on the treated is identified:

$$\mathbb{E}\left[\mathbb{E}[Y_{ik} - Y_{ik-1}|D_{ik} = 1, X_i] - \mathbb{E}[Y_{ik} - Y_{ik-1}|D_{ik} = 0, X_i]\,|D_{ik} = 1\right] = \mathbb{E}\left[Y^1_{ik} - Y^0_{ik}|D_{ik} = 1\right].$$

This identification strategy is powerful: it allows for different growth rates of the outcome variables in the treated and untreated groups, due to different observed initial conditions. It is even possible to accommodate interaction between observed variables and fixed effects and to allow the treatment effects to depend on unobservables (i.e. essential heterogeneity à la Heckman, Urzua, and Vytlacil (2006) can be allowed for in this setting): $Y_{it} = g^a(D_{it}, d_t, X_i, \epsilon^a_{it}, D_{it}\mu^1_i) + g^b(X_i, \epsilon^b_{it}, D_{it}\mu^1_i + (1 - D_{it})\mu^0_i)$ and $D_{ik} = \mathbb{1}\left[h(\mu^1_i, \mu^0_i, X_i, u_{ik}) \geq 0\right]$. The independent increment condition is also fulfilled in this model, meaning that selection on unobserved gains to the treatment does not jeopardize the identification strategy. We have thus proved the following result:

Result 1 *DID-matching identifies treatment on the treated if control variables are fixed through time, unobserved fixed time and individual effects are additively separable and selection bias is only due to the unobserved fixed effects terms.*

Crucial to this result is the fact that μ_i and d_t are linearly separable in the outcome equation. This means that DID-matching is not stable to monotonic nonlinear transformations of the outcome variable (Athey and Imbens, 2006). It is for example well-known that applying DID in logs and in levels may lead to opposite signs for the estimated treatment effects Meyer, Viscusi, and Durbin (1995). A second restriction of the setting presented so far is that we have only considered control variables that are fixed through time. In many applied settings, there exists potential control variables that may vary between $k-1$ and k. In labor economics, age is a natural candidate as a time-varying control variable, but pre-treatment outcomes (past wages and employment status) are also frequently considered as control variables. In the studies of firms, past levels of capital stock and number of employees are also often included as controls. Do we really know that DID-matching still identifies treatment effects in this case? In the remaining of this section, I carefully examine whether the crucial independent increments conditions is fulfilled for three types of time-varying candidate variables: fully exogenous variables, variables correlated to the fixed effects and past-outcomes.

2.1 Exogenous time-varying covariates

Let's write outcome and selection as a function of time varying covariates: $Y_{it} = g(D_{it}, d_t, X_{it}, \epsilon_{it}) + \mu_i$ and $D_{ik} = \mathbb{1}\left[h(\mu_i, X_{ik-1}, u_{ik}) \geq 0\right]$. In this example, selection into the program in period k depends on period $k-1$ control variables: this is a setting that is often found in practice. The analysis could also be done with selection driven by period k controls without much alteration in the conclusions. In this setting, there is a selection problem due to time-varying covariates only if there is true state dependence in the process generating these variables. I write $X_{it} = i(X_{it-1}, D_{it}, \eta_{it})$: covariates in period t depend on past values of the covariates, thereby creating a selection on observables problem, on current treatment status, allowing potential effect of the treatment, and on an i.i.d. shock independent of all other variables in the model. Note that treatment has no effect on covariates in period $k-1$: I assume no anticipation of the treatment. Before we proceed with the analysis of DID-matching in this case, it is to be noted that the general framework adopted here has consequences for the definition of the individual level treatment effect. We have to make a distinction between a partial treatment effect, where X_{it} is held constant ($\Delta^x Y_{it} = g(1, d_t, x, \epsilon_{it}) - g(0, d_t, x, \epsilon_{it})$) and a complete treatment effect where the indirect effect of treatment mediated through X_{it} is taken into account ($\Delta Y_{it} = g(1, d_t, i(X_{it-1}, 1, \eta_{it}), \epsilon_{it}) - g(0, d_t, i(X_{it-1}, 0, \eta_{it}), \epsilon_{it})$).

The most natural control variable in this setting is X_{ik-1}. It is easy to show in that case that the independent increments condition is satisfied conditional on X_{ik-1}:

$$\begin{aligned}
\mathbb{E}\left[Y^0_{ik} - Y^0_{ik-1} | D_{ik} = d, X_{ik-1} = x\right] &= \mathbb{E}\left[g(0, d_k, i(x, 0, \eta_{ik}), \epsilon_{ik}) | \mathbb{1}\left[h(\mu_i, x, u_{ik}) \geq 0\right] = d\right] - \mathbb{E}\left[g(0, d_{k-1}, x, \epsilon_{ik-1})\right] \\
&= \mathbb{E}\left[g(0, d_k, i(x, 0, \eta_{ik}), \epsilon_{ik})\right] - \mathbb{E}\left[g(0, d_{k-1}, x, \epsilon_{ik-1})\right].
\end{aligned}$$

The last equality follows because μ_i and u_{ik} are independent of η_{ik} and ϵ_{ik}.

However, controlling only for X_{ik} leads to a biased estimation for two reasons: first, X_{ik} is partly determined by the treatment, so that there is an overcontrol bias. But even if there is no effect of treatment on X_{ik} (i.e. if the function i is everywhere constant in its second argument), this estimation strategy would be biased because the independent increments condition would not be fulfilled:

$$\begin{aligned}
\mathbb{E}\left[Y^0_{ik} - Y^0_{ik-1} | D_{ik} = d, X_{ik} = x\right] &= \mathbb{E}\left[g(0, d_k, x, \epsilon_{ik})\right] \\
&\quad - \mathbb{E}\left[g(0, d_{k-1}, X_{ik-1}, \epsilon_{ik-1}) | \mathbb{1}\left[h(\mu_i, X_{ik-1}, u_{ik}) \geq 0\right] = d, X_{ik-1} = i(x, 0, \eta_{ik})\right].
\end{aligned}$$

These increments depend on d because the term on the second line is not equal to its unconditional version: treated people with $X_{ik} = x$ have different X_{ik-1} than untreated people with the same value for X_{ik}. This result applies only to variables who have a stochastic component to their relationship with past outcomes. Variables with non-stochastic dynamic relationships (e.g. age) are not subject to bias if the observation from the proper period is not used as a control.

Finally, we could control for all the observed variables. In that case, the independent increments condition is fulfilled. If X_{ik} depends on treatment status, there is overcontrol bias, but the estimated parameter can be given causal content: it is the direct treatment effect on the treated ($\mathbb{E}[\Delta^x Y_{ik} | D_{ik} = 1]$). If X_{ik} does not depend on treatment status, treatment on the treated is identified. We thus have proved the following result holds:

Result 2 *DID-matching identifies treatment on the treated if time-varying control variables are independent of the individual fixed effects. Controlling for the value of the covariates at the period when treatment is decided is sufficient and necessary.*

2.2 Endogenous time-varying covariates

In this section, we allow X_{it} to depend on μ_i: $X_{it} = i(X_{it-1}, \mu_i, \eta_{it})$. Contrary to the previous result, controlling for X_{ik-1} does not solve the problem of selection bias: the independent increments property is not fulfilled in that case:

$$\mathbb{E}\left[Y^0_{ik} - Y^0_{ik-1} | D_{ik} = d, X_{ik-1} = x\right] = \mathbb{E}\left[g(0, d_k, i(x, \mu_i, \eta_{ik}), \epsilon_{ik}) | \mathbb{1}\left[h(\mu_i, x, u_{ik}) \geq 0\right] = d\right] - \mathbb{E}\left[g(0, d_{k-1}, x, \epsilon_{ik-1})\right]$$

The previous equality depends on d because the first element on the right hand side is not equal to its unconditional version. The conditioning set restricts the values of μ_i, and as μ_i is in the expectation term, this conditional expectation depends on d. Thus result 2 breaks down when time varying control variables are endogenous. The intuition for this result is quite subtle: when comparing treated and untreated people with the same value for X_{ik-1}, we do not compare the "same" average people: they differ in their mean level of μ_i: those who participate do so because they have different values of the unobserved individual fixed effects. In the exogenous case, this is not a problem as all the effect of different μ_i is canceled out by first-differencing the outcome: time persistent unobservables are differenced out. But in the endogenous controls case, different μ_i mean different X_{ik}: treated and untreated units have different long term values for the control variables, and they do converge to these different values between $k-1$ and k. By matching treated and untreated on X_{ik}, we render them comparable in an observed dimension, but they still differ in their unobserved dynamic behavior.

Surprisingly enough, controlling for time-varying covariates in both periods restores the independent increments property and the unbiasedness of DID-matching.

$$\mathbb{E}\left[Y_{ik}^0 - Y_{ik-1}^0 | D_{ik} = d, X_{ik} = x, X_{ik-1} = \tilde{x}\right] = \mathbb{E}\left[g(0, d_k, x, \epsilon_{ik})\right] - \mathbb{E}\left[g(0, d_{k-1}, \tilde{x}, \epsilon_{ik-1})\right].$$

If X_{ik} is influenced by the treatment, the estimated treatment effect is the average partial effect. If X_{ik} is not determined by D_{ik}, the estimated treatment effect is the average complete effect.

Result 3 *With endogenous time-varying covariates, DID-matching controlling only on covariates at the time of selection is biased. When controlling on covariates both at the time of selection and at the time when the outcome is measured, DID-matching is unbiased for the average partial effect on the treated. If treatment does not affect the control variables in the period when the outcome is measured, the estimated treatment effect is the complete effect.*

2.3 Past outcomes as covariates

We can now study the case of particular covariates: past outcomes. Controlling for past outcomes in the matching process and then using DID is unnecessary relative to the same period: matching eliminates all differences in past outcomes between the treated and their matched counterparts. Matching and DID-matching are equivalent when past outcomes are included as control variables.[4] I consider three distinct cases: past outcomes determine participation or current outcomes or both.

2.3.1 Past outcomes only determine selection

Let's write $D_{ik} = \mathbb{1}\left[h(Y_{ik-1}, \mu_i, u_{ik}) \geq 0\right]$ and $Y_{it} = g(D_{it}, d_t, \epsilon_{it}) + \mu_i$. In that setting, and quite surprisingly as it determines selection, controlling on past outcome leads to bias while not controlling leads to unbiased estimation of the treatment on the treated parameter. To prove this result, first note that DID-matching when controlling on Y_{ik-1} is equivalent to simple matching. Unbiasedness in this setting thus means that the conditional expectation of Y_{ik}^0 conditional on past outcome must not depend on treatment status:

$$\mathbb{E}\left[Y_{ik}^0 | D_{ik} = d, Y_{ik-1} = y\right] = \mathbb{E}\left[g(0, d_k, \epsilon_{it})\right] + \mathbb{E}\left[\mu_i | \mathbb{1}\left[h(y, \mu_i, u_{ik}) \geq 0\right] = d\right].$$

This equality depends on d because the second term on the right hand side is not equal to its unconditional counterpart: units selected into the treatment converge to a different long term value of the outcome than units not selected into the treatment but with the same pre-treatment outcome. It is on the contrary obvious that in this case simple DID without matching (or matching only on observable characteristics fixed through time) is not biased, because the independent increments condition is verified unconditionally. This setting is thus an occurrence of biased generated by the desire to control "too much". We generate bias by trying to render treated and untreated units identical along observed dimensions: by doing so, we generate time varying bias that cannot be corrected by time-differencing. This is to my knowledge the first instance of bias due to too many controls in the panel data setting.[5] This also sheds light on whether one should use proxies

[4]This has nevertheless been somewhat overlooked in the literature: Heckman, Ichimura, Smith, and Todd (1998); Heckman, Ichimura, and Todd (1997); Heckman and Smith (1999) use earnings in the month of enrollment as a control variable and estimate how the bias term varies through time, including in the quarter containing the month of enrollment. Smith and Todd (2005) use both income in 1974 and 1975 as control variables in some of their specifications while first-differencing across the same years.

[5]See Heckman and Navarro-Lozano (2004) and Wooldridge (2005) for examples in the cross-section case.

as controls (variables correlated with μ_i but that have no direct effect on outcomes): the answer is no, as long as the additive separability is maintained. Otherwise, both approaches are biased.

Result 4 *If past outcomes do not directly determine current outcomes, but determine selection into the treatment, controlling for past outcomes biases DID-matching, while not doing so preserves unbiasedness.*

2.3.2 Past outcomes only determine current outcomes

Let's write $D_{ik} = \mathbb{1}\left[h(\mu_i, u_{ik}) \geq 0\right]$ and $Y_{it} = g(Y_{it-1}, D_{it}, d_t, \epsilon_{it}) + \mu_i$. In that setting, matching on Y_{ik-1} is equivalent to DID-matching and is biased, because the fixed effect term is not differenced out. The bias term is equal to the mean difference in the unobserved fixed effects between treated and untreated with the same value of the past outcome, integrated over the distribution of past outcomes in the treated group. This bias appears because identical people (in terms of past outcomes) are different in terms of fixed effects and thus have different long run value for their outcomes. They happen to have the same outcomes when they enter the treatment because of different histories of shocks. They start converging to their different long run values just after receiving the treatment.

Simple DID is also biased in this case, because first-differencing does generally not eliminate all the influence of the individual fixed effects. If there exists a long run equilibrium point or path to which the Y_{it} process converges, this long run value generally depends on μ_i. Then, two cases can be considered in turn. First, if Y_{ik-1} is not a random point around the long run equilibrium (i.e. the process has just started a few periods ago: individuals have just entered the labor market without specific knowledge of μ_i), then the process will start its convergence toward the long run equilibrium along different paths for treated and untreated individuals, since they have different long run values for their equilibrium level. In this case, the independent increments condition is violated. Second, if Y_{ik-1} is a random point around the long run equilibrium, individuals will start converging toward their long run equilibrium and the bias of DID depends on the speed of this process, which in turns depend on the first derivative of the g function with respect to its first argument. Generally, this first derivative does depend on Y_{ik-1}, and thus on μ_i: mean speed of convergence will vary between treated and untreated groups, and DID will be biased. An important special feature of this second case appears if g is linear in its first argument: Blundell and Bond (1998) show that in this case the average speed of adjustment is independent from μ_i and thus both group follow the same mean path toward their long run equilibrium. In the case of an additive autoregressive process that starts around its long run equilibrium, simple DID is thus not biased whereas controlling on past outcomes would be.

Result 5 *If past outcomes directly determine current outcomes, but do not directly determine selection into the treatment, both controlling and not controlling for past outcomes biases DID-matching. In the special case of a linear autoregressive process starting at a random point around its long run equilibrium, not controlling on past outcomes preserves unbiasedness.*

2.3.3 Past outcomes determine both selection and current outcomes

Let's write $D_{ik} = \mathbb{1}\left[h(Y_{ik-1}, \mu_i, u_{ik}) \geq 0\right]$ and $Y_{it} = g(Y_{it-1}, D_{it}, d_t, \epsilon_{it}) + \mu_i$. In that setting, both approaches are generally biased. Matching on past outcomes is biased because people having the same value of past outcomes have different values of μ_i and thus converge toward different long run equilibria after receiving the treatment. Not matching on past outcomes is also biased. The easiest way to see this is perhaps to show that even in the case of an additive autoregressive process starting around its long run equilibrium, DID would be biased. Let's write: $Y^0_{ik} = \rho Y_{ik-1} + d_k + \epsilon_{ik} + \mu_i$ and $Y_{ik-1} = \frac{\mu_i}{1-\rho} + \epsilon_{ik-1}$. In this case, the bias of the DID estimator (equal to the difference in mean increments between treated and untreated) is equal to: $(\rho - 1)(\mathbb{E}[\epsilon_{ik-1}|D_{ik} = 1] - \mathbb{E}[\epsilon_{ik-1}|D_{ik} = 0])$. This term is in general not null because D_{ik} depends directly on Y_{ik-1} and thus on ϵ_{ik-1}.

Result 6 *If past outcomes directly determine both current outcomes and selection into the treatment, both controlling and not controlling for past outcomes biases DID-matching.*

Thus far, we have mainly established impossibility results: in the most general case of endogenous time-varying covariates influenced by the treatment, DID-matching is biased. We have proved that controlling for past outcomes can be harmful in some cases, whereas not controlling could avoid bias in some special cases. But we have no general hint at whether one should control or not in general cases. I establish more precise results in the case of the evaluation of a labor program in the next section.

3 The Evaluation of Job Training Programs: a Structural Model

3.1 A simple model of the wage process and of entry in a job training program

This section heavily borrows on the setting presented in Heckman, LaLonde, and Smith (1999). In their handbook chapter, the authors show that with perfect credit markets, entry into a job market program in period k should be governed by the following rule for individual i:

$$D_{ik} = \mathbb{1}\left[\frac{\alpha_i}{r} \geq c_i + Y^0_{ik}\right], \tag{1}$$

where α_i is the individual level wage gain from the program, r is the interest rate, c_i are the administrative costs of entering the program (including transfers partially compensating the opportunity cost) and Y^0_{ik} are log-earnings in period k outside of the program (they measure the opportunity cost of entering the program during one period). Earnings are not observed to the econometrician for those individuals entering the program, but are assumed known to the individual when she decides to enter the program. I follow the suggestion in Heckman, LaLonde, and Smith (1999) in specifying the wage equation:

$$Y^0_{it} = g^0(X_i, d_t) + \mu_i + U_{it}, \text{ with } U_{it} = \rho U_{it-1} + \underbrace{m_1 v_{it-1} + m_2 v_{it-2} + m_0 v_{it}}_{\nu_{it}}, \tag{2}$$

where the exact formulation for the residual process comes from MaCurdy (1982)'s preferred specification for the dynamics of log-wage earnings (ARMA(1,2)). Note that I abstract from the problem of time varying covariates other than past outcomes in this analysis. The analyst wishes to estimate the effect of entering the program in period k on period $k+\tau$ mean outcome and that she considers controlling for past outcomes observed in period $k-\tau'$. Direct control for the direct determinants of entry into the program (Y_{ik}) is indeed impossible, wage in period k being unobserved because the individual participate in the program. Only proxies for this unobserved determinants can be used in the analysis. In the remaining of this section, I consider in turn two possible avenues for implementing DID-matching in this setting: controlling for past outcomes (which amounts to only using simple matching) *vs* not controlling for past outcomes (pure DID-matching).

3.2 Bias from pure DID-matching (not controlling for past outcomes)

If we just control for X_i and not for past outcomes and apply DID to the matched sample between periods $k+\tau$ and $k-\tau'$, the bias term is as follows (for $\tau > 1$):[6]

$$\begin{aligned}
B^{did}_x &= \mathbb{E}\left[U_{ik+\tau} - U_{ik-\tau'}|X_i = x, D_{ik} = 1\right] - \mathbb{E}\left[U_{ik+\tau} - U_{ik-\tau'}|X_i = x, D_{ik} = 0\right] \\
&= \left(\rho^{\tau+\tau'} - 1\right)\left(\mathbb{E}\left[U_{ik-\tau'}|X_i = x, D_{ik} = 1\right] - \mathbb{E}\left[U_{ik-\tau'}|X_i = x, D_{ik} = 0\right]\right) && \text{(3a)} \\
&+ \rho^{\tau+\tau'-2}\begin{pmatrix} (\rho m_1 + m_2)\left(\mathbb{E}\left[v_{ik-\tau'}|X_i = x, D_{ik} = 1\right] - \mathbb{E}\left[v_{ik-\tau'}|X_i = x, D_{ik} = 0\right]\right) \\ +\rho m_2\left(\mathbb{E}\left[v_{ik-\tau'-1}|X_i = x, D_{ik} = 1\right] - \mathbb{E}\left[v_{ik-\tau'-1}|X_i = x, D_{ik} = 0\right]\right)\end{pmatrix} && \text{(3b)} \\
&+ \rho^{\tau-2}\left(\rho^2 + \rho m_1 + m_2\right)\sum_{j=0}^{\tau'-1}\rho^j\left(\mathbb{E}\left[v_{ik-j}|X_i = x, D_{ik} = 1\right] - \mathbb{E}\left[v_{ik-j}|X_i = x, D_{ik} = 0\right]\right). && \text{(3c)}
\end{aligned}$$

The first part of the bias term (3a) is due to initial differences in the history of shocks up to period $k-\tau'$ between treated and untreated. These initial differences persist until period k, leading to different participation decisions, but progressively fade away as long as we get away from the enrollment period. This progressive return to the mean wage is confounded with a causal effect of the treatment by the DID estimator because it fails to take into account that some of the initial differences in earnings are due to

[6] If $\tau = 1$, in the last term, the term $(\rho^2 + \rho m_1 + m_2)$ is deleted and replaced inside the sum by $(\rho^2 + \rho m_1 + \mathbb{1}[j > 1]\, m_2)$.

transient shocks. This leads to an upward bias in the estimation of the treatment on the treated parameter. The farther away the two periods, the higher the bias. The dependence of this bias term on ρ, the persistence of shocks, is difficult to determine: on the one hand, the higher ρ, the lower the discrepancy between the two periods, because the initial differences tend to persist. On the other hand, a higher ρ leads to higher variance of the error terms, and thus may increase the size of the bias term.

The second part of the bias (3b) reflects the fact that some of the negative shocks to wages will get corrected in the immediate aftermath of period $k-\tau$, through the action of the MA terms. The estimated MA terms are generally negative: the negative shocks leading people to participate do not fully persist until period $k+\tau$. This bias term will also be positive because it leads to an overestimate of the initial mean wage differences between participants and non participants.

The last part of the bias term (3c) is the consequence of shocks that have happened after the pre-treatment control period ($k-\tau'$). This bias term will generally be negative because the AR term dominates the MA terms (with the estimates in MaCurdy (1982), we have $\rho^2+\rho m_1+m_2 \approx 0.5$) and because mean shocks are negatively correlated to participation through the participation equation (the lower the wage, the higher the probability of participation). So this term may offset the two previous terms, all the more so if τ' is large. Note finally that in the case of a pure white noise process ($\rho = m_1 = m_2 = 0$), the three bias terms are null because earnings in period t are not related to past shocks.

It is possible to derive closed form expressions for these bias terms if we make the assumption that the i.i.d MA terms are normal with variance σ^2 (see appendix A.1 for the derivation). From (3), we can derive the bias terms for the DID case (for $\tau,\tau' \geq 2$): [7, 8]

$$\frac{B_x^{did}}{\frac{\sigma^2}{\sigma^2_{D^*}}\left(\frac{\phi(A_x)}{1-\Phi(A_x)}+\frac{\phi(A_x)}{\Phi(A_x)}\right)} = -\left(\rho^{\tau+\tau'}-1\right)\rho^{\tau'-1}\left(\rho\frac{\sigma_U^2}{\sigma^2}+m_1+m_2\left(m_1+\rho\right)\right) \tag{4a}$$

$$-\rho^{\tau+2\tau'-4}(\rho m_1+m_2(\rho^2+1))\left(m_2+\rho m_1+\rho^2\right) \tag{4b}$$

$$-\rho^{\tau-2}\left(\rho^2+\rho m_1+m_2\right)\left(\frac{1-\rho^{2\tau'}}{1-\rho^2}+m_1\rho\frac{1-\rho^{2\tau'-2}}{1-\rho^2}+m_2\rho^2\frac{1-\rho^{2\tau'-4}}{1-\rho^2}\right) \tag{4c}$$

with $A_x = \frac{\bar{c}-\frac{\bar{\alpha}}{r}+g(x,d_k)}{\sigma_{D^*}}$, $\sigma^2_{D^*} = \sigma_U^2+\sigma_\mu^2+\sigma_c^2+\frac{\sigma_\alpha^2}{r^2}+2(\sigma_{\mu c}-\frac{\sigma_{\mu\alpha}}{r})$ and ϕ (resp. Φ) is the density (resp. cumulative distribution function) of the standard normal.

These bias terms have the expected signs: the first term (4a) is positive because $\rho<1$ and $\frac{\sigma_U^2}{\sigma^2}>1$, so that in general $\rho\frac{\sigma_U^2}{\sigma^2}+m_1+m_2\left(m_1+\rho\right)$ is positive. The second term (4b) is positive because $m_1<m_2<0$. The third term (4b)is negative because the AR term dominates the MA term, leading to positive terms inside the brackets.

Figure 1 presents the values of the three components of this bias term for the value of the parameters of the wage process estimated by MaCurdy (1982). Three conclusions emerge from this figure:

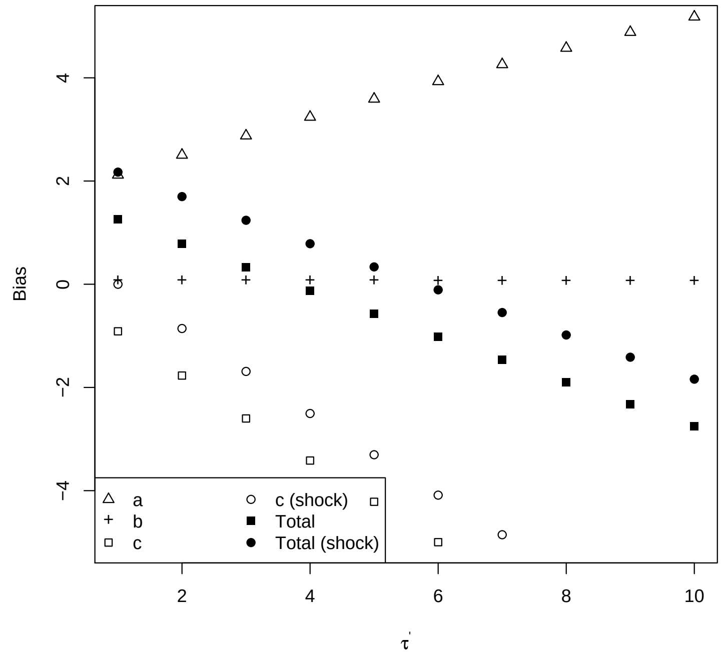
1. The main components of bias are the terms (4a) and (4c). They have opposite signs and both increase with τ', i.e. with the distance between treatment and the pre-treatment control period.

2. Total bias is null when $\tau = \tau'$: symmetric DID is unbiased though non symmetric DID is biased. The intuition for this result is the following: correlation between Y_{ik-t}, ($t=-\tau,\tau'$) and D_{ik} is due to the correlation between Y_{ik} and Y_{ik-t}. When $t=-\tau$, the correlation between $Y_{ik+\tau}$ and D_{ik} is weakened (compared to that between Y_{ik} and D_{ik}) because of the occurrence of τ periods of i.i.d. shocks. The same is true for $t=\tau'$: the τ' periods of shocks that have not yet occurred weaken the relationship between $Y_{ik+\tau'}$ and D_{ik}. When $\tau=\tau'$, the correlation between $Y_{ik-\tau'}$ and D_{ik} on the one hand and $Y_{ik+\tau}$ and D_{ik} on the other hand have been weakened by the same amount compared to the correlation between Y_{ik} and D_{ik}. This gives a rationale to the use of symmetric matching DID by Heckman, Ichimura, Smith, and Todd (1998); Heckman, Ichimura, and Todd (1997).

[7] When $\tau'=1$, the second component of the bias (4b) is equal to $-\rho^{\tau-1}\left((\rho m_1+m_2)(m_1+\rho)+\rho m_2(m_2+\rho m_1+\rho^2)\right)$ and the third component (4c) is equal to $-\rho^{\tau-2}(\rho^2+\rho m_1+\mathbb{1}[\tau>1]\,m_2)$.

[8] If $\tau=1$ and $\tau'>1$, the third component (4c) is equal to $\rho+m_1+\left(\rho^2+\rho m_1+m_2\right)\left(\rho\frac{1-\rho^{2\tau'-2}}{1-\rho^2}+m_1\frac{1-\rho^{2\tau'-2}}{1-\rho^2}+m_2\rho\frac{1-\rho^{2\tau'-4}}{1-\rho^2}\right)$.

Figure 1 – Bias from DID-matching (not controlling on past outcomes) for $\tau = 4$



Note: *a*, *b* and *c* stand respectively for the components (4a), (4b) and (4c) of the bias term of DID matching. (choc) corresponds to the value of the corresponding terms when v_{ik} is not observed when the individual decides to participate. All values have been computed according to MaCurdy (1982)'s estimates of the parameters: $\rho = 0.99$, $m_1 = -0.4$, $m_2 = -0.01$.

3. This result of minimal bias when implementing symmetric DID critically hinges on the assumption that individuals perfectly know the wages they would have earned if they have remained on their job at period k. If, for example, we assume that they do not know v_{ik} when deciding whether or not they enter the program, figure 1 (terms "shock") show that symmetric DID is no longer the least biased estimator.

This case study has thus proved the following result:

Result 7 *When evaluating a job training program, symmetric DID is the least biased estimator among the simple DID estimators when the agents know perfectly the wage they would have had in the period the program takes place if they had not entered the program.*

3.3 Bias from pure matching (controlling for past outcomes)

When controlling for past outcomes, a very usual relationship can be extracted from the common factor representation of the wage process in equation (2), making use of the fact that $U_{it} = Y^0_{it} - g^0(X_i, d_t) - \mu_i$:

$$Y^0_{ik+\tau} = g^0(X_i, d_{k+\tau}) + \rho^{\tau+\tau'} Y_{ik-\tau'} - \rho^{\tau+\tau'} g^0(X_i, d_{k-\tau'}) + \mu_i(1-\rho^{\tau+\tau'}) + \sum_{j=0}^{\tau+\tau'} \rho^j v_{ik+\tau-j}.$$

From this formula, we can derive the bias that arises from matching on X_i and $Y_{ik-\tau'}$ (for $\tau \geq 2$):

$$B^m_{xy} = \left(1-\rho^{\tau+\tau'}\right)\left(\mathbb{E}\left[\mu_i|X_i=x,Y_{ik-\tau'}=y,D_{ik}=1\right]-\mathbb{E}\left[\mu_i|X_i=x,Y_{ik-\tau'}=y,D_{ik}=0\right]\right) \tag{5a}$$

$$+\rho^{\tau+\tau'-2}\begin{pmatrix}(\rho m_1+m_2)\left(\mathbb{E}\left[v_{ik-\tau'}|X_i=x,Y_{ik-\tau'}=y,D_{ik}=1\right]-\mathbb{E}\left[v_{ik-\tau'}|X_i=x,Y_{ik-\tau'}=y,D_{ik}=0\right]\right)\\ +\rho m_2\left(\mathbb{E}\left[v_{ik-\tau'-1}|X_i=x,Y_{ik-\tau'}=y,D_{ik}=1\right]-\mathbb{E}\left[v_{ik-\tau'-1}|X_i=x,Y_{ik-\tau'}=y,D_{ik}=0\right]\right)\end{pmatrix} \tag{5b}$$

$$+\rho^{\tau-2}\left(\rho^2+\rho m_1+m_2\right)\sum_{j=0}^{\tau'-1}\rho^j\left(\mathbb{E}\left[v_{ik-j}|X_i=x,D_{ik}=1\right]-\mathbb{E}\left[v_{ik-j}|X_i=x,D_{ik}=0\right]\right). \tag{5c}$$

The first part of the bias term is due to differences in unobservables fixed through time between treated and untreated. This term arises because participation into the program is partly determined by differences in μ_i because of the opportunity cost of participating in the program in terms of foregone wages. The second part of the bias term arises because of the moving average components of the wage process. In terms of the participation equation, this term means that among people with the same pre-treatment wage, only those with a recent positive shock to their wages have decided to enroll into the program. This shock being transitory (40% of the shock will have disappeared after one year), these people tend to have lower wages in period k after a wage surge in period $k-1$. With lower potential wages in period k, they thus decide to enroll into the program, whose opportunity cost has decreased. After the adjustment in period k due to the MA(1) term, the rest of the positive wage shock tend to persist through time, due to the large AR(1) term. Wages for participants thus persist, but at a lower level than those of untreated people who had the same wages in period $k-1$. We thus tend to underestimate the impact of the program by confouding it with the fading out of a transient shock that lead people to participate in the first place. The last component of the bias does not depend on $Y_{ik-\tau'}$ because it is due to shocks posterior to period $k-\tau'$. This part of the bias is thus identical to the bias in the simple DID case when we do not control on $Y_{ik-\tau'}$.

In order to be able to compare the two estimators, we have to derive bias terms conditional on the same conditioning set (here $X_i=x$). We can derive expressions for the unconditional bias term of simple matching if we assume that the error terms are normally distributed. The bias term for the simple matching case is derived in appendix A.2. After integrating out $Y_{ik-\tau'}|D_{ik}=1,X_i=x$, we have:

$$B^m_x = -\frac{1-\rho^{\tau+\tau'}}{1-\Phi(A_x)}\left(\frac{\sigma_{D^*,\mu}}{\sigma_{D^*}}\phi(A_x)+\frac{\sigma_{D^*,\mu}-\frac{\sigma^2_\mu\sigma_{D^*,Y_{k-\tau'}}}{\sigma^2_Y}}{\sigma_{D^*}\sigma_Y\sqrt{1-\frac{\sigma^2_{D^*,Y_{k-\tau'}}}{\sigma^2_{D^*}\sigma^2_Y}}}\int_{-\infty}^{+\infty}\phi\left(\frac{y-g(x,d_{k-\tau'})}{\sigma_Y}\right)\phi(A_{xy})\frac{1-\Phi(A_{xy})}{\Phi(A_{xy})}dy\right) \tag{6a}$$

$$-\frac{\rho^{\tau+\tau'-2}}{1-\Phi(A_x)}\left[\left((\rho m_1+m_2)\frac{\sigma_{U_t,v_{t-\tau'}}}{\sigma_{D^*}}+\rho m_2\frac{\sigma_{U_t,v_{t-\tau'-1}}}{\sigma_{D^*}}\right)\phi(A_x)+\left((\rho m_1+m_2)\frac{\sigma_{U_t,v_{t-\tau'}}-\frac{\sigma^2\sigma_{D^*,Y_{k-\tau'}}}{\sigma^2_Y}}{\sigma_{D^*}\sigma_Y\sqrt{1-\frac{\sigma^2_{D^*,Y_{k-\tau'}}}{\sigma^2_{D^*}\sigma^2_Y}}}\right.\right.$$

$$\left.\left.+\ \rho m_2\frac{\sigma_{U_t,v_{t-\tau'-1}}-\frac{m_1\sigma^2\sigma^2_{D^*,Y_{k-\tau'}}}{\sigma^2_Y}}{\sigma_{D^*}\sigma_Y\sqrt{1-\frac{\sigma^2_{D^*,Y_{k-\tau'}}}{\sigma^2_{D^*}\sigma^2_Y}}}\right)\int_{-\infty}^{+\infty}\phi\left(\frac{y-g(x,d_{k-\tau'})}{\sigma_Y}\right)\phi(A_{xy})\frac{1-\Phi(A_{xy})}{\Phi(A_{xy})}dy\right] \tag{6b}$$

$$-\rho^{\tau-2}\left(\rho^2+\rho m_1+m_2\right)\left(\frac{1-\rho^{2\tau'}}{1-\rho^2}+m_1\rho\frac{1-\rho^{2\tau'-2}}{1-\rho^2}+m_2\rho^2\frac{1-\rho^{2\tau'-4}}{1-\rho^2}\right)\frac{\sigma^2}{\sigma^2_{D^*}}\left(\frac{\phi(A_x)}{1-\Phi(A_x)}+\frac{\phi(A_x)}{\Phi(A_x)}\right) \tag{6c}$$

The first part of each component of the bias term is identical to the DID case. The second part of each component is an integration of the distribution of the error term among the untreated conditional on the value of past outcomes with respect to the distribution of past outcomes in the treated population. In order to compute this integral, I use 32 points Gauss-Hermite integration rule.

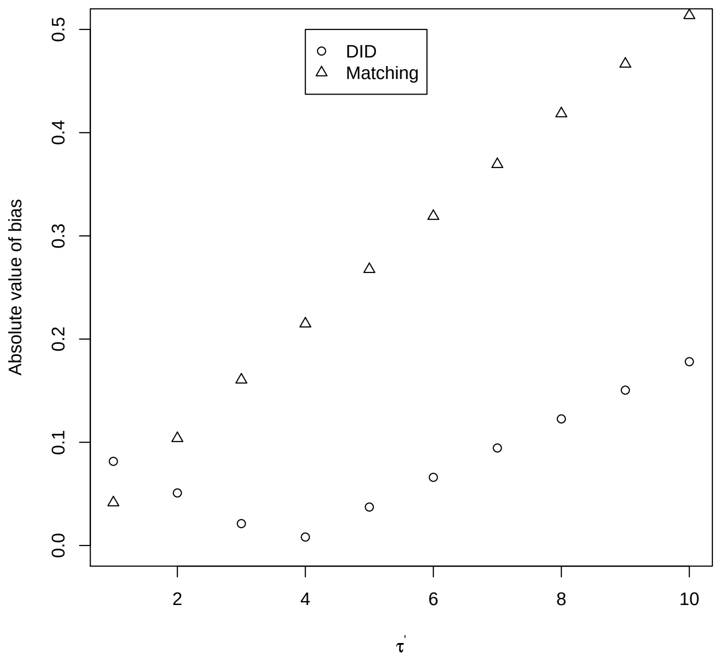
Figure 2 presents the absolute values of the bias terms of DID matching (equation 4) and simple matching (equation 6) with the values of the parameters estimated by MaCurdy (1982). Three important results emerge from this figure:

1. Not controlling on past outcomes has generally lower bias than controlling on past outcomes.
2. Controlling on past outcomes is the least biased when controlling for the period closest to enrollment (i.e. $\tau' = 1$). In that case, matching controlling on past outcome is less biased than DID matching with respect to the same period.
3. Symmetric DID matching is the least biased estimator: it is less biased than matching controlling on the period closest to enrollment.

All these results rely on the bias reduction property of symmetric simple DID: the mean pre-treatment difference in outcomes between treated and untreated τ' periods before enrollment is a good estimate of the mean post treatment difference in outcomes τ' periods after enrollment. We thus have proved the following result:

Result 8 *When evaluating a job training program, symmetric DID is less biased than matching controlling on past outcomes.*

Figure 2 – Comparison of bias when controlling (simple matching) and not controlling (DID matching) on past outcomes for $\tau = 4$



Note: the value of the bias for DID (resp. simple) matching are the absolute values of the terms in equation (4) (resp. 6) computed according to MaCurdy (1982)'s estimates of the parameters: $\rho = 0.99$, $m_1 = -0.4$, $m_2 = -0.01$, $\sigma^2 = 0.055$, $\sigma^2_\mu = 0$. The rest of the parameters takes the following values: $\bar{c} = -2$, $\sigma^2_c = 0.05$, $\sigma_{\mu c} = 0$, $\bar{\alpha} = 0.1$, $\sigma^2_\alpha = 0$, $\sigma_{\mu\alpha} = 0$, $r = 0.05$, $g(x, d_k) = g(x, d_{k-\tau'}) = 0$.

This result enable us to reinterpret some of the results in Heckman, Ichimura, Smith, and Todd (1998) and Smith and Todd (2005). Heckman, Ichimura, Smith, and Todd (1998) compares the relative ability of different set of control variables to reproduce the experimental results of the evaluation of 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 estimator is lower than that obtained thanks to the set of variables with the higher predicting power including wages at enrollment (see their table XII p.1062).

Smith and Todd (2005) use two set 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 lower than when they use the second set of controls (see their table 6, p.340).

4 Conclusion

In this paper, I study whether is always better to control for more covariates in a dynamic context. I study the relative biases case of matching on past covariates *vs* Difference in Difference matching. I exhibit two special cases where controlling on past outcomes is biased whereas implemented DID matching is not. In the general case where both estimators are biased, I study the special case of the evaluation of a job training program. Borrowing a credible selection rule from Heckman, LaLonde, and Smith (1999) and relying on the parameters of the wage process estimated by MaCurdy (1982), I derive closed forms for the bias terms of the two estimators when the error terms are normally distributed. I show that DID matching performs better when used symmetrically around the period of enrollment, as implemented by Heckman, Ichimura, Smith, and Todd (1998). Matching is less biased than DID-matching when considering to use the first pre-enrollment period, but symmetric DID-matching is less biased than the less biased matching estimator. These results are to my knowledge the first one that prove in a practical application that not controlling for past outcomes may be better than controlling for them.

This work can naturally be extended to controlling for covariates that are not past outcomes, but are correlated to them, like quasi-fixed factors when evaluating the effect of a treatment on variable factors. For example, in the evaluation of investment or hiring subsidies, should we control for past values of the capital stock? A second extension to this work would include examining the relative performances of other estimators when faced with the same problem: the Change in Change (CIC) estimator of Athey and Imbens (2006), the exchangeable estimator of Altonji and Matzkin (2005).

References

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

ALTONJI, J. G., AND R. L. MATZKIN (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73(4), 1053–1102.

ARNOLD, B., R. BEAVER, R. GROENEVELD, AND W. MEEKER (1993): "The Nontruncated Marginal of a Truncated Bivariate Normal Distribution," *Psychometrika*, 58(3), 471–488.

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

ATHEY, S., AND G. W. IMBENS (2006): "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica*, 74(2), 431–497.

BLUNDELL, R., AND S. BOND (1998): "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 87(1), 115–143.

DEHEJIA, R. H., AND S. WAHBA (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94(448), 1053–1062.

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

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

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

HECKMAN, J. J., AND J. SMITH (1999): "The Pre-Programme Earnings Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies," *Economic Journal*, 109(457), 313–348.

HECKMAN, J. J., S. URZUA, AND E. VYTLACIL (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88, 389–432.

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

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

MEYER, B. D., W. K. VISCUSI, AND D. L. DURBIN (1995): "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *The American Economic Review*, 85(3), 322–340.

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.

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

A.1 Not controlling for past outcomes

It is possible to derive closed form expressions for the bias terms of section 3.2 if we make the assumption that the i.i.d MA terms are normal with variance σ^2. I assume that the process generating the outcomes as begun 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 pose that $\alpha_i = \alpha$ is a constant, and that c_i is a normal variable with variance σ_c^2, independent of μ_i, whose variance is σ_μ^2. To obtain the biased terms, I study the joint distribution of for normal variables conditional on $X_i = x$, under the assumption that X_i is independent from μ_i. The observed variables we want to condition on are:

$$\begin{aligned} D_{ik}^* &= c_i - \frac{\alpha_i}{r} + g(x, d_k) + \mu_i + U_{ik} \\ Y_{ik-\tau'} &= g(x, d_{k-\tau'}) + \mu_i + U_{ik-\tau'} \\ Y_{ik-\tau'-1} &= g(x, d_{k-\tau'-1}) + \mu_i + U_{ik-\tau'-1}. \end{aligned}$$

So we need to derive the joint distribution of the following error terms: $\left(U_{ik}, U_{ik-\tau'}, U_{ik-\tau'-1}, \{v_{ik-j}\}_{0\le j\le\tau'-1}, v_{ik-\tau'}, v_{ik-\tau'-1}\right)$. This distribution is a centered normal with covariance matrix Σ_1:

$$\Sigma_1 = \begin{pmatrix} \sigma_U^2 & & & & & \\ \rho^{\tau'-1}\sigma_{U_t,U_{t-1}} & \sigma_U^2 & & & & \\ \rho^{\tau'}\sigma_{U_t,U_{t-1}} & \sigma_{U_t,U_{t-1}} & \sigma_U^2 & & & \\ \sigma_{U_t,v_{t-j}} & 0 & 0 & \sigma^2 & & \\ \sigma_{U_t,v_{t-\tau'}} & \sigma^2 & 0 & 0 & \sigma^2 & \\ \sigma_{U_t,v_{t-\tau'-1}} & \sigma^2(m_1+\rho) & \sigma^2 & 0 & 0 & \sigma^2 \end{pmatrix},$$

with $\sigma_U^2 = \sigma^2\left(1 + (m_1+\rho)^2 + \frac{\rho^4(\rho^2+\rho m_1+m_2)^2}{1-\rho^2}\right)$, $\sigma_{U_t,U_{t-1}} = \rho\sigma_U^2 + \sigma^2\left(m_1 + m_2\left(m_1+\rho\right)\right)$ and $\sigma_{U_t,v_{t-j}} = \rho^{j-2}(\mathbb{1}[j\neq 0]\,\mathbb{1}[j\neq 1]\,m_2 + \mathbb{1}[j\neq 0]\,\rho m_1 + \rho^2)\sigma^2$. The latter term comes from the fact that the shocks at t and $t-1$ are not fully adjusted for at period t: the first shock only enters directly while the second shock enters through the AR term and the MA term. All previous shocks enter in the same way, according to a weighted average of the ARMA terms. To obtain σ_U^2, note that $U_{it} = (\rho^2+\rho m_1+m_2)\sum_{j=2}^{\infty}\rho^j v_{it-j} + v_{it} + (m_1+\rho)v_{it-1}$. The variance of U_{it} is the sum of the variances of these three terms. The variance of the first term is: $\mathrm{Var}(\sum_{j=2}^{\infty}\rho^j v_{it-j}) = \sigma^2\sum_{j=2}^{\infty}\rho^{2j}$. As $\rho^2<1$ we can write: $\sum_{j=2}^{\infty}\rho^{2j} = \sum_{j=0}^{\infty}\rho^{2j} - 1 - \rho^2 = \frac{1-(1+\rho^2)(1-\rho^2)}{1-\rho^2} = \frac{\rho^4}{1-\rho^2}$, which gives the result. To obtain $\sigma_{U_t,U_{t-1}}$, note that $\mathrm{Cov}(U_t,U_{t-1}) = \mathrm{Cov}(\rho U_{it-1} + v_{it} + m_1 v_{it-1} + m2 v_{it-2}, U_{it-1}) = \rho\sigma_U^2 + m_1\mathrm{Cov}(v_{it-1},U_{it-1}) + m_2\mathrm{Cov}(v_{it-2},U_{it-1})$. As $U_{it-1} = \rho^2 U_{it-3} + v_{it-1} + m_1 v_{it-2} + m_2 v_{it-3} + \rho(v_{it-2} + m_1 v_{it-3} + m_2 v_{it-4})$, this leads to the result.

From these expressions, we can readily obtain the bias terms for the DID case (for $\tau,\tau'\ge 2$): [9, 10]

$$\begin{aligned} \frac{B_x^{did}}{\frac{\sigma^2}{\sigma_{D^*}^2}\left(\frac{\phi(A_x)}{1-\Phi(A_x)} + \frac{\phi(A_x)}{\Phi(A_x)}\right)} &= -\left(\rho^{\tau+\tau'}-1\right)\rho^{\tau'-1}\left(\rho\frac{\sigma_U^2}{\sigma^2} + m_1 + m_2\left(m_1+\rho\right)\right) \\ &\quad - \rho^{\tau+2\tau'-4}(\rho m_1 + m_2(\rho^2+1))\left(m_2+\rho m_1+\rho^2\right) \\ &\quad - \rho^{\tau-2}\left(\rho^2+\rho m_1+m_2\right)\left(\frac{1-\rho^{2\tau'}}{1-\rho^2} + m_1\rho\frac{1-\rho^{2\tau'-2}}{1-\rho^2} + m_2\rho^2\frac{1-\rho^{2\tau'-4}}{1-\rho^2}\right), \end{aligned}$$

with $A_x = \frac{\bar{c}-\frac{\bar{\alpha}}{r}+g(x,d_k)}{\sigma_{D^*}}$ and $\sigma_{D^*}^2 = \sigma_U^2 + \sigma_\mu^2 + \sigma_c^2 + \frac{\sigma_\alpha^2}{r^2} + 2(\sigma_{\mu c} - \frac{\sigma_{\mu\alpha}}{r})$.

If parts (4a) and (4b) of the bias term are quite self-explanatory, some calculation details are needed for part (4c). This term is the sum of the expectations of shocks after the control period τ' until k conditional on

[9] When $\tau'=1$, the second component of the bias (4b) is equal to $-\rho^{\tau-1}\left((\rho m_1+m_2)(m_1+\rho) + \rho m_2(m_2+\rho m_1+\rho^2)\right)$ and the third component (4c) is equal to $-\rho^{\tau-2}(\rho^2+\rho m_1+\mathbb{1}[\tau>1]\,m_2)$.

[10] If $\tau=1$ and $\tau'>1$, the third component (4c) is equal to $\rho + m_1 + \left(\rho^2+\rho m_1+m_2\right)\left(\rho\frac{1-\rho^{2\tau'-2}}{1-\rho^2} + m_1\frac{1-\rho^{2\tau'-2}}{1-\rho^2} + m_2\rho\frac{1-\rho^{2\tau'-4}}{1-\rho^2}\right)$.

participating (resp. not participating) in period k. Both conditional expectations depend on the covariance $\sigma_{U_k,v_{k-j}}$. Indeed, after factorization of $\frac{\sigma^2}{\sigma_D^2}\left(\frac{\phi(A_x)}{1-\Phi(A_x)}+\frac{\phi(A_x)}{\Phi(A_x)}\right)$, the sum term that remains is: $\sum_{j=0}^{\tau'-1}\rho^j\frac{\sigma_{U_k,v_{k-j}}}{\sigma^2} = \sum_{j=0}^{\tau'-1}\rho^{2j-2}(\mathbb{1}[j\neq 0]\,\mathbb{1}[j\neq 1]\,m_2+\mathbb{1}[j\neq 0]\,\rho m_1+\rho^2) = \sum_{j=0}^{\tau'-1}(\rho^2)^j + m_1\rho\sum_{j=1}^{\tau'-1}(\rho^2)^{j-1}+m_2\rho^2\sum_{j=2}^{\tau'-1}(\rho^2)^{j-2} = \sum_{k=0}^{\tau'-1}(\rho^2)^k + m_1\rho\sum_{k=0}^{\tau'-2}(\rho^2)^k + m_2\rho^2\sum_{k=0}^{\tau'-3}(\rho^2)^k$. If the decision maker does not know period k information (i.e. shock v_{ik}) when deciding to enter the program, the first term of the previous sum is changed to $\sum_{j=1}^{\tau'-1}(\rho^2)^j = \rho^2\sum_{j=1}^{\tau'-1}(\rho^2)^{j-1} = \rho^2\sum_{k=0}^{\tau'-2}(\rho^2)^k = \rho^2\frac{1-\rho^{2\tau'-2}}{1-\rho^2}$.

These bias terms have the expected signs: the first term (4a) is positive because $\rho<1$ and $\frac{\sigma_U^2}{\sigma^2}>1$, so that in general $\rho\frac{\sigma_U^2}{\sigma^2}+m_1+m_2\left(m_1+\rho\right)$ is positive. The second term (4b) is positive because $m_1<m_2<0$. The third term (4b)is negative because the AR term dominates the MA term, leading to positive terms inside the brackets.

A.2 Controlling for past outcomes

To derive a closed form expression for the bias term of DID-matching when controlling on past outcomes (i.e. simple matching), we have to derive the joint distribution of μ_i, $v_{ik-\tau'}$,$v_{ik-\tau'-1}$,D^*_{ik} and $Y_{ik-\tau'}$, which is a normal with mean $\mu_2=\left(0,0,0,\bar{c}+\frac{\bar{\alpha}}{r}+g(x,d_k),g(x,d_{k-\tau'})\right)$ and covariance Σ_2:

$$\Sigma_2=\begin{pmatrix} \sigma_\mu^2 & & & & \\ 0 & \sigma^2 & & & \\ 0 & 0 & \sigma^2 & & \\ \sigma_{D^*,\mu} & \sigma_{U_t,v_{t-\tau'}} & \sigma_{U_t,v_{t-\tau'-1}} & \sigma_{D^*}^2 & \\ \sigma_\mu^2 & \sigma^2 & m_1\sigma^2 & \sigma_{D^*,Y_{k-\tau'}} & \sigma_Y^2 \end{pmatrix},$$

with $\sigma_{D^*,Y_{k-\tau'}}=\sigma_{\mu c}-\frac{\sigma_{\mu\alpha}}{r}+\sigma_\mu^2+\rho^{\tau'-1}\sigma_{U_t,U_{t-\tau'}}$, $\sigma_{D^*,\mu}=\sigma_\mu^2+\sigma_{\mu c}-\frac{\sigma_{\mu\alpha}}{r}$ and $\sigma_Y^2=\sigma_\mu^2+\sigma_U^2$. The distribution of μ_i, $v_{ik-\tau'}$,$v_{ik-\tau'-1}$,D^*_{ik} conditional on $Y_{ik-\tau'}=y$ is normal, with mean μ_2^y calculated as the linear projection of these variables onto $Y_{ik-\tau'}$ and covariance matrix Σ_2^y which is the Schur complement of $(\sigma_\mu^2+\sigma_U^2)$ in Σ_2. We thus have:

$$\mu_2^{xy}=\begin{pmatrix} (y-g(x,d_{k-\tau'}))\frac{\sigma_\mu^2}{\sigma_Y^2} \\ (y-g(x,d_{k-\tau'}))\frac{\sigma^2}{\sigma_Y^2} \\ (y-g(x,d_{k-\tau'}))\frac{m_1\sigma^2}{\sigma_Y^2} \\ \bar{c}+\frac{\bar{\alpha}}{r}+g(x,d_k)+(y-g(x,d_{k-\tau'}))\frac{\sigma_{D^*,Y_{k-\tau'}}}{\sigma_Y^2} \end{pmatrix},$$

$$\Sigma_2^{xy}=\begin{pmatrix} \sigma_\mu^2\left(1-\frac{\sigma_\mu^2}{\sigma_Y^2}\right) & & & \\ -\sigma^2\frac{\sigma_\mu^2}{\sigma_Y^2} & \sigma^2\left(1-\frac{\sigma^2}{\sigma_Y^2}\right) & & \\ -\frac{m_1\sigma^2\sigma_\mu^2}{\sigma_Y^2} & -\frac{m_1\sigma^4}{\sigma_Y^2} & \sigma^2\left(1-\frac{m_1^2\sigma^2}{\sigma_Y^2}\right) & \\ \sigma_{D^*,\mu}-\frac{\sigma_\mu^2\sigma_{D^*,Y_{k-\tau'}}}{\sigma_Y^2} & \sigma_{U_t,v_{t-\tau'}}-\frac{\sigma^2\sigma_{D^*,Y_{k-\tau'}}}{\sigma_Y^2} & \sigma_{U_t,v_{t-\tau'-1}}-\frac{m_1\sigma^2\sigma_{D^*,Y_{k-\tau'}}}{\sigma_Y^2} & \sigma_{D^*}^2-\frac{\sigma_{D^*,Y_{k-\tau'}}^2}{\sigma_Y^2} \end{pmatrix}.$$

The first part of the bias term of matching on past values of the outcomes is thus:

$$B_{xy}^{m1}=-(1-\rho^{\tau+\tau'})\frac{\sigma_{D^*,\mu}-\frac{\sigma_\mu^2\sigma_{D^*,Y_{k-\tau'}}}{\sigma_Y^2}}{\sqrt{\sigma_{D^*}^2-\frac{\sigma_{D^*,Y_{k-\tau'}}^2}{\sigma_Y^2}}}\left(\frac{\phi\left(A_{xy}\right)}{1-\Phi\left(A_{xy}\right)}+\frac{\phi\left(A_{xy}\right)}{\Phi\left(A_{xy}\right)}\right),$$

where $A_{xy}=\frac{\bar{c}-\frac{\bar{\alpha}}{r}+g(x,d_k)+(y-g(x,d_{k-\tau'}))\frac{\sigma_{D^*,Y_{k-\tau'}}}{\sigma_Y^2}}{\sqrt{\sigma_{D^*}^2-\frac{\sigma_{D^*,Y_{k-\tau'}}^2}{\sigma_Y^2}}}$.

The second part of the bias term is:

$$B_{xy}^{m2} = -\rho^{\tau+\tau'-2}\left((\rho m_1 + m_2)\frac{\sigma_{U_t,v_{t-\tau'}} - \frac{\sigma^2\sigma_{D^*,Y_{k-\tau'}}}{\sigma_Y^2}}{\sqrt{\sigma_{D^*}^2 - \frac{\sigma_{D^*,Y_{k-\tau'}}^2}{\sigma_Y^2}}} + \rho m_2\frac{\sigma_{U_t,v_{t-\tau'-1}} - \frac{m_1\sigma^2\sigma_{D^*,Y_{k-\tau'}}}{\sigma_Y^2}}{\sqrt{\sigma_{D^*}^2 - \frac{\sigma_{D^*,Y_{k-\tau'}}^2}{\sigma_Y^2}}}\right)\left(\frac{\phi\left(A_{xy}\right)}{1-\Phi\left(A_{xy}\right)} + \frac{\phi\left(A_{xy}\right)}{\Phi\left(A_{xy}\right)}\right).$$

In order to obtain bias terms that are comparable to those calculated for DID matching, we have to integrate B_{xy}^{m1} and B_{xy}^{m2} with respect to the distribution $F_{Y_{ik-\tau'}|D_{ik}=1,X_i=x}(y)$. This distribution has the following density (Arnold, Beaver, Groeneveld, and Meeker, 1993):

$$f_{Y_{ik-\tau'}|D_{ik}=1,X_i=x}(y) = \frac{1}{\sigma_Y}\phi\left(\frac{y-g(x,d_{k-\tau'})}{\sigma_Y}\right)\frac{1-\Phi(A_{xy})}{1-\Phi(A_x)}.$$

We can derive expressions for the two unconditional bias terms after integrating out $Y_{ik-\tau'}|D_{ik}=1, X_i = x$:

$$B_x^{m1} = -\frac{1-\rho^{\tau+\tau'}}{1-\Phi(A_x)}\left(\frac{\sigma_{D^*,\mu}}{\sigma_{D^*}}\phi(A_x) + \frac{\sigma_{D^*,\mu} - \frac{\sigma_\mu^2\sigma_{D^*,Y_{k-\tau'}}}{\sigma_Y^2}}{\sigma_{D^*}\sigma_Y\sqrt{1-\frac{\sigma_{D^*,Y_{k-\tau'}}^2}{\sigma_{D^*}^2\sigma_Y^2}}}\int_{-\infty}^{+\infty}\phi\left(\frac{y-g(x,d_{k-\tau'})}{\sigma_Y}\right)\phi(A_{xy})\frac{1-\Phi(A_{xy})}{\Phi(A_{xy})}dy\right),$$

$$B_x^{m2} = -\frac{\rho^{\tau+\tau'-2}}{1-\Phi(A_x)}\left[\left((\rho m_1+m_2)\frac{\sigma_{U_t,v_{t-\tau'}}}{\sigma_{D^*}} + \rho m_2\frac{\sigma_{U_t,v_{t-\tau'-1}}}{\sigma_{D^*}}\right)\phi(A_x) + \right.$$
$$\left.\left((\rho m_1+m_2)\frac{\sigma_{U_t,v_{t-\tau'}} - \frac{\sigma^2\sigma_{D^*,Y_{k-\tau'}}}{\sigma_Y^2}}{\sigma_{D^*}\sigma_Y\sqrt{1-\frac{\sigma_{D^*,Y_{k-\tau'}}^2}{\sigma_{D^*}^2\sigma_Y^2}}} + \rho m_2\frac{\sigma_{U_t,v_{t-\tau'-1}} - \frac{m_1\sigma^2\sigma_{D^*,Y_{k-\tau'}}^2}{\sigma_Y^2}}{\sigma_{D^*}\sigma_Y\sqrt{1-\frac{\sigma_{D^*,Y_{k-\tau'}}^2}{\sigma_{D^*}^2\sigma_Y^2}}}\right)\int_{-\infty}^{+\infty}\phi\left(\frac{y-g(x,d_{k-\tau'})}{\sigma_Y}\right)\phi(A_{xy})\frac{1-\Phi(A_{xy})}{\Phi(A_{xy})}dy\right].$$