Nonlinear Difference-in-Differences in Repeated Cross Sections with Continuous Treatments

Xavier D'Haultfœuille CREST
Stefan Hoderlein Boston College
Yuya Sasaki* Johns Hopkins

March 29, 2015

Abstract

This paper studies the identification of nonseparable models with continuous, endogenous regressors, also called treatments, using repeated cross sections. We show that several treatment effect parameters are identified under two assumptions on the effect of time, namely a weak stationarity condition on the distribution of unobservables, and time variation in the distribution of endogenous regressors. Other treatment effect parameters are set identified under curvature conditions, but without any functional form restrictions. This result is related to the difference-in-differences idea, but does neither impose additive time effects nor exogenously defined control groups. Furthermore, we investigate two extrapolation strategies that allow us to point identify the entire model: using monotonicity of the error term, or imposing a linear correlated random coefficient structure. Finally, we develop nonparametric estimators of the treatment effects and illustrate our results by studying the effect of mother's age on infants' birth weight.

Keywords: identification, repeated cross sections, nonlinear models, continuous treatment, random coefficients, endogeneity, difference-in-differences.

[*] Xavier D'Haultfœuille: Centre de Recherche en Économie et Statistique (CREST), 15 Boulevard Gabriel Péri 92254 Malakoff Cedex, email: `xavier.dhaultfoeuille@ensae.fr`. Stefan Hoderlein: Boston College, Department of Economics, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA, email: `stefan_hoderlein@yahoo.com`. Yuya Sasaki: Johns Hopkins University, Department of Economics, 440 Mergenthaler Hall, 3400 N Charles Street, Baltimore, MD 21218 USA, email: `sasaki@jhu.edu.` While not at the core of this paper, some elements are taken from the earlier draft "On the role of time in nonseparable panel data models" by Hoderlein and Sasaki, which is now retired. We have benefited from helpful comments from seminar participants at Boston College, GRIPS, and Chicago. The usual disclaimer applies.

1 Introduction

Using the time dimension to correct for the influence of correlated but time invariant unobservables has a long tradition in econometrics. When panel data are available, fixed effects or first differencing transformations are commonly used to purge the model from the influence of correlated unobserved heterogeneity. However, panel data sets are not always available. For many questions arising in applications, they simply do not exist or are of limited usefulness, because the covariates of interest are time invariant for a given individual. Prominent examples for these types of limitations include the returns to schooling and the evaluation of gender or racial wage gaps. Indeed, a similar issue arises in our application, where we study the effect of mother's age on the birth weight of their first child. Moreover, even if panels are available, they have several potential drawbacks. In particular, panels frequently suffer from nonrandom attrition, a phenomenon that is hard to control for[1], and they frequently cover only short time spans.

An alternative is to rely on repeated cross sections, i.e., a data set that covers the same population, but not necessarily the same individual, repeatedly. More formally, as econometricians we have access to the distributions $(F_{Y_1,X_1}, ..., F_{Y_T,X_T})$ of outcomes and explanatory variables. Contrary to the case of panel data, the joint distribution $F_{Y_1,X_1,...,Y_T,X_T}$ is not identified. Many prominent data sets are repeated cross sections, e.g., the FES in the UK, or the CEX in the US. We argue in this paper that many of the strong identification results obtained for panel data, e.g., concerning the correlated random coefficients panel data model (Chamberlain, 1982, Graham & Powell, 2012) have close correspondences in repeated cross section (RCS). In particular, in a RCS it is possible to obtain causal effects of a continuous variable of interest X_t on an outcome Y_t, while allowing for an unobservable A_t that is contemporaneously arbitrarily correlated with X_t, and does not even need to be time invariant. Note that the data structure precludes the use of past X_t for the same individual to construct control variables, as in Altonji & Matzkin (2005).

Formally, we consider as a general framework the single-equation structure of the form

$$Y_t = g_t(X_t, A_t) \qquad t = 1, ..., T \tag{1.1}$$

where $Y_t \in \mathbb{R}$ is the outcome, $X_t = (X_{1t}, ..., X_{kt}) \in \mathbb{R}^k$ is a set of explanatory variables and A_t are unobserved heterogeneous factors which may be correlated with X_t, causing endogeneity and invaildating causal inference. Observe that the structural function g_t is allowed to depend on the time period t, e.g., whether we are in a boom or an in a crisis in the business cycle.

[1] See nonetheless, among others, Hausman & Wise (1979), Kyriazidou (1997), Hirano et al. (2001), Das (2004), Bhattacharya (2008) or Sasaki (2013) for proposals on how to deal with endogenous attrition.

However, we do place restrictions on the time evolution of g_t by requiring that it is comprised of a monotonic transformation m_t and a time invariant base function g, i.e. $Y_t = m_t(g(X_t, A_t))$. This transformation extends typical additive time dummy specifications that are meant to capture macro shocks, and allows for the macro shocks to have different effects on different parts of the distribution of the "detrended" variable $\widetilde{Y}_t \equiv g(X_t, A_t)$. These macro shocks could for instance only affect individuals with high values of $\widetilde{Y}_t$.

In this setup, we focus on the identification of several parameters that take the form of average and quantile treatment effects. Generally, we establish that several treatment effect parameters are point identified. This requires in a first step to establish identification of the time dependent transformation function m_t. Based on this result, we establish point identification of a number of treatment effect parameters. Some parameters, like average partial effects at arbitrary positions $X_t = x$, however, are not covered by our point identification results. In those cases we derive bounds under additional conditions.

Finally, we clarify the identifying role of linear correlated random coefficient specifications considered for instance by Chamberlain (1982), Wooldridge (2003), Murtazashvili & Wooldridge (2008) or Graham & Powell (2012). We show that in our setting, such specifications allow to extrapolate and thus obtain point identification of average partial effects across the entire population.[2] A similar remark applies to assuming monotonicity in a scalar A_t - we are again able to point identify a structural model across the entire population.

Our key identifying assumptions are the following. First, we impose the stationarity condition on the error term that for almost all v and all $(s, t) \in \{1, ..., T\}^2$,

$$A_t | V_t = v \sim A_s | V_s = v,$$

where $V_t = \mathbf{F}_t(X_t) = (F_{X_{1t}}(X_{1t}), ..., F_{X_{kt}}(X_{kt}))$. We also need the source of exogenous variation in our model, time itself, to have some effect on the continuous explanatory variable. Since we are not following individuals over time, these variations should be at the distributional level. Namely, $\mathbf{F}_t \neq \mathbf{F}_s$ is necessary to obtain nontrivial identification results on the effect of X_T. We also require that there be at least one crossing point x^* between $\mathbf{F}_t$ and $\mathbf{F}_s$. Note that this is a testable property, analogous to a rank condition in IV.

To illustrate the content of these assumption, consider the textbook example where Y_t is unemployment duration, A_t is ability and X_t is unemployment benefits. Suppose also that, as in many countries, the benefit is tied to past income through a fixed monotonic tax schedule. Moreover, assume that this schedule is modified between two periods s and t. Finally, suppose

[2] We also show identification of treatment effects in a polynomial correlated random coefficient model provided that the order of the polynomial is less or equal to T. This result is related to Florens et al. (2008) except that time is discrete and does not act as a standard instrument here.

that the schedule remains unchanged for, say, low-income earners. In this setting, the second and third assumptions are likely to hold thanks to the nature of the change in the schedule: the distribution of unemployment benefits is likely to change, but the cdf may remain stable at the bottom of the distribution. The first assumption requires that people at a given rank of the income distribution (e.g., earning the median income) should have the same distribution of ability over time. This assumption holds if, basically, the unconditional distribution of ability remains stable between the two periods, and the way ability affects our ranking in terms of wages is also unchanged. Importantly, this assumption may still hold even if income itself is substantially modified between the two periods because of, say, macro shocks.

The main idea behind our identification result is to first isolate the effect of time, in our notation the monotonic function m_t. This is achieved by realizing that we can construct a control group at the crossing point x^*. Since X_t is time invariant at this point, and the associated rank and hence the distribution of unobservables A_t does not change either, we conclude that any effect on the outcome distribution must have been generated by time itself. The combination of this insight with the monotonicity assumption allows one to recover m_t. The next step is to purge Y_t from the influence of time, thus removing from the treated groups the effect that time alone had on them. The new variable, $\tilde{Y}_t$ can now be used to point or set identify any of the various causal effects parameters described below. At this stage, the key insight is that under our assumptions, in particular time invariance of the conditional distribution of A_t, time plays the role of an instrument. We can therefore use exogenous variation in the distribution of X_t over to time to identify causal effects.

Related Literature: Our setup is most related to the difference-in-difference (DiD) framework introduced by Ashenfelter & Card (1985). In its standard version, the difference-in-difference method also works with repeated cross sections, though it applies to binary treatments, and assumes a linear fixed coefficients structure. The idea is that there are two well-defined groups, namely the control and treatment group, and while none of them are treated at period 1, the treatment group becomes treated at period 2. If the effect of time is the same for both groups ("the common trend assumption"), it can be identified using the control group. The effect of the treatment is then obtained using the treatment group and the detrended variable.[3] The broad identification strategy we develop here is similar, though there are important differences, most notably that we consider a continuous endogenous regressor (treatment). But this is by no means the only important difference. Other crucial differences include the following: First, our model is fully nonlinear in both the continuous regressor and the potentially high dimensional unobservables. Second, the effect of time in particular is al-

[3] Extensions of this strategy to account for time effects that depend on covariates are considered by Heckman et al. (1997) and Abadie (2005).

lowed to be nonadditive in our model. This makes our model closer to the changes-in-changes model developed by Athey & Imbens (2006) for binary treatments (see also de Chaisemartin & D'Haultfoeuille, 2014, for an extension to fuzzy settings). Different, however, from the entire literature, including Athey & Imbens (2006), is that the control group is data-dependent in our context, whereas it is defined ex ante in the DiD framework.

As already discussed, we use exogenous variations of X_t due to time. This idea has already been put forward in the literature on repeated cross sections. Previous contributions include Deaton (1985), Moffitt (1993), Verbeek & Nijman (1992, 1993), Verbeek (1996), Collado (1997), McKenzie (2004) and Devereux (2007). Compared to this literature, our contribution is twofold. First, we dispense with the common linear or parametric framework that they consider. Our model is nonlinear and nonparametric, and allows for high dimensional heterogeneity. Second, our identification strategy does not exclude time from affecting the outcome directly. A last important difference between our work and the classical literature on repeated cross sections is the focus. While we are concerned with contemporaneous causal effects, the literature usually focuses on the identification of the joint distribution of $(Y_1, X_1, ..., Y_T, X_T)$, or features of it, from the marginal distributions of (Y_1, X_1),..., (Y_T, X_T), usually to derive dynamic effect, see Moffitt & Ridder (2007) for a survey.

Our work is also related to general work on high dimensional heterogeneity in panel and cross section data, starting with the seminal work by Chamberlain (1982, 1984). Important references in the class of panel data models include Arellano & Bonhomme (2012), Altonji & Matzkin (2005), Graham & Powell (2012), Hoderlein & White (2012) and Chernozhukov et al. (2013). All of these papers consider special cases or similar structures as defined in Equation (1.1), but they do not allow the structural function to depend on time. Instead of our time invariance assumption, all of these references assume, for $(s, t) \in \{1, ..., T\}^2$ and almost all $(x_1, ..., x_T)$

$$A_t | X_1 = x_1, ..., X_T = x_T \sim A_s | X_1 = x_1, ..., X_T = x_T.$$

This condition neither nests nor is nested in our assumption, as we argue below. In addition, Altonji & Matzkin (2005) assume an exchangeability condition that allows to construct a control function that makes A_t conditionally independent of X_t, while Graham & Powell (2012) assume a linear random coefficients structure, arguably a crucial special case that we will also analyze in detail. Evdokimov (2011) imposes the error term to be scalar and to have a monotonic effect. Under monotonicity, we also obtain full identification with only repeated cross sections over two time periods, as opposed to panel data with three periods in his case. On the other hand, we obtain our result under time invariance conditions that are not imposed in his setting. Finally, many of the treatment parameters we are considering appear in these references, but have also figured prominently in the cross section literature, see Imbens & Newey (2009), Schennach et al. (2012), or Hoderlein & Mammen (2007).

Structure of the Paper In section 2, we introduce the model formally, including all major assumptions and the parameters of interest, and discuss them thoroughly. In the third section, we present the main identification result. In the fourth section, we discuss two extrapolation strategies. We consider a linear correlated random coefficient structure and a model where g depends monotonically on a scalar A_t. We show that in both cases, these restrictions yield point identification of the structural effect across the entire population. In the fifth section, we consider estimators of the average and quantile treatment effects that were shown to be identified in Section 3, and show their asymptotic normality. Finally, in the sixth section, we apply our methodology to the effect of maternal age on birth weight of the first child. This is typically an example where maternal age is endogenous, an instrument might be difficult to find and panel data are useless, because the maternal age at the first birth does not vary within individuals.

2 The Model and Formal Assumptions

In this section, we formally introduce the model and the main assumptions. Since the model is nonparametric and heterogeneous, the parameters of interest are not obvious. We start out by formally introducing these parameters. We then proceed to present and discuss the main assumptions we employ.

2.1 Parameters of interest

We are especially interested in the following average and quantile treatment on the treated effects:

$$\begin{aligned}
\Delta^{ATT}(x,x') &\equiv E\left[g_T(x',A_T)-g_T(x,A_T)|X_T=x\right],\\
\Delta_j^{AME}(x) &\equiv E\left[\frac{\partial g_T}{\partial x_j}(x,A_T)|X_T=x\right],\\
\Delta^{QTT}(p,x,x') &\equiv F^{-1}_{g_T(x',A_T)|X_T}(p|x)-F^{-1}_{g_T(x,A_T)|X_T}(p|x),\\
\Delta_j^{QME}(p,x) &\equiv \frac{\partial F^{-1}_{g_T(x',A_T)|X_T}(p|x)}{\partial x'_j}|x'=x,
\end{aligned}$$

for any $x=(x_1,...,x_k)$ and $x'=(x'_1,...,x'_k)$ in the support of X_T and $j\in\{1,...,k\}$. These parameters correspond to the effect of exogenous shifts of X_T on Y_T. The first two effects are average effects, while the latter two effects are their quantile analogs. The former two effects are related to treatment effects on the treated in that they provide averages over causal effects for a subpopulation with treatment intensity $X_T = x$. To understand this better, consider the first parameter of interest, $\Delta^{ATT}(x, x')$. To fix ideas, think of A_t as ability in period t, and X_t as schooling. Obviously, we would believe ability to be heterogeneously distributed across the population, as well as contemporaneously correlated with schooling. For an individual with ability level $A_t = a$ in period t, the effect of changing exogenously the amount of schooling she receives from x to x' would be

$$g_T(x', a) - g_T(x, a).$$

A very natural parameter for a decision maker to be interested in is some form of average across a heterogeneous population. Since X_t and A_t are correlated, the natural question is which type of average one would like to consider. In this paper, we advocate the use of $F_{A_t|X_t}$ as a weighting scheme. The reason is simple, and easily understood in our example. Suppose $X_t = x$ corresponds to 4 years of university, and the question is to determine effect of the introduction of ninth semester (i.e., $x' = x + 0.5$) as a policy measure. In this case it does not make sense to weigh with the unconditional distribution of A_t as there are many individuals, presumably frequently with lower levels of ability, who never complete four years of college. Hence, it is natural to average the causal effect with the weighting scheme $F_{A_t|X_t}(.|x)$, since this is really the subpopulation primarily affected by the policy measure of changing X_t exogenously from x to x'. This corresponds, in period T, to the effect

$$\int (g_T(x', a) - g_T(x, a)) F_{A_T|X_T}(da; x) = E\left[g_T(x', A_T) - g_T(x, A_T)|X_T = x\right].$$

Very analogous arguments apply to the marginal effect $\Delta_j^{AME}(x)$. The analysis of this effect has a long history in econometrics, starting with the seminal work by Chamberlain (1982, 1984), who called this marginal effect the local average response. Important references are Altonji & Matzkin (2005), Wooldridge (2005), Graham & Powell (2012), Hoderlein & White (2012) and Chernozhukov et al. (2013) in the panel data literature, and Hoderlein & Mammen (2007), Imbens & Newey (2009), Schennach et al. (2012) in the IV literature.

An interesting consequence of obtaining $\Delta_j^{AME}(x)$ is that

$$\int \Delta_j^{AME}(x) f_X(x) dx = E\left[\frac{\partial g_T}{\partial x}(X_T, A_T)\right]$$

provides the overall average partial effect (see Chamberlain, 1984). This parameter corresponds to the thought experiment of increasing schooling marginally across the entire population, and averaging the effect across the various levels of eduction and ability.

The quantile effects $\Delta^{QTT}(p, x, x')$ and $\Delta_j^{QME}(p, x)$ provide causal effects on the counterfactual marginal distributions. This is different from obtaining the distribution of causal effects, but both effects are widely analyzed, see Abadie et al. (2002) and Chernozhukov et al. (2013), amongst many others.

Finally, we consider all effects for period T as we believe there are the most natural to compute in general. However, the result of Theorem 1 below implies that we can actually identify similar effects at any date.

2.2 Assumptions

The broad idea for identifying these parameters is to restrict the way time affects both observed and unobserved variables. More specifically, we impose hereafter three restrictions. The first is a stationarity condition on the observed and unobserved determinants of the outcome. The second restricts the way time is affecting the outcome itself. The third restricts the way the distribution of X_t moves over time. We discuss them in turns, using the notations $\mathbf{F}_t(x) = (F_{X_{1t}}(x_1), ..., F_{X_{kt}}(x_k))$ for any $x = (x_1, ..., x_k)$, $V_t = \mathbf{F}_t(X_t)$ and $\mathcal{V}_t = \text{supp}(V_t)$. The first assumption is:

Assumption 1. *The distribution of X_t is absolutely continuous with a convex support, and for all $(s,t) \in \{1, ..., T\}^2$ and almost all $v \in \mathcal{V}_T$,*

$$A_t|V_t = v \sim A_s|V_s = v.$$

To fix ideas, consider the returns to education example, and suppose that A_t comprises an ability term correlated with education, and an idiosyncratic term independent of ability and education. Assumption 1 means in this context that the distribution of ability conditional on a given rank in the distribution of education remains stable over time.

This stationarity condition is different from the condition

$$A_s|X_1, ..., X_T \sim A_t|X_1, ..., X_T, \tag{2.1}$$

commonly assumed in panel data (see, e.g., Manski, 1987, Honore, 1992, Graham & Powell, 2012 and Chernozhukov et al., 2013). To understand the differences between the two, consider two polar cases. In the first, endogeneity stems from contemporanous simultaneity, as is often the case with variables that are jointly determined, while $(A_t, V_t)_{t=1...T}$ are i.i.d. across time. If so, Assumption 1 is satisfied. On the other hand, (2.1) does not hold, unless A_t is independent of V_t, because the distribution of A_s conditional on $(X_1, ..., X_T)$ is a function of X_s only, i.e., $f_{A_s|X_1,...,X_T}(a|x_1, ..., x_T) = f_{A_s|X_s}(a|x_s)$, while the conditional distribution A_t is a function of X_t only, and they do generally not coincide if $x_s \neq x_t$. Assuming (A_s, V_s) independent of (A_t, V_t) is of course often unrealistic, but the same conclusion would hold with, say, a vector autoregressive structure. In the second case, $A_t = (A, U_t)$ where A is a fixed effect potentially correlated with $X_1, ..., X_T$ and $(U_t)_{t=1}^T$ are i.i.d. idiosyncratic shocks that are independent of $(A, X_1, ..., X_T)$. In this case, the condition (2.1) is always satisfied. On the other hand, Assumption 1 holds only under a special correlation structure between A and $(X_1, ..., X_T)$: $A|V_t = v \sim A|V_s = v$, which for instance imposes $Cov(A, V_t) = Cov(A, V_s),\ s \neq t$. While this still allows for arbitrary contemporaneous correlation between A and V_t, respectively V_s, it limits the time evolution of this covariance. It is this type of time invariance of the correlated unobservables that an applied researcher has to check, and, if adopted, defend.

This time invariance is somewhat mitigated by the fact that we allow for the function g_t to vary with time. To see this, let us first state the extent to which we allow for time dependence formally:

Assumption 2. *For all $t \in \{1, ..., T\}$, $g_t = m_t \circ g$, where m_t is strictly increasing. Without loss of generality, we let $m_T(y) = y$ for all $y \in supp(Y_T)$.*

Assumption 2 generalizes the standard translation model $m_t(u) = \delta_t + u$ to allow for heterogeneous effects of time. Allowing for the effect of time on the structural relationship seems quite important. For instance, in the returns to education example, the effect of education on wage may vary according to the state of the business cycle. Our specification allows for these macroeconomic shocks to have heterogeneous effects on individuals. To understand the extent to which is the case, think of $\widetilde{Y}_t = g(X_t, A_t)$ as the latent, long run wage which is free of seasonal or business cycle effects. Then, our specification allows in particular for the effect of an economic downturn on lower $\widetilde{Y}$ individuals to be stronger (or less strong). But it still places restriction on the way time affects the outcome. In particular, while allowing for contractions and expansions of the wage distribution, we cannot assume that the effect of time is such that the ordering of any two individuals is reversed if neither their observables nor unobservables change over time.

On the positive side, this assumption allows to overcome some of the restrictiveness of the fact that $\mathrm{Cov}(A, V_t) = \mathrm{Cov}(A, V_s)$, $s \neq t$. To understand this, suppose that the structural model is given by $Y_t = \delta_t(\alpha A + \beta h_1(X_t) + \gamma A h_2(X_t)) = \alpha_t A + \beta_t k_1(V_t) + \gamma_t A k_2(V_t)$, where $h_j, k_j, j = 1, 2$ are increasing transformations, and $\gamma_t = \delta_t \gamma$. This specification allows for some interaction effect between between A and V_t, with a time heterogeneous impact on Y_t. In the example of returns to eduction, even if the correlation between ranks in the education distribution and unobserved ability is time invariant, the effect of having high education combined with high ability could be higher in, say, an economic upswing.

Finally, our last assumption concerns the independent variation that identifies the model. Given the highly nonlinear setup we are considering, it comes in the form of a distributional assumption. It allows for the construction of a "control group" that identifies the effect of time on the outcome (the function m_t), analogously to the DiD literature.

Assumption 3. *For all $t \in \{1, ..., T\}$, there exists $x_t^* \in \mathbb{R}^k$ such that $\mathbf{F}_t(x_t^*) = \mathbf{F}_T(x_t^*) \in (0,1)^k$.*

Several remarks are in order: first, Assumption 3 is directly testable in the data. It allows for any change in the distribution of X_t, provided that there is a crossing between the cumulative distribution function of X_{jT} and X_{jt}, for all $j \in \{1, ..., k\}$ and $t < T$.[4] Roughly speaking, this means that time has an heterogenous effect on the distribution of X_t. It fails to hold in the pure location model $X_t = \gamma_t + B_t$, where the distribution of B_t is stationary with support $\mathbb{R}^k$. On the other hand, it holds in the location-scale model $X_t = \gamma_t + \Sigma_t B_t$ if Σ_t is diagonal with diagonal terms σ_{jt} that are distinct at each time period. In such a case x_t^* is unique and satisfies

$$x_t^* = \left(\frac{\gamma_t - \gamma_T}{\sigma_{1T} - \sigma_{1t}}, ..., \frac{\gamma_t - \gamma_T}{\sigma_{kT} - \sigma_{kt}} \right).$$

Note that if $\mathbf{F}_t$ remains constant with t, Assumption 3 is satisfied but we identify only trivial parameters such as $\Delta^{ATT}(x, x)$. Nontrivial parameters are identified only when $\mathbf{F}_t$ changes with t. This contrasts with the kind of variations in the individual value of the treatment over time that is typically required with panel data, the fixed effects absorbing any variable that is constant across time. The distribution of X_t can move over time even if X_t is constant for each individual, provided new generations are involved at date t compared to date s. Our application below is an example of such a situation. On the other hand, compared to panel data, we do not identify anything, apart from the time effect m_t, when the treatment changes at an individual level but the distribution of X_t remains constant over time. This is one different aspect of our identification strategy from panel data based strategies.

Summarizing our model, the overall idea is that an exogenous shock, such as a policy change, affects the distribution of X, so that $\mathbf{F}_t \neq \mathbf{F}_T$, while units at the same rank of X_t remain comparable in the sense that their unobservable characteristics have the same distribution, $A_t | V_t = v \sim A_T | V_T = v$. There may be some aggregate shocks on the outcome Y_t, still, and we allow these shocks to affect units in a nonadditive way. On the other hand, we restrict them to affect individuals solely through the index $g(X_t, A_t)$. In this sense, our model considers the effect of time in a similar way as Athey & Imbens (2006) or de Chaisemartin & D'Haultfoeuille (2014),

[4] We assume for simplicity crossings between X_T and the other cdf, but actually, $T - 1$ crossings are fine provided that we can "relate" them to each other, for instance if the cdf of X_t crosses the one of X_{t+1} for $1 \leq t < T$. With only one crossing between $\mathbf{F}_s$ and $\mathbf{F}_t$, we can still identify the effect of time between these two periods ($m_t \circ m_s^{-1}$) and then identify some treatment effects.

where the outcome is a function of time and a scalar index. Such a restriction implies that two individuals with the same initial outcome and for whom both X_t and A_t remain constant would be affected by time in the same way, even if they have different X's and unobserved characteristics A.

3 Identification results

3.1 Point Identified Effects

The first idea that drives our results is that the effect of time can be obtained using individuals for which $X_T = X_t = x_t^*$. These individuals, though possibly different across time periods, have under Assumption 1 the same distribution of unobservables and the same value of the treatment. For them, differences between Y_T and Y_t can only stem from the effect of time itself. This is the reason why we call them the "control group". Formally,

$$\begin{aligned}
P\left(Y_T \leq y | X_T = x_t^*\right) &\overset{A.2}{=} P\left(g(x_t^*, A_T) \leq y | V_T = \mathbf{F}_T(x_t^*)\right) \\
&\overset{A.1}{=} P\left(g(x_t^*, A_t) \leq y | V_t = \mathbf{F}_T(x_t^*)\right) \\
&\overset{A.3}{=} P\left(g(x_t^*, A_t) \leq y | V_t = \mathbf{F}_t(x_t^*)\right) \\
&\overset{A.2}{=} P\left(m_t \circ g(x_t^*, A_t) \leq m_t(y) | X_t = x_t^*\right) \\
&\overset{A.2}{=} P\left(Y_t \leq m_t(y) | X_t = x_t^*\right),
\end{aligned}$$

the first equality following because m_T is the identity function. As a result, m_t is identified by

$$m_t(y) = F^{-1}_{Y_t|X_t}\left[F_{Y_T|X_T}(y|x_t^*)|x_t^*\right]. \tag{3.1}$$

This transformation is similar in spirit to a transformation in Athey & Imbens (2006). However, it differs in the crucial aspect that we are not exogenously given a treatment and, in particular, a control group, but endogenously obtain the control group through our assumptions. We conjecture that there are more general ways of constructing a control group, in particular if there are more than two time periods available, but we leave this issue for future research.

Beyond the identification of m_t, (3.1) reveals that the model is testable if there are more than one crossing point between $\mathbf{F}_t$ and $\mathbf{F}_T$, say x_t^* and x_t^{**}. In such a case, we have, for all y,

$$F^{-1}_{Y_t|X_t}\left[F_{Y_T|X_T}(y|x_t^*)|x_t^*\right] = F^{-1}_{Y_t|X_t}\left[F_{Y_T|X_T}(y|x_t^{**})|x_t^{**}\right],$$

which can be tested in the data.

Next, we consider the identification of the treatment effects introduced in Subsection 2.1. Let us consider the transformed outcome $\widetilde{Y}_t = m_t^{-1}(Y_t)$, which is purged of the influence of time in the sense that by Assumption 1, time has no direct effect on $\widetilde{Y}_t$. In other words, variations in X_t provided by time are now exogenous in the sense that they do not affect the distribution of unobservables. Time can thus be considered to act like an instrument. As already mentioned, implicitly similar ideas have been used in the panel data literature, though using different and non-nested assumptions (see, e.g., Manski, 1987, Honore, 1992, Graham & Powell, 2012, Hoderlein & White, 2012, Chernozhukov et al., 2013), which all consider the effect of time variations on X_t and Y_t.

To proceed with the identification of our model, let $q_t(x)$ denote the value of X_t (say, income in period t) for an individual at the same rank as another individual whose period T income is $X_T = x$, $x \neq x_t^*$. Formally, let $q_{jt} = F_{X_{jt}}^{-1} \circ F_{X_{jT}}$ for $j \in \{1, ..., k\}$ and

$$q_t(x) = (q_{1t}(x_1), ..., q_{kt}(x_k)).$$

We have then that

$$\begin{aligned} E\left[\widetilde{Y}_t | X_t = q_t(x)\right] &= E\left[g(q_t(x), A_t) | V_t = \mathbf{F}_T(x)\right] \\ &\overset{A.1}{=} E\left[g(q_t(x), A_T) | V_T = \mathbf{F}_T(x)\right] \\ &= E\left[g(q_t(x), A_T) | X_T = x\right]. \end{aligned}$$

The latter is the mean counterfactual outcome at period T for individuals with $X_T = x$ if X_T was moved exogenously to $q_t(x)$. We can therefore identify $\Delta^{ATT}(x, q_t(x))$, the average effect of moving X_T from their initial value x to $q_t(x)$, by

$$\begin{aligned} \Delta^{ATT}(x, q_t(x)) &\overset{A.2}{=} E\left[g(q_t(x), A_T) - g(x, A_T) | X_T = x\right] \\ &= E\left[\widetilde{Y}_t | X_t = q_t(x)\right] - E\left[\widetilde{Y}_T | X_T = x\right], \end{aligned}$$

where the first equality comes from the normalization in assumption 2, implying that $g_T = m_T \circ g = g$, and hence $ATT(x, x') \equiv E\left[g_T(x', A_T) - g_T(x, A_T) | X_T = x\right] = E\left[g(x', A_T) - g(x, A_T) | X_T = x\right].$ This means that we can obtain $\Delta^{ATT}(x, x')$ for any pair $x, x' = q_t(x)$, and $x \neq x_t^*$. Note that we cannot point identify $\Delta^{ATT}(x, x')$ for $x' \neq q_t(x)$, but we will show in the following subsection that we can at least set identify these parameters under plausible curvature restrictions. Also, we cannot identify any effect $\Delta^{ATT}(\xi, \xi')$ with $\xi' \neq \xi$ if $\mathbf{F}_t(\xi) - \mathbf{F}_T(\xi) = 0$. As mentioned above, we need the distribution of X_t to change with time.

When X_T is multivariate, it may be difficult to interpret $\Delta^{ATT}(x, q_t(x))$ because it corresponds to the effect of a change of potentially all components of X_T. However, still using the crossing points, we can identify some partial effects. To see this, consider ${}_j\widetilde{x}_t = (x_{1t}^*, ..., x_{j-1t}^*, x_j, x_{j+1t}^*, ..., x_{kt}^*)$ for some $x_j \neq x_{jt}^*$. Then, by definition of x_t^*,

$$q_t({}_j\widetilde{x}_t) = \left(x_{1t}^*, ..., x_{j-1t}^*, q_{jt}(x_j), x_{j+1t}^*, ..., x_{kt}^*\right).$$

This means that $\Delta^{ATT}({}_j\widetilde{x}_t, q_t({}_j\widetilde{x}_t))$ corresponds to the average partial effect of exogenously shifting X_{jT} from x_j to $q_{jt}(x_j)$.

For people at the crossing points x_t^*, we do not learn anything from the above reasoning, because $\Delta^{ATT}(x_t^*, q_t(x_t^*)) = \Delta^{ATT}(x_t^*, x_t^*) = 0$ by construction. On the other hand, under mild regularity condition (see Assumption 4 below), we can identify the average marginal effects for this population provided that q_t differs from the identity function in the neighborhood of x_t^*. The intuition behind the latter result is that we can find values x close to x_t^* and such that $q_t(x) - x$ is close to zero, but not exactly zero. Then, if X_t is univariate (the multivariate case can be handled similarly),

$$\frac{g(q_t(x), A_T) - g(x, A_T)}{q_t(x) - x} \simeq \frac{\partial g}{\partial x}(x_t^*, A_{1t}). \tag{3.2}$$

Moreover, if the conditional distribution of A_T is regular, conditioning on $X_T = x$ becomes the same as conditioning on $X_T = x_t^*$, so that

$$\frac{\Delta^{ATT}(x)}{q_t(x) - x} \simeq \Delta_1^{AME}(x)(x_t^*).$$

Formally, identification of the marginal effect is achieved on the set $\mathcal{X}_0$ defined by

$$\begin{aligned}\mathcal{X}_0 \quad = \quad & \Big\{ x \in \mathbb{R}^k : \exists (t, (x_n)_{n\in\mathbb{N}}) \in \{1, ..., T-1\} \times \left(\mathbb{R}^k\right)^{\mathbb{N}} : \; q_t(x) = x, \lim_{n\to\infty} x_n = x \\ & \quad \text{and } q_{jt}(x_{jn}) \neq x_{jn} \text{ for all } j = 1, ..., k \Big\}.\end{aligned}$$

$\mathcal{X}_0$ is the union of fixed points of $q_2, ..., q_T$, once we exclude points x^* such that in their neighborhood, $q_{jt}(x_j) = x_j$ for some $j \in \{1, ..., k\}$. See Figure 1 for an illustration in the univariate case. To make the preceding argument rigorous, the following technical conditions are also required.

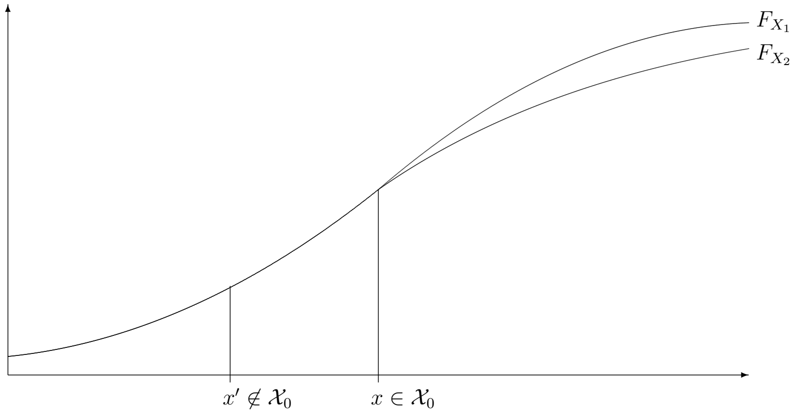


Figure 1: Example of points belonging or not to $\mathcal{X}_0$

Assumption 4. *(Regularity conditions) For all points $x_t^* \in \mathcal{X}_0$, there exists a neighborhood $\mathcal{N}$ such that:*

(i) almost surely, $x \mapsto g(x, A_T)$ is continuously differentiable on $\mathcal{N}$.

(ii) the distribution of A_T conditional on X_T is continuous with respect to the Lebesgue measure and $x \mapsto f_{A_T|X_T}(a|x)$ is continuous at x_t^*.

(iii) For all $j \in \{1, ..., k\}$, $\int |\sup_{x' \in \mathcal{N}} \partial g / \partial x_j(x', a)| \left|\sup_{x' \in \mathcal{N}} f_{A_T|X_T}(a|x')\right| da < \infty$.

(iv) For all $x \in \mathcal{N}$ and $j \in \{1, ..., k\}$, $x' \mapsto F^{-1}_{g(x', A_T)|X_T}(p|x)$ is differentiable at x_t^*. $(x, x') \mapsto \frac{\partial F^{-1}_{g(x'', A_T)|X_T}(p|x)}{\partial x''_j} | x'' = x'$ is continuous on $\mathcal{N}^2$.

Finally, we can apply the same reasoning to the quantile function. We can recover $F^{-1}_{g_T(q_t(x), A_T)|X_T}(p|x)$ by $F^{-1}_{\widetilde{Y}_t|X_t}(p|q_t(x))$, which implies that $\Delta^{QTT}(p, x, q_t(x))$ is identified. We also identify $\Delta_j^{QME}(p, x_t^*)$ by a similar argument as above.

Theorem 1 summarizes all findings of this section:

Theorem 1. *Under Assumptions 1-3, we identify, for all $x \in supp(X_T)$, $p \in (0, 1)$ and $t \in \{1, ..., T-1\}$, the functions m_t and the average and quantile treatment effects $\Delta^{ATT}(x, q_t(x))$ and $\Delta^{QTT}(p, x, q_t(x))$. If Assumption 4 holds as well, we also identify $\Delta_j^{AME}(x)(x_t^*)$ and $\Delta_j^{QME}(p, x_t^*)$ for all $x_t^* \in \mathcal{X}_0$ and all $j \in \{1, ..., k\}$.*

3.2 Partial Identification of Other Treatment Effects

Theorem 1 implies that we can point identify some but not all average treatment effects $\Delta^{ATT}(x, x')$. Similarly, we point identify the average marginal effects only at some particular points. We show in this subsection that with three or more periods of observation and an univariate X_t, we can get bounds for many other points under a weak local curvature condition.[5] Let us consider the average marginal effect for instance. The idea is that if $g(., A_t)$ is locally concave (say) and $q_t(x) < x$, then $\frac{g(q_t(x), A_T) - g(x, A_T)}{q_t(x) - x}$ is an upper bound of $\frac{\partial g}{\partial x}(x, A_T)$. Similarly, if $q_s(x) > x$, then $\frac{g(q_s(x), A_T) - g(x, A_T)}{q_s(x) - x}$ is a lower bound for $\frac{\partial g}{\partial x}(x, A_T)$ (see Figure 2). By integrating over A_T, we can therefore bound $\Delta_j^{AME}(x)$ by some appropriate $\Delta^{ATT}(x, q_t(x))/(q_t(x) - x)$. The same idea can be used to obtain bounds $\Delta^{ATT}(x, x')$ for $x' \notin \{q_t(x), t = 2...T\}$.

The above argument works even if we do not know a priori whether g is concave or convex. Using the minimum and the maximum of the local discrete treatment effect will be sufficient to obtain bounds, provided that g is locally concave or locally convex around x. We therefore adopt henceforth the following definition.

[5] The reasoning developed here also works when X_t is multivariate, but only applies to $\Delta^{ATT}({}_j\widetilde{x}_t^*, {}_j\widetilde{x}_t^{*\prime})$, where ${}_j\widetilde{x}_t^*$ is defined as before and ${}_j\widetilde{x}_t^{*\prime}$ is similar to ${}_j\widetilde{x}_t^*$, except that its j-th component is x'_j instead of x_j.

Definition 1. *g is locally concave or convex on $[\widetilde{x}, \widetilde{x}']$ if $x \mapsto g(x, A_t)$ is twice differentiable and*

$$\frac{\partial^2 g}{\partial x^2}(x, A_t) \leq 0 \; \forall x \in [\widetilde{x}, \widetilde{x}'] \textit{ a.s. or } \frac{\partial^2 g}{\partial x^2}(x, A_t) \geq 0 \; \forall x \in [\widetilde{x}, \widetilde{x}'] \textit{ a.s.}$$

Let us introduce, for all $(x, x') \in \text{supp}(X_T)$, $(\underline{x}_T(x'), \overline{x}_T(x'))$ defined by

$$\begin{aligned} \underline{x}_T(x') &= \max\{q_t(x), t \in \{1, ..., T-1\} : \; q_t(x) \neq x \text{ and } q_t(x) < x'\}, \\ \overline{x}_T(x') &= \min\{q_t(x), t \in \{1, ..., T-1\} : \; q_t(x) \neq x \text{ and } q_t(x) > x'\}. \end{aligned}$$

If the sets are empty we let $\underline{x}_T(x') = -\infty$ and $\overline{x}_T(x') = +\infty$.

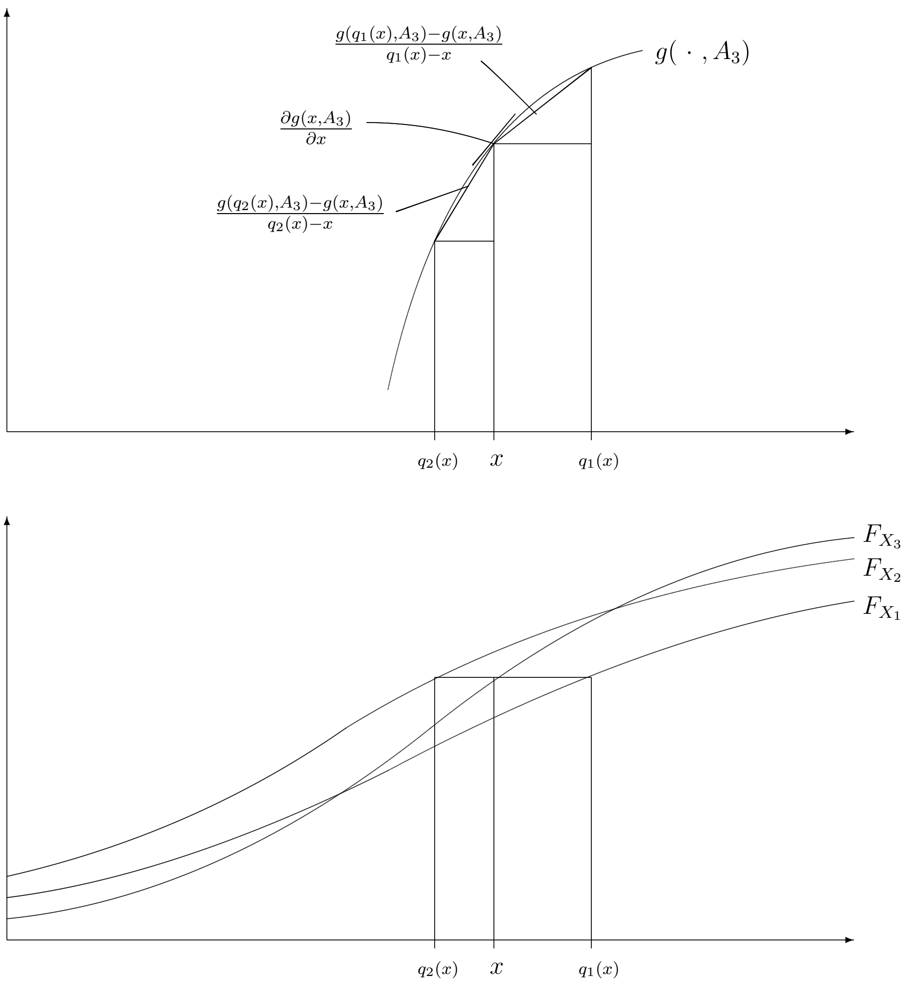

$\frac{g(q_1(x),A_3)-g(x,A_3)}{q_1(x)-x}$
$g(\,\cdot\,, A_3)$
$\frac{\partial g(x,A_3)}{\partial x}$
$\frac{g(q_2(x),A_3)-g(x,A_3)}{q_2(x)-x}$
$q_2(x)$
x
$q_1(x)$
F_{X_3}
F_{X_2}
F_{X_1}
$q_2(x)$
x
$q_1(x)$

Figure 2: Bounds under the local curvature condition

Theorem 2. *If $k = 1$ and under Assumptions 1-3,*

- *for any $x < x'$, if g is locally concave or convex on $[\min(x, \underline{x}_T(x')), \overline{x}_T(x')]$, then*

$$\begin{aligned} & (x'-x)\min\left\{\frac{\Delta^{ATT}(x,\underline{x}_T(x'))}{\underline{x}_T(x')-x}, \frac{\Delta^{ATT}(x,\overline{x}_T(x'))}{\overline{x}_T(x')-x}\right\} \leq \Delta^{ATT}(x,x') \\ \leq \quad & (x'-x)\max\left\{\frac{\Delta^{ATT}(x,\underline{x}_T(x'))}{\underline{x}_T(x')-x}, \frac{\Delta^{ATT}(x,\overline{x}_T(x'))}{\overline{x}_T(x')-x}\right\}. \end{aligned}$$

- *If g is locally concave or convex on $[\underline{x}_T(x), \overline{x}_T(x)]$, then*

$$\min\left\{\frac{\Delta^{ATT}(x,\underline{x}_T(x))}{\underline{x}_T(x)-x}, \frac{\Delta^{ATT}(x,\overline{x}_T(x))}{\overline{x}_T(x)-x}\right\} \leq \Delta_1^{AME}(x) \leq \max\left\{\frac{\Delta^{ATT}(x,\underline{x}_T(x))}{\underline{x}_T(x)-x}, \frac{\Delta^{ATT}(x,\overline{x}_T(x))}{\overline{x}_T(x)-x}\right\}.$$

where the bounds are understood to be infinite when either $\underline{x}_T(x') = -\infty$ or $\overline{x}_T(x') = +\infty$ (whether $x' > x$ or $x' = x$).

Both bounds are finite provided that there exists t, t' such that $q_t(x) < x < q_{1t'}(x)$, which implies that $T \geq 3$. More generally, the bounds improve with T, because $(\underline{x}_T(x'))_{T\in\mathbb{N}}$ and $(\overline{x}_T(x'))_{T\in\mathbb{N}}$ are by construction increasing and decreasing, respectively. The local curvature condition becomes less and less restrictive as T increases, because the interval on which g has to satisfy this condition decreases. It seems particularly credible, if $q_t(x) \mapsto \Delta(x, q_t(x))/(q_t(x) - x)$ is monotonic, because such a pattern is implied by global concavity or global convexity.

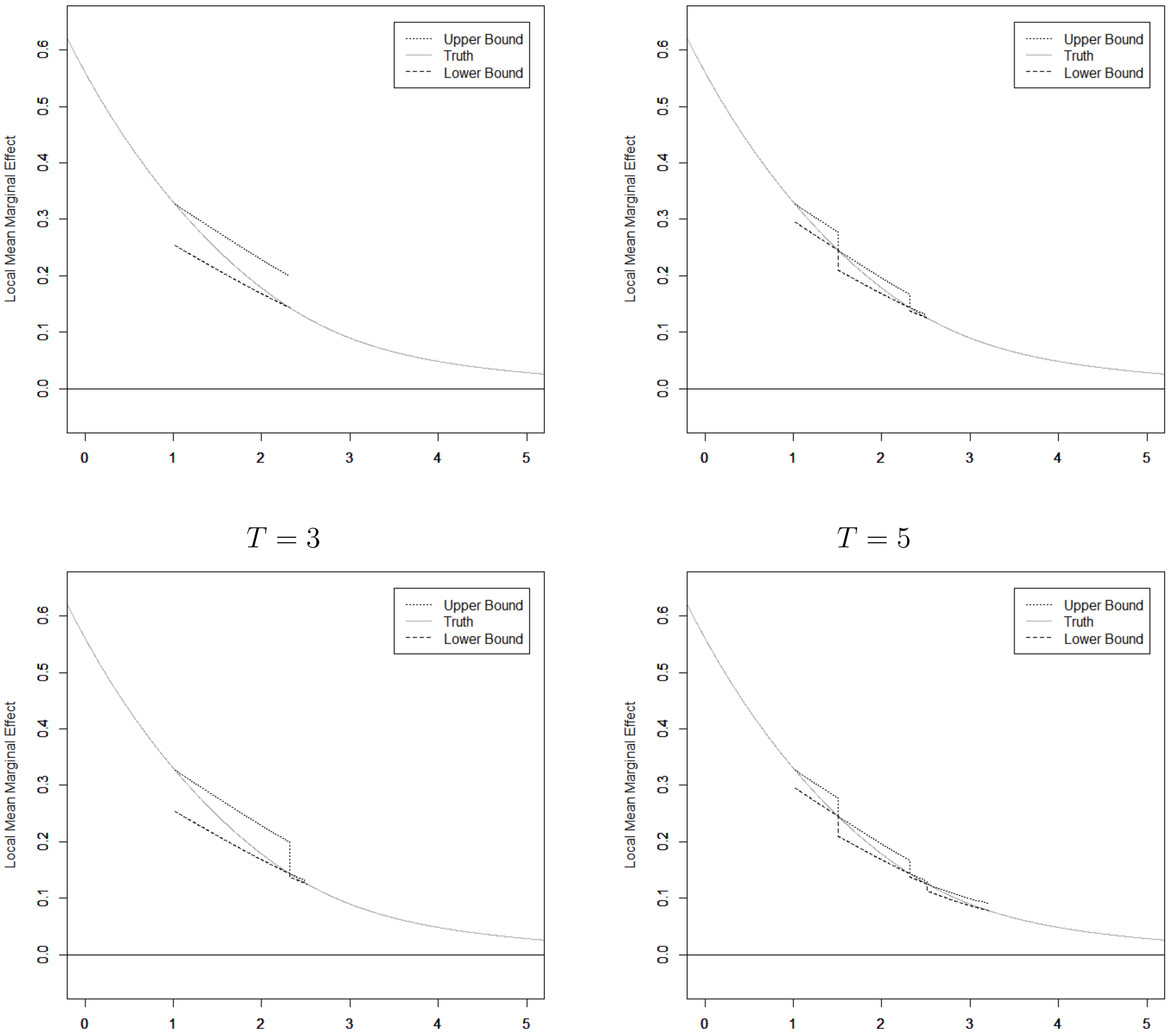
To illustrate Theorem 2, we consider the following example:

$$\begin{aligned} Y_t &= 1 - \exp(-0.5(\delta_t + X_t + A_t)) \\ X_t &= \mu_t + \sigma_t \Phi^{-1}(V_t), \end{aligned}$$

where $V_t \sim U[0,1]$ and $A_t|V_t \sim \mathcal{N}(V_t, 1)$. We also suppose that

$$\begin{aligned} \mu_T = 2.5, &\qquad \mu_t \sim \mathcal{N}(\mu_T, 1) \text{ for } t > 1, \\ \sigma_T = 1, &\qquad \sigma_t \sim \chi^2(1) \text{ for } t < T, \\ \delta_T = 0, &\qquad \delta_t \sim \mathcal{N}(0,1) \text{ for } t < T. \end{aligned}$$

In this example, Assumptions 1, 2 (with $m_t(y) = 1 - \exp(-0.5\delta_t)(1-y)$) and 3 are satisfied, the latter because $\sigma_t \neq \sigma_T$ almost surely. The local curvature condition also holds, since $u \mapsto 1 - \exp(-0.5u)$ is concave. Figure 3 displays the bounds on $\Delta_1^{AME}(x)(x)$ for $T = 3, 4, 5$ and 6. Note that the bounds coincide for $T - 1$ points. This simply reflects our previous point identification result. Each $\mathbf{F}_t$ crosses once $\mathbf{F}_T$ and each at a different point. By Theorem 1, point identification is achieved at these $T-1$ crossing points. We also see that in the interval where we get finite bounds, that is to say the interval for which $-\infty < \underline{x}_T(x) < \overline{x}_T(x) < \infty$, the bounds are quite informative even with $T=3$. Figure 3 also shows that as T increases, both the bounds shrink and the interval on which we get finite bounds increase. For $T=6$, we get informative bounds for $x \in [1,\ 3.85]$, which corresponds roughly to 85% of the population. This means that we could also obtain finite bounds for the average partial effect for this large fraction of the total population.



$T=3$

$T=5$

$T=4$

$T=6$

Figure 3: Example of bounds on $\Delta_1^{AME}(x)$ for different values of x and $T=3,4,5$ and 6.

3.3 Point Identification with Exogenous Covariates

We consider here the case where exogenous covariates Z_t also affect Y_t, so that the model now writes

$$Y_t = g_t(X_t, Z_t, A_t) \qquad t = 1, ..., T. \tag{3.3}$$

We still focus on the effect of X_t hereafter. In this case, the preceding analysis can be conducted conditionally on Z_t. We briefly discuss this extension here, by considering only the discrete average and quantile effects

$$\begin{aligned} \Delta^{ATT}(x, x', z) &\equiv E\left[g_T(x', z, A_T) - g_T(x, z, A_T)|X_T = x, Z_T = z\right], \\ \Delta^{QTT}(p, x, x', z) &\equiv F^{-1}_{g_T(x',z,A_T)|X_T,Z_T}(p|x, z) - F^{-1}_{g_T(x,z,A_T)|X_T,Z_T}(p|x, z). \end{aligned}$$

The marginal effects can be handled similarly.

We first restate our previous conditions in this context. The rank variable is now defined conditionally on Z_t, $V_t = \mathbf{F}_{t|Z_t}(X_t)$ with

$$\mathbf{F}_{t|Z_t}(x) = \left(F_{X_{1t}|Z_t}(X_{1t}|Z_t), ..., F_{X_{kt}|Z_t}(X_{kt}|Z_t)\right).$$

Assumption 1'. *The conditional distributions of $X_t|Z_t = z$ is absolutely continuous with a convex support, supp$((V_t, Z_t))$ does not depend on t and for all $(s, t) \in \{1, ..., T\}^2$ and almost all $(v, z) \in$ supp(V_t, Z_t),*

$$A_s|V_s = v, Z_s = z \ \sim \ A_t|V_t = v, Z_t = z.$$

Next, we consider two versions of Assumptions 2 and 3. The trade-off between these two versions is basically between the generality of the model and data requirement. In the first version, we allow for more general time effects but the corresponding crossing condition is more demanding, because we should observe a crossing point for each value of z.

Assumption 2'. *We have either*
(i) for all t, $g_t(X_t, Z_t, A_t) = m_t(Z_t, g(X_t, Z_t, A_t))$, where $m_t(Z_t, .)$ is strictly increasing. Without loss of generality, we let $m_T(z, y) = y$ for all $(y, z) \in$ supp$((Y_T, Z_T))$;
or (ii) for all t, $g_t(X_t, Z_t, A_t) = m_t(g(X_t, Z_t, A_t))$, where m_t is strictly increasing. Without loss of generality, we let $m_T(y) = y$ for all $y \in$ supp(Y_T).

Assumption 3'. *We have either:*
(i) for all $(z, t) \in$ supp$(Z_T) \times \{1, ..., T - 1\}$, there exists $x^*_t(z)$ such that $\mathbf{F}_{T|Z_T}(x^*_t(z)|z) = \mathbf{F}_{t|Z_t}(x^*_t(z)|z) \in (0, 1)$.
or (ii) for all t, there exists (x^*_t, z^*_t) such that $\mathbf{F}_{T|Z_T}(x^*_t|z^*_t) = \mathbf{F}_{t|Z_t}(x^*_t|z^*_t) \in (0, 1)$.

These two sets of assumptions lead to the same results, which are qualitatively very similar to those of Theorem 1. The proof, which is very similar to the one of Theorem 1, is omitted.

Theorem 3. *Suppose that Assumption 1' and either Assumptions 2' (i) -3' (i) or Assumptions 2' (ii) -3' (ii) hold. Then, for almost all $(x, z) \in supp((X_T, Z_T))$, all $p \in (0, 1)$ and all $t \in \{1, ..., T-1\}$, the functions m_t and the average and quantile treatment effects $\Delta^{ATT}(x, q_t(x), z)$ and $\Delta^{QTT}(p, x, q_t(x), z)$ are identified.*

4 Extrapolation

As we have established in Theorem 1, we can point identify several treatment effect parameters under the relatively mild restrictions A1 to A3, but, as pointed out, these are by no means all possible causal effects one may be interested in. As we have seen in the previous section, many more treatment parameters can be set identified under often plausible curvature restrictions, in particular average marginal effects and effects of the form $\Delta^{ATT}(x, x')$. However, in any given application, these bounds may be wide, and to conduct inference may be cumbersome, or even impractical. Hence it makes sense to search for additional assumptions that yield point identification of average structural effects across the entire population, or even of all structural functions.

In the following, we propose two sets of non-nested restrictions that allow us to achieve point identification. The main restriction in the first approach constrains the heterogeneity term A_t to be scalar and have a monotonic effect on g. The main restriction in the second approach constrains X_t to have a linear or polynomial effect on Y_t. On the other hand, the coefficients on the explanatory variables are allowed to be random and correlated with X_t. These two approaches can be seen as providing a trade off. We either limit the extent of unobserved heterogeneity while allowing for flexibility in the way X_t enters the function or impose a functional form restriction on g but allow for a rich heterogeneity structure.

4.1 Scalar Monotonic Heterogeneity

In this subsection, we assume that heterogeneity is scalar and has a monotonic effect on the outcome. More formally:

Assumption 5. *$A_t \in \mathbb{R}$ and $g(X_t, .)$ is strictly increasing in its second argument.*

An example of model satisfying Assumption 5 is the linear quantile regression: $g(X_t, A_t) = X_t'\beta_{A_t}$, where $a \mapsto X_t'\beta_a$ is strictly increasing almost surely (i.e, there is comonotonicity). However, linearity is really not the essence here.

We also rely on the following technical restrictions:

Assumption 6. *(i) $X_t \in \mathbb{R}$ and its support $\mathcal{X} = [\underline{x}, \overline{x}]$ (with $-\infty \leq \underline{x} < \overline{x} \leq +\infty$) does not depend on t.*
(ii) A_t is uniformly distributed.
(iii) $(a, v) \mapsto F_{A_T|V_T}(a|v)$ is continuous on $(0, 1)^2$ and $a \mapsto F_{A_T|V_T}(a|v)$ is strictly increasing on $(0, 1)$ for all $v \in (0, 1)$.
(iv) $g(., .)$ is continuous on $\mathcal{X} \times (0, 1)$.
(v) q_t has a finite number of fixed points.

Under these additional conditions, we obtain

Theorem 4. *Under Assumptions 1-3, 5-6, m_t and g are identified.*

The proof relies on the observation that we have a triangular system

$$\begin{cases} \widetilde{Y}_t &= g(X_t, A_t) \\ X_t &= h(t, V_t) \end{cases}$$

where $h(t, v) = F_{X_t}^{-1}(v)$. This is a nonseparable triangular model where X_t is endogenous and t may be seen as an instrument. In this context, the usual exogeneity condition translates into time invariance of the distribution of (A_t, V_t). Because both $g(X_t, .)$ and $h(t, .)$ are strictly increasing, we can then use the identification results of D'Haultfoeuille & Février (2015) or Torgovitsky (2015). Note that under additional conditions, we could also obtain full identification when X_t is multivariate, using Theorem 5.2 of D'Haultfoeuille & Février (2015).

The reason why monotonicity makes a difference in our context is that we can then directly relate $g(q_t(x), a)$ with $g(x, a)$:

$$g(q_t(x), a) = Q_{q_t(x),x} \circ g(x, a),$$

where $Q_{q_t(x),x}$ is identified. This shows, as before, that $\Delta^{ATT}(x, q_t(x))$ is identified, but also that we can iterate, and relate $g(q_t \circ q_t(x), a)$ to $g(x, a)$, so that $\Delta^{ATT}(x, q_t \circ q_t(x))$ is identified as well. By repeating this argument, and using fixed points of q_t, we can show that the model is fully identified. Because the model is actually identified with $T = 2$, it may well be the case that identification is possible even without any fixed points when $T > 2$. This issue is left for future research.

It is instructive to relate Theorem 4 to results for nonlinear panel data models. The closest paper is the one of Evdokimov (2011), who considers the nonseparable model $Y_t = g_t(X_t, A_t)$ where A_t also satisfies Assumption 5 in his model. Compared to us, he imposes $A_t = U + \varepsilon_t$

and identification is achieved using the entire joint distribution of $(Y_1, X_1, ..., Y_T, X_T)$ and with $T \geq 3$. On the other hand, he does not impose any time invariance restriction on ε_t, nor does he put restriction on the effect of time on Y_t. Other related work is quantile regressions with "fixed effects". Rosen (2012) considers the model $Y_t = X_t'\beta_p + \alpha_p + \varepsilon_{tp}$, with $F^{-1}_{\varepsilon_{tp}|X_t,\alpha}(p|X_t, \alpha) = 0$ and where α_p may be correlated with X_t. He shows that β_p is not point identified for a fixed T. So it might seem surprising that with only $T = 2$, without panel data, and even without assuming linearity, identification can be achieved in such quantile regression models. Once more, the key difference between our setting and the one of Rosen (2012) is the time invariance condition that we impose on the error term.

4.2 Linear Correlated Random Coefficient Model

The second possible route for extrapolation is a random coefficient linear model of the form:

$$Y_t = \delta_t + A_{0t} + X_t'A_t, \tag{4.1}$$

where $A_t = (A_{1t}, ..., A_{kt})'$. Under this structure, the vector $E[A_T|X_T = x]$ is the vector of average marginal effects for individuals at x:

$$E[A_T|X_T = x] = (\Delta_1^{AME}(x), ..., \Delta_k^{AME}(x))'.$$

Moreover,

$$\Delta^{ATT}(x, q_t(x)) = (q_t(x) - x)'E[A_T|X_T = x].$$

Let us define the matrix $\mathbf{Q}(x)$ and the vector $\mathbf{\Delta}(x)$ as

$$\mathbf{Q}(x) = \begin{bmatrix} (q_1(x) - x)' \\ \vdots \\ (q_{T-1}(x) - x)' \end{bmatrix}, \quad \mathbf{\Delta}(x) = \begin{pmatrix} \Delta^{ATT}(x, q_1(x)) \\ \vdots \\ \Delta^{ATT}(x, q_{T-1}(x)) \end{pmatrix}.$$

If $\mathbf{Q}(x)$ is full column rank, we can identify $E[A_T|X_T = x]$ by

$$E[A_T|X_T = x] = (\mathbf{Q}(x)'\mathbf{Q}(x))^{-1}\mathbf{Q}(x)'\mathbf{\Delta}(x). \tag{4.2}$$

Apart from the vector of average marginal effects, we can then identify $\Delta^{ATT}(x, x')$, for any x', by

$$\Delta^{ATT}(x, x') = (x' - x)'E[A_T|X_T = x].$$

Note that the rank condition implies that $T - 1 \geq k$. It also implies that the distribution of X_t differs at each date, so that $q_s(x) \neq q_t(x)$. It makes sense that with several endogenous variables, more time variation on X_t is needed to identify causal effects.

Finally, if $\mathbf{Q}(X_T)$ is full rank almost surely, we point identify the vector of average marginal effect over the whole population, $\Delta^{AME} = (\Delta_1^{AME}, ..., \Delta_k^{AME})'$, by

$$\Delta^{AME} = E\left[A_T\right] = E\left[\left(\mathbf{Q}(X_T)'\mathbf{Q}(X_T)\right)^{-1}\mathbf{Q}(X_T)'\boldsymbol{\Delta}(X_T)\right].$$

We summarize these finding in the following theorem.

Theorem 5. *Under Assumptions 1-3 and Equation (4.1), δ_t, $\Delta^{ATT}(x, x')$ and $\Delta_j^{AME}(x)$ are identified for all x such that $\mathbf{Q}(x)$ is full column rank, and for any x' and $j \in \{1, .., k\}$. If $\mathbf{Q}(X_T)$ is full column rank almost surely, Δ_j^{AME} is point identified as well for $j \in \{1, .., k\}$.*

Thus, we recover the same parameter as Graham & Powell (2012), who also consider a random coefficient linear model similar to (4.1). They obtain identification with panel data, relying on first-differencing. Compared to them, we rely on variations in the cdf of X_t rather than on individual variations. We rely on a different, non-nested, restriction on the distribution of the error term. In particular, for the same individual, $A_{1t} - A_{1s}$ could be correlated with X_t in our framework.

Apart from identification, Equation (4.2) implies that the linearity assumption can be testable when $T - 1 > k$, because the system of equation is overidentified. In the univariate case, for instance, Equation (4.2) implies

$$\frac{\Delta^{ATT}(x, q_s(x))}{q_s(x) - x} = \frac{\Delta^{ATT}(x, q_t(x))}{q_t(x) - x} \quad \forall s \neq t.$$

We can use additional periods to identify higher moments of the distribution of the coefficients. For instance, with $k = 1$, $V(A_{01}|X_T = x)$, $V(A_{1T}|X_T = x)$ and $\text{Cov}(A_{01}, A_{1T}|X_T = x)$ can be shown to be identified with $T = 3$ as soon as $x, q_{12}(x)$ and $q_{13}(x)$ are distinct. Alternatively (here still with $k = 1$ to simplify), we can identify the random coefficient polynomial model of order T

$$Y_t = \delta_t + A_{0t} + A_{1t}X_t + ... + A_{Tt}X_t^T. \tag{4.3}$$

Identification works the same way as before. At the end, we recover not only average marginal effect, but actually $E(A_{kt}|X_t = x)$ for all $k = 1...T$ and all x such that $(x, q_{12}(x), ..., q_{1T}(x))$ are all distinct. Identification of Model (4.3) was studied before by Florens et al. (2008), but with cross-sectional data and under assumptions that typically rule out discrete instruments (see also Heckman & Vytlacil (1998) for any study of the identification of Model (4.1) with instruments). In contrast, we allow here for a time effect and rely only on a finite number of time periods, which would be equivalent to a discrete instrument.

5 Estimation of Average Treatment Effects and Large Sample Properties

We consider in this section estimators of the parameters $\Delta^{ATT}(x, q_t(x))$ and $\Delta^{QTT}(\tau, x, q_t(x))$ that are shown to be identified in Theorem 1. We suppose for that purpose to observe two independent samples corresponding to the periods t and T. For simplicity, we suppose hereafter that $k = 1$ and that the two corresponding sample sizes are identical.

Assumption 7. *We observe the two independent samples $(Y_{it}, X_{it})_{i=1...n}$ and $(Y_{iT}, X_{iT})_{i=1...n}$, which are both samples of i.i.d. random variables with cdf F_{Y_t,X_t} and F_{Y_T,X_T} respectively.*

Our estimator follows closely our identification strategy. Let us define

$$\Psi_n(x) = \widehat{F}_{X_T}(x) - \widehat{F}_{X_t}(x),$$

where $\widehat{F}_{X_T}$ (resp. $\widehat{F}_{X_t}$) denotes the empirical cdf of X_T (resp. X_t). We first estimate x_t^* by

$$\widehat{x}_t^* = \min\left\{x \in \left[\widehat{F}_{X_t}^{-1}(\underline{p}), \widehat{F}_{X_t}^{-1}(\overline{p})\right] : |\Psi_n(x)| \leq |\Psi_n(x')| \; \forall x' \in \left[\widehat{F}_{X_t}^{-1}(\underline{p}), \widehat{F}_{X_t}^{-1}(\overline{p})\right]\right\}, \qquad (5.1)$$

where $\widehat{F}_{X_t}^{-1}$ denotes the empirical quantile function and $0 < \underline{p} < \overline{p} < 1$ are two given constants used to avoid reaching the boundaries of the support of X_t. Note that the minimum in (5.1) is well defined because Ψ_n is left continuous.

Next, we estimate $q_t(x) = F_{X_t}^{-1} \circ F_{X_T}(x)$ by its empirical counterpart $\widehat{q}_t(x) = \widehat{F}_{X_t}^{-1} \circ \widehat{F}_{X_T}(x)$. We then estimate m_t using an empirical counterpart of (3.1). For that purpose, we estimate the conditional cdf $F_{Y_\tau|X_\tau}$, for $\tau \in \{t, T\}$, by

$$\hat{F}_{Y_\tau|X_\tau}(y|x) = \frac{\sum_{i=1}^n \mathbb{1}\{Y_{i\tau} \leq y\} K\left(\frac{x - X_{i\tau}}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x - X_{i\tau}}{h_n}\right)},$$

where K is a kernel function and h_n denotes the bandwidth. We then let $\hat{F}_{Y_\tau|X_\tau}^{-1}(.|x)$ denote the generalized inverse of $\hat{F}_{Y_\tau|X_\tau}(.|x)$. We estimate m_t by

$$\widehat{m}_t(y) = \hat{F}_{Y_t|X_t}^{-1}\left[\hat{F}_{Y_T|X_T}(y|\widehat{x}_t^*)|\widehat{x}_t^*\right].$$

$\Delta^{ATT}(x, q_t(x))$ and $\Delta^{QTT}(\tau, x, q_t(x))$ satisfy, under Assumptions 1-3,

$$\Delta^{ATT}(x, q_t(x)) = E[m_t(Y_t)|X_t = q_t(x)] - E(Y_T|X_T = x),$$
$$\Delta^{QTT}(\tau, x, q_t(x)) = F_{m_t(Y_t)|X_t}^{-1}(\tau|q_t(x)) - F_{Y_T|X_T}^{-1}(\tau|x).$$

We estimate these two parameters by

$$\widehat{\Delta}^{ATT}(x,q_t(x)) = \frac{\sum_{i=1}^n \hat{m}_t^{-1}(Y_{it})K\left(\frac{x-X_{it}}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_{it}}{h_n}\right)} - \frac{\sum_{i=1}^n Y_{iT}K\left(\frac{x-X_{iT}}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_{iT}}{h_n}\right)},$$

$$\widehat{\Delta}^{QTT}(\tau,x,q_t(x)) = \widehat{F}^{-1}_{\widehat{m}_t(Y_t)|X_t}(\tau|\widehat{q}_t(x)) - \widehat{F}^{-1}_{Y_T|X_T}(\tau|x).$$

For simplicity, we chose here the same kernels and bandwiths for each nonparametric terms, though we could obviously consider different ones. We establish below that $\widehat{\Delta}^{ATT}(x,q_t(x))$ and $\widehat{\Delta}^{QTT}(\tau,x,q_t(x))$ are consistent and asymptotically normal. Our result is based on the following conditions.

Assumption 8. *(Conditions for the root-n consistency of $\widehat{x}_t^*$ and $\widehat{q}_t(x)$)*
(i) There exists a unique x_t^* satisfying $F_{X_t}(x_t^*) = F_{X_T}(x_t^*) \in (0,1)$ and $F_{X_t}(x_t^*) \in (\underline{p},\overline{p})$.
(ii) For $\tau \in \{t,T\}$, X_τ admits a continuous density f_{X_τ} satisfying, for all x in the interior of $\mathcal{X}$, $f_{X_\tau}(x) > 0$. Moreover, $f_{X_t}(x_t^*) \neq f_{X_T}(x_t^*)$.

Assumption 9. *(Regularity conditions on (X_τ, Y_τ), for $\tau \in \{t,T\}$)*
(i) supp$(X_\tau,Y_\tau) = \mathcal{X}\times\mathcal{Y}$ with $\mathcal{Y} = [\underline{y},\overline{y}]$ with $-\infty < \underline{y} < \overline{y} < +\infty$.
(ii) $F_{Y_\tau|X_\tau}(.|.)$ is continuously differentiable with, for every $x \in \mathcal{X}$, $\inf_{y\in\mathcal{Y}} f_{Y_\tau|X_\tau}(y|x) > 0$. (iii) For every $y \in \mathcal{Y}$, $F_{Y_\tau|X_\tau}(y|.)$ and f_{X_τ} are twice differentiable. f_{X_τ}, $|f'_{X_\tau}|$ and $|f''_{X_\tau}|$ are bounded. $\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}} |\partial_x F_{Y_\tau|X_\tau}(y|x)| < \infty$ and $\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}} |\partial_{xx} F_{Y_\tau|X_\tau}(y|x)| < \infty$.

Assumption 10. *(Conditions on the kernels and bandwidths)*
(i) $nh_n^3/|\log(h_n)| \to +\infty$, $nh_n^5 \to 0$.
(ii) K has a compact support, is differentiable with K' of bounded variation and satisfies $K(y) \geq 0$ for all y. Besides, $\int K(y)dy = 1$ and $\int yK(y)dy = 0$.

Assumption 8-(i) strengthens Assumption 3, by assuming the unicity of the crossing point. We make this assumption for the sake of simplicity. If we allowed for several crossing points, we would have to estimate the set of such crossing points, instead of a single point. Assumption 8-(ii) is a mild regularity condition on F_{X_T} and F_{X_t}. As Lemmas 1 and 2 in Appendix B show, these two restrictions ensure that $\widehat{x}_t^*$ and $\widehat{q}_t(x)$ are root-N consistent. Assumption 9 provides a set of conditions ensuring that $\widehat{m}_t$ is consistent and asymptotically normal. Conditions (i) and (ii) are also made by Athey & Imbens (2006), without any X_τ in their case, in another context where quantile-quantile transforms must be estimated. Condition (iii) is required as well here as we deal with nonparametric cdfs rather than empirical cdfs, as Athey & Imbens (2006). Finally, Assumption 10 is a standard condition on the bandwidths and the kernels appearing in the nonparametric estimators . We impose $nh_n^5 \to 0$ here in order to avoid any asymptotic bias on $\widehat{\Delta}^{ATT}(x,q_t(x))$ and $\widehat{\Delta}^{QTT}(p,x,q_t(x))$.

Theorem 6. *Suppose that Assumptions 1-3 and 7-10 are satisfied for each $\tau \in \{t, T\}$. Then, for any $x \in \mathcal{X}$ such that F_{X_t} is differentiable at $q_t(x)$ with $F'_{X_t}(q_t(x)) > 0$,*

$$\sqrt{nh_n}\left(\widehat{\Delta}^{ATT}(x, q_t(x)) - \Delta^{ATT}(x, q_t(x))\right) \xrightarrow{d} \mathcal{N}(0, V_1)$$
$$\sqrt{nh_n}\left(\widehat{\Delta}^{QTT}(\tau, x, q_t(x)) - \Delta^{QTT}(\tau, x, q_t(x))\right) \xrightarrow{d} \mathcal{N}(0, V_2),$$

for some V_1, V_2.

A proof of this result is provided in Section B in the appendix. We do not provide here the form of the asymptotic variance as it involves many terms, because of the multiple nonlinear compositions of nonparametric estimators. In practice, we suggest to rely on bootstrap, as we do in the application below. We conjecture that the bootstrap is consistent in our setting, though a formal proof of its validity is beyond the scope of this paper. The main issue for establishing its validity would be to prove the (conditional) weak convergence of the process

$$G^*_{nx\tau} = \sqrt{nh_n}\left(\widehat{F}^*_{Y_\tau|X_\tau}(.|x) - \widehat{F}_{Y_\tau|X_\tau}(.|x)\right), \ \tau \in \{t, T\},$$

where $\widehat{F}^*_{Y_\tau|X_\tau}$ is the bootstrap counterpart of $\widehat{F}_{Y_\tau|X_\tau}$. Up to our knowledge, such a result is not available in the literature yet.

6 Application to the Effect of Maternal Age on Birth Weight

In most industrialized economies, there is a pronounced trend towards a later age at which a family is established. In particular, mother's childbearing age is steadily increasing. This phenomenon is well documented, and the individual and social costs have been extensively studied (see, e.g., Heffner, 2004, for a medical perspective and Hofferth, 1998, for an economic overview). In this section, we want to focus on one aspect that has received less attention, which we think is important: the ceteris paribus effects of mother's age at first birth, denoted X_t, on infant birth weight Y_t. The reason why we focus on birth weights is that infant birth weight plays a very important role in the literature on health economics. In particular, infant birth weights are often thought of as playing a dual role, both as an output and as an input. On one hand, birth weights are used as a measure of an outcome, namely infant health, that involves maternal behaviors and environments as primitive inputs (see, e.g., Rosenzweig & Schultz, 1983, Corman et al., 1987, Grossman & Joyce, 1990, Geronimus & Korenman, 1992, Rosenzweig & Wolpin, 1991, Rosenzweig & Wolpin, 1995, Evans & Ringel, 1999, Currie & Moretti, 2003 and Camacho, 2008). On the other hand, birth weight is itself used as a measure for the initial input, the condition of an individual at birth, that eventually "produces" educational attainment, employment, and earnings as outcomes (see, e.g., Behrman et al., 1994, Currie & Hyson, 1999, Behrman & Rosenzweig, 2004, Black et al., 2007). Both aspects make understanding the causal determinants of a child's birth weight an issue of first order importance.

In most economic approaches, maternal age and the decision to give birth are made endogenously through life cycle plans made by forward-looking decision makers. The key econometric issue is to separate the physiological effects of mother's age from the effects of the economic environment that is associated with a mother's age. Standard panel data approaches may not allow one to recover this causal effect of maternal age for several reasons. First, we would have to focus on mothers with at least two children, clearly a selected subpopulation. Second, the time span between the birth of the first and the second child (which would be the regressor in a first-differences model) is likely to not be exogenous. Third, the model may be dynamic because the first pregnancy may also have an effect on subsequent pregnancies, including the timing.

Another option for identifying the causal effect of maternal age would be to use a standard instrumental variable. It is however challenging, in particular with administrative data which do not contain a lot of information about the socioeconomic background, to find a variable affecting maternal age but not the infant's weight directly. All of this taken together may help to explain the notable absence of research on this issue.

In our empirical study, we use the repeated cross sections of the Natality Vital Statistics System of the National Center for Health Statistics for years 1990-2010. Following our notation, we let X_t and Y_t denote mother's age and infant birth weight, respectively, where t denotes the index for the years 1990-2010. To mitigate the endogenous effects of dynamic optimization, we focus on the subsample consisting of first births. To mitigate the effect of smoking and education, which may affect birthweight and whose distribution has changed substantially over the period, we focus on non-smoking mothers with twelve years or more of education. They represent 64.7% of the population of mothers of first born children. Finally, we exclude preterm birth, namely infants born after less than 37 weeks of gestation, which represent 8.8% of the sample. We do so because we expect that the time affects preterm newborns very differently from other infants, which may cause a violation of Assumption 2. By technical change, more preterm newborns are saved nowadays, which means that preterm newborns are on average lighter today. But as we shall see, technical change has a positive effect on small infants when excluding preterm births. Of course, maternal age is also associated with preterm birth, so including it might increase the magnitude of the effect of maternal age. We consider this overall effect of maternal age at the end of the section.

Table 1 shows summary statistics of (Y_t, X_t), for the repeated cross sections of first births.

The displayed values are the sample means of age and birth weight, with the sample standard errors shown in parentheses. These aggregate statistics suggest two time trends – the mean infant birth weight is decreasing over time, and mean age of mother at first birth is increasing over time. This simple observation alone, however, does not allow us to conclude the causal effects of mother's age on infant birth weight due to omitted explanatory variables which may also follow certain time trends. Our approach controls for these omitted variables, as well as latent time trends that may reflect time progresses of medical technologies.

Year	1990	2000	2010
Birth Weight	3440.8 (507.6)	3408.2 (501.6)	3355.7 (469.1)
Maternal Age	25.28 (4.86)	26.06 (5.43)	26.32 (5.54)
Sample Size	590,224	722,304	597,834

Source: Natality Vital Statistics System of the National Center for Health Statistics (subsamples of non-smoking, with twelve years or more of education mothers at first birth, with gestation greater than 36 weeks).

Notes: we display the sample means, with standard errors in parentheses.

Table 1: Summary statistics of the main variables for years between 1990 and 2010.

To see whether our approach is applicable, we first consider the time shift of the cumulative distribution of mother's age at first birth. We focus on the pair of most distant years in our data set, namely 1990 and 2010, but we will later use cross sections of the other years for robustness checks. Figure 6 shows the cdfs of maternal age at first birth in the years 1990 and 2010, smoothed by interpolation of the discretely supported X_t. Observe that they are almost confounded for ages between 15 and 20, while X_{2010} first-order stochastically dominates X_{1990} otherwise, confirming the trend observed on average maternal age. Assumption 3 is therefore satisfied, and we can use $x^* = 20$, for instance, to form the temporal control group to disentangle the time effects from the effect of age for those mothers older than 20.

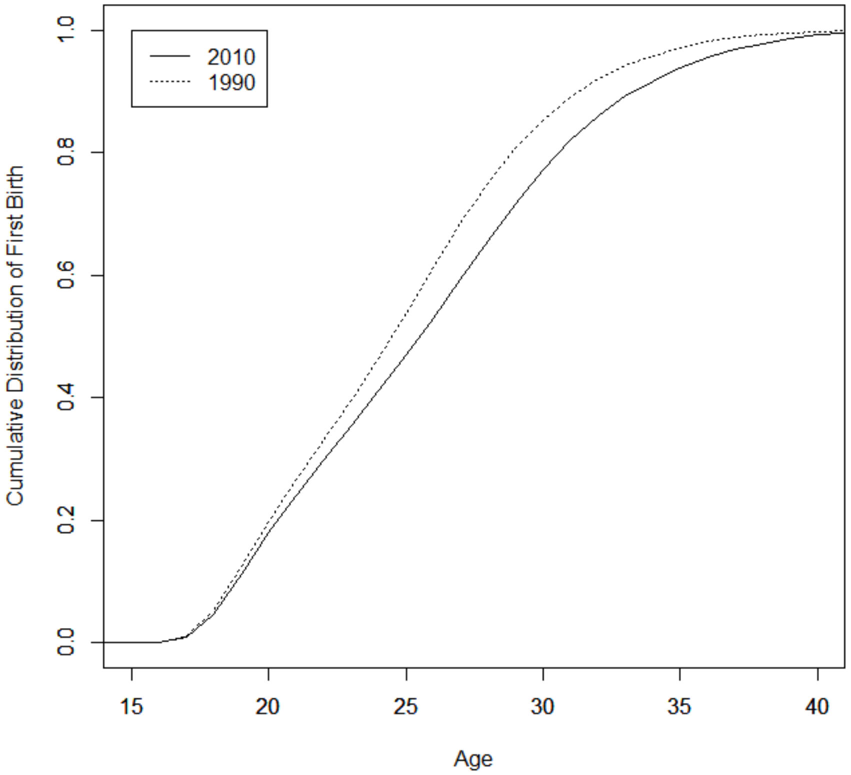
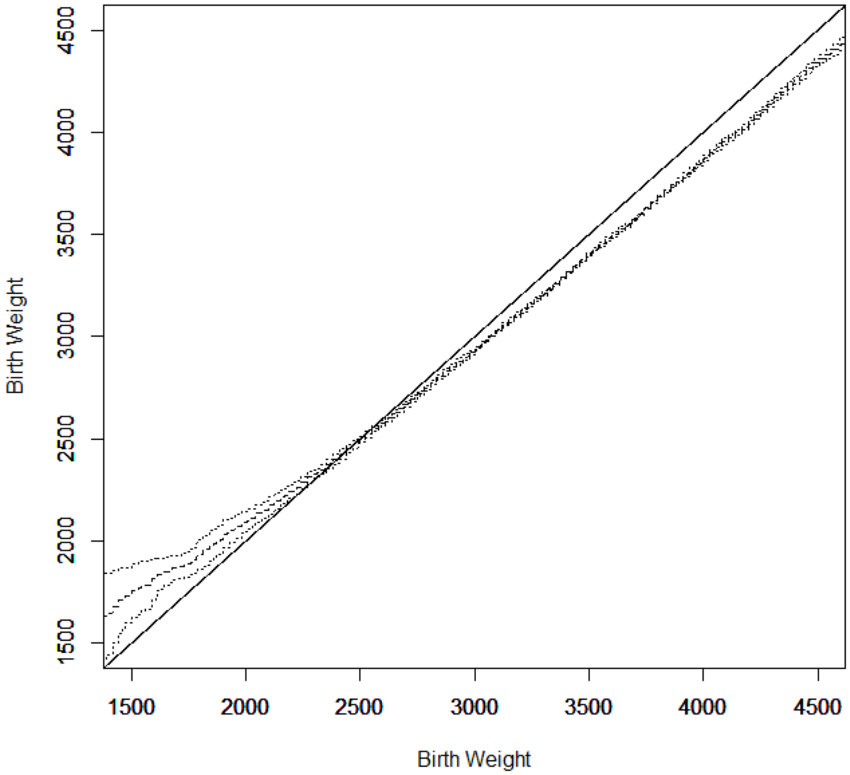


Figure 4: Cumulative distribution of maternal age at first birth (1990-2010).

Our approach relies on the conditional stationarity assumption 1, which states that the distribution of unobservables A_t of all mothers who have rank $V_t = v$ in the first birth age distribution is the same as the distribution of A_s given $V_s = v$. To understand this requirement, think of A_t as variable that captures the healthiness of the lifestyle. Endogeneity arises here because the distribution of healthy lifestyles of mothers who have their first child at 18 (think of teenage pregnancies) is likely to be different from the one of mothers who have their first child at 28, for instance. Then, our identifying assumption says that mothers at the third quartile (say) of first birth age in 1990 (which is 28), have the same distribution in terms of healthy lifestyles, as the mothers at the third quartile of first birth age in 2010 (which is 30). This is plausible if, loosely speaking, these two subpopulations are at the same position in the distribution of alcohol consumption, physical activity, educational backgrounds, etc., as is likely the case given the close proximity of the two cdfs, and the not too distant time periods.

Given the crossing point $x^* = 20$ that we use to construct the "nontreated" control group, we use the two conditional cumulative distributions, $F_{Y_{1990}|X_{1990}=20}$ and $F_{Y_{2010}|X_{2010}=20}$ to identify the effect of time in isolation (see Figure 5). Interestingly, we point out an heterogenous time effect on birth weight. Namely, we estimate an increase in the weight of small infants, but a decrease in the weight of large ones. The increase for small infants may be seen as a positive effect of technical change. This may also be the case for large weights, because large babies may cause trouble at delivery (see, e.g., Stotland et al., 2004). The estimated decrease of birth weight for such large babies may be due to increased numbers of inductions and C-sections, which causes overweight babies that would have been delivered naturally in 1990 to be induced earlier and at lower weight in 2010 - thus resulting in overall reduction of birth weights particularly toward the right tail of the distribution. This example illustrates the potential importance of not restricting oneself to additive constant time effects ex ante.

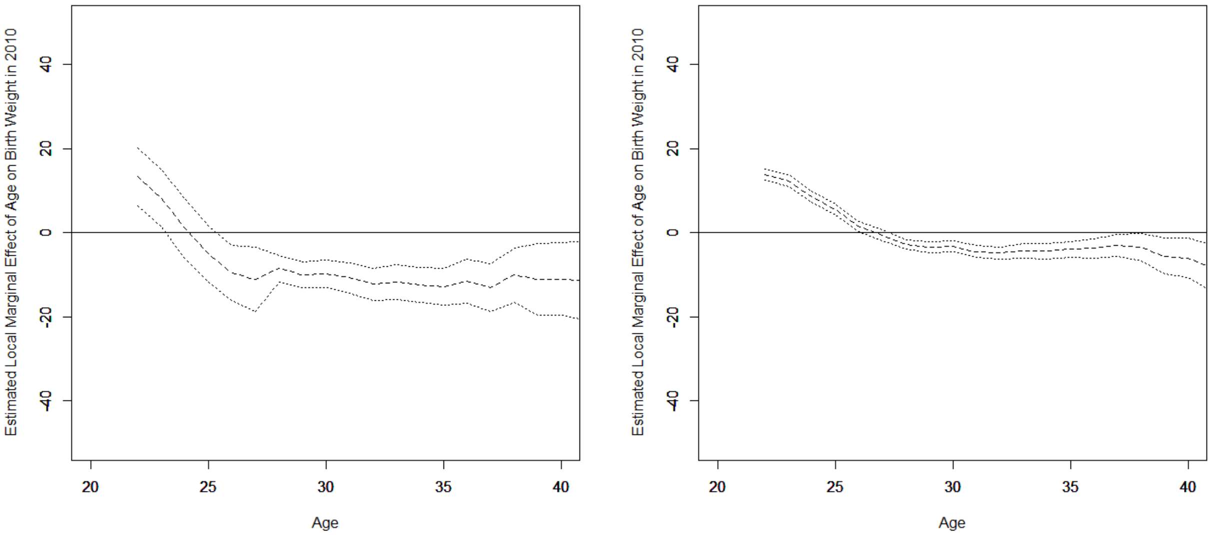


The horizontal and vertical axes correspond respectively to weight in 1990 and 2010. The time effect is obtained using Age=20 as the control group. The displayed curves indicate the estimates and pointwise 95% confidence intervals, obtained by bootstrap.

Figure 5: Time effects on birth weight in grams from 1990 to 2010.

Using the estimated time effects, we in turn estimate the marginal effects of interest. As we have $F_{X_{1990}}(x) \approx F_{X_{2010}}(x)$ for all $x \leq 20$ and all $x > 40$, it is only for $20 < x \leq 40$ that heterogeneous marginal effects can be obtained, which is of course a very large part of the population. Figure 6 (a) shows the average estimated effects $\Delta^{ATT}(x, q(x))/(q(x) - x)$ for the year 2010 together with 95% bootstrap confidence intervals. Note that because $q(x)$ is close to x, these effects are likely to approximate well the average marginal effects $\Delta_j^{AME}(x)(x)$, and with slight abuse of language we refer to them as marginal effects hereafter. The mean estimates are negative throughout most of the effective domain of mother's age. Furthermore, these marginal effects are significantly negative at the five percent level for 26- through 40-year old mothers, implying that adverse physiological effects of aging on birth weight are likely to exist.

Note that this result accounts for the endogeneity of mother's age at first birth, which may for instance be the result of family planning by forward-looking individuals, or may be the result of the tendency that wealthy mothers delay first birth. To see the degree to which this endogeneity would affect estimates of the marginal effects, if not properly taken care of, we also compute a naive cross section estimate of the marginal effects, assuming that mother's age at first birth were exogenous, i.e., the effect of an exogenous shift from x to x' is analyzed using $\mathrm{E}[Y_{2010} \mid X_{2010} = x'] - \mathrm{E}[Y_{2010} \mid X_{2010} = x]$, instead of $\mathrm{E}[F^{-1}_{Y_{2010}|X_{2010}=20} \circ F_{Y_{1990}|X_{1990}=20}(Y_{1990}) \mid X_{1990} = x'] - \mathrm{E}[Y_{2010} \mid X_{2010} = x]$. Figure 6 (b) shows these "naive" estimates. Compared with Figure 6 (a), which accounts for endogeneity, the mean estimates in Figure 6 (b) are much smaller in absolute value. Furthermore, these naively estimated marginal effects are strongly significantly positive until age 26. One possible explanation of this outcome is that wealthier and more educated women, i.e., women with a healthier lifestyle on average, who may tend to have newborns with higher birth weights are likely to defer childbearing in these early ages. An estimator that does not account for endogeneity might wrongly classify this to be a positive marginal effect of mother's age on birth weight.



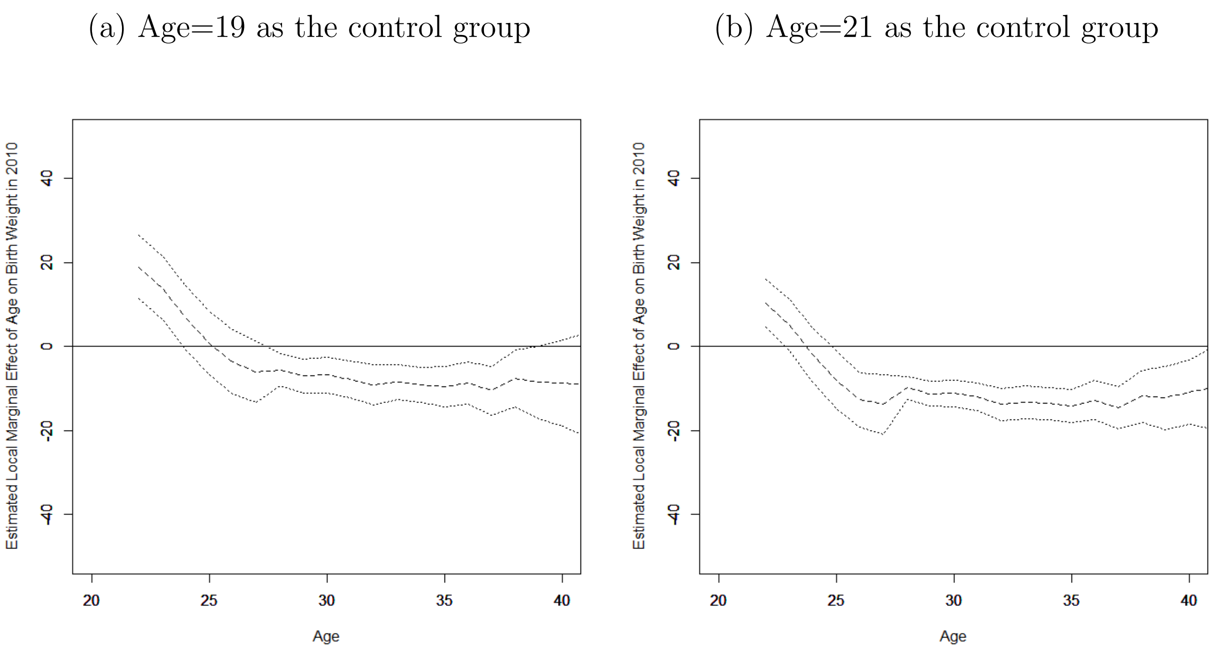
(a) Mother's age may be endogenous.

(b) Mother's age assumed to be exogenous.

Notes: Effects for first birth in 2010. The displayed curves indicate the estimates and pointwise 95% confidence intervals obtained by bootstrap.

Figure 6: Estimated marginal effects of mother's age on infant birth weights.

All of these graphs show estimates of the marginal effects in 2010 using age=20 as a control group, and 1990 as the reference period in our analysis. To check the robustness of our specification, we next demonstrate that the qualitative patterns are similar even if we use alternative definitions of control groups or switch the reference periods. Figure 7 shows the marginal effects of age with alternative control groups in terms of age. Figure 8 considers different second periods, instead of 1990. Overall, all these estimates are very consistent with our previous results, suggesting that our method is robust in this application. Finally, we also compute the marginal effects of maternal age when including preterm birth. Because maternal age also affects preterm birth, the computation of the overall effect is complicated in this case, and we have to rely on further assumptions (details are provided in Appendix B, but the results are essentially the same (see Figure 9 in Appendix B).

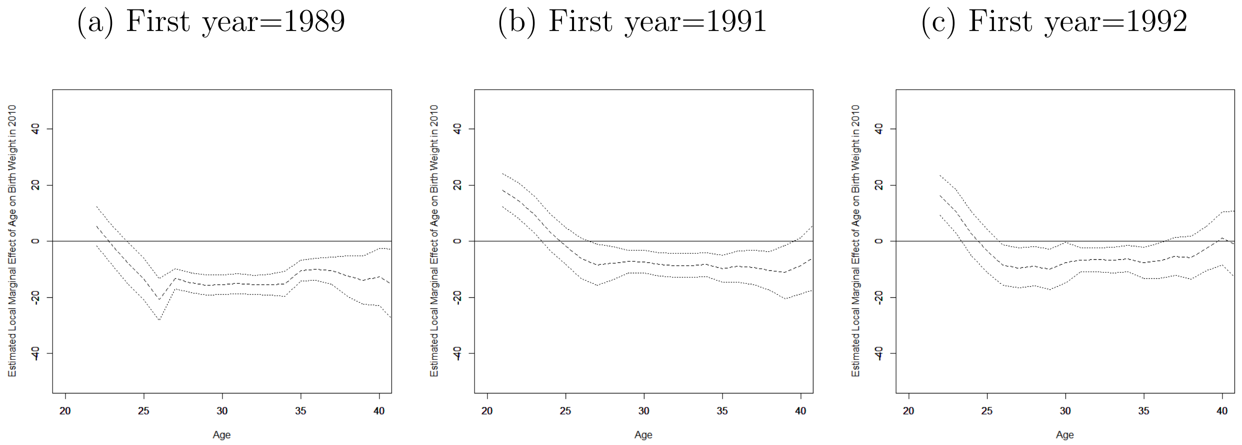


(a) Age=19 as the control group

(b) Age=21 as the control group

Notes: Effects for first birth in 2010. The displayed curves indicate the estimates and pointwise 95% confidence intervals obtained by bootstrap.

Figure 7: Influence of the control group on marginal effects



(a) First year=1989

(b) First year=1991

(c) First year=1992

Notes: effects for first birth in 2010, with bootstrap confidence intervals.

Figure 8: Influence of the initial year on marginal effects.

Many economies have seen the ongoing trend of delaying marriage and first birth. Social costs of this tendency have been discussed extensively, but we found that there may be costs to the health of children, at least in as far as they are reflected in reduced birth weights. Based on our mean estimates, delaying first birth by two year results in about 20 gram loss of birth weights for mothers older than 26. This number is statistically significant, and is larger than the weight reduction that could result from one additional cigarettes smoked by a smoking pregnant mother per day (see Hoderlein & Sasaki, 2013). Couples may want to consider these potential health costs when making decisions to delay marriage and child birth. Furthermore, since these immediate health costs can have significant losses in the long run (see, e.g. Behrman et al., 1994, Currie & Hyson, 1999, Behrman & Rosenzweig, 2004, Black et al., 2007), our empirical results may inform policy makers on how to improve welfare outcomes by affecting the age of first birth.

7 Conclusion

In sharp contrast to panel data, repeated cross sections are seldom considered as an alternative to instruments when endogeneity is suspected. In this paper, we show that repeated cross sections can resolve the endogeneity issue in a way that is reminiscent of a difference-in-difference approach, even if every individual is only observed once and the variable (treatment) of interest is continuous. Importantly, this is possible even if time has a nonlinear and heterogeneous effect, meaning that the additive decomposition typically assumed with difference-in-differences is not a necessary condition to conduct such an analysis. However, other conditions are important: The first key assumption is a time invariance condition, which - as we argue - differs from the one usually assumed in panel data models. The second is a crossing condition, which basically holds when the distribution of the treatment is shifted by time in a non-homogeneous way. Under these conditions, a number treatment effect parameters are point identified, while others are set identified. Moreover, we propose two distinct additional set of restrictions that yield point identification of most commonly analyzed treatment effects. The first such additional set of restrictions is a linear correlated random coefficient model recently considered in the panel data literature (see, e.g., Arellano & Bonhomme, 2012, Graham & Powell, 2012). The second does not impose linearity, but restricts the error term to be scalar, in line with the literature on nonseparable models. We show that such an approach works well in an application that discusses the effect of maternal age at first birth on the birth weight of a newborn, and uncovers, as we feel, interesting details.

A Proofs for Identification Results

A.1 Proof of Theorem 1

The result for m_t and $\Delta^{ATT}(x, q_t(x))$ has already been proved in the text. As for $\Delta^{QTT}(x, q_t(x))$, we have

$$\begin{aligned} F^{-1}_{\widetilde{Y}_t|X_t}(p|q_t(x)) &= F^{-1}_{g(q_t(x),A_t)|V_t}(p|\mathbf{F}_T(x)) \\ &\overset{A.1}{=} F^{-1}_{g(q_t(x),A_T)|V_T}(p|\mathbf{F}_T(x)) \\ &= F^{-1}_{g(q_t(x),A_t)|X_T}(p|x). \end{aligned}$$

The result follows.

Now consider marginal effects. Consider a sequence $(x_n)_{n\in\mathbb{N}}$ such that for all $i \in \{1, ..., K\}$, $i \neq j$, $x_{in} = x^*_{it}$ and $q_{jt}(x_{jn}) \neq x_{jn}$. We have

$$\frac{\Delta_j^{ATT}(x_n, q_t(x_n))}{q_{jt}(x_n) - x_{jn}} = \int \frac{g(q_t(x_n), a) - g(x_n, a)}{q_{jt}(x_n) - x_{jn}} f_{A_T|X_T}(a|x_n)da.$$

By Assumption 4-(i) and (ii), we have, for almost all a,

$$\begin{aligned} \frac{g(q_t(x_n), a) - g(x_n, a)}{q_{jt}(x_n) - x_{jn}} f_{A_T|X_T}(a|x_n) &= \frac{\partial g}{\partial x_j}(\widetilde{x}_n, a) f_{A_T|X_T}(a|x_n) \\ &\longrightarrow \frac{\partial g}{\partial x_j}(x^*_t, a) f_{A_T|X_T}(a|x^*_t), \end{aligned}$$

where $\widetilde{x}_n$ is such that $\widetilde{x}_{in} = x^*_{it}$ for all $i \neq j$ and $\widetilde{x}_{jn} \in [x_{jn}, q_{jt}(x_{jn})]$. Moreover, for n large enough, x_n and $\widetilde{x}_n$ belong to the neighborhood $\mathcal{N}$ considered in Assumption 4. Thus, for n large enough,

$$\left| \frac{\partial g}{\partial x_j}(\widetilde{x}_n, a) f_{A_T|X_T}(a|x_n) \right| \leq \left| \sup_{x'\in\mathcal{N}} \frac{\partial g}{\partial x_j}(x', a) \right| \left| \sup_{x'\in\mathcal{N}} f_{A_T|X_T}(a|x') \right|.$$

The right-hand side is integrable by Assumption 4-(iii). Thus, by the dominated convergence theorem,

$$\int \frac{g(q_t(x_n), a) - g(x_n, a)}{q_t(x_n) - x_n} f_{A_T|X_T}(a|x_n)da \longrightarrow \int \frac{\partial g}{\partial x_j}(x^*_t, a) f_{A_T|X_T}(a|x^*_t)da = \Delta_j^{AME}(x^*_t).$$

Finally, let us turn to $\Delta_j^{QME}(p, x)$. We have

$$\begin{aligned} \frac{\Delta_j^{QTT}(p, x_n, q_t(x_n))}{q_{jt}(x_{jn}) - x_{jn}} &= \frac{F^{-1}_{g(q_t(x_n),A_T)|X_T}(p|x_n) - F^{-1}_{g(x_n,A_T)|X_T}(p|x_n)}{q_{jt}(x_{jn}) - x_{jn}} \\ &= \frac{\partial F^{-1}_{g(x',A_T)|X_T}(p|x_n)}{\partial x'_j}_{|x'=\widetilde{x}'_n}, \end{aligned}$$

where $\widetilde{x}'_n$ is such that $\widetilde{x}'_{in} = x^*_{it}$ for all $i \neq j$ and $\widetilde{x}'_{jn} \in [x_{jn}, q_{jt}(x_{jn})]$. By Assumption 4-(iv), the last derivative converges to

$$\frac{\partial F^{-1}_{g(x',A_T)|X_T}(p|x^*_t)}{\partial x'_j}_{|x'^*_t} = \Delta_j^{QME}(p, x^*_t).$$

A.2 Proof of Theorem 2

Suppose first that g is locally concave on $[\min(x, \underline{x}_T(x')), \overline{x}_T(x')]$. Then, for all $x_1 \leq x' \leq x_2$, almost surely,

$$\frac{g(x_2, A_T) - g(x, A_T)}{x_2 - x} \leq \frac{g(x', A_T) - g(x, A_T)}{x' - x} \leq \frac{g(x_1, A_T) - g(x, A_T)}{x_1 - x}. \qquad \text{(A.1)}$$

Taking $x_1 = \underline{x}_T(x')$ and $x_2 = \overline{x}_T(x')$ and integrating conditional on $X_T = x$, we obtain

$$(x' - x)\frac{\Delta^{ATT}(x, \overline{x}_T(x'))}{\overline{x}_T(x') - x} \leq \Delta^{ATT}(x, x') \leq (x' - x)\frac{\Delta^{ATT}(x, \underline{x}_T(x'))}{\underline{x}_T(x') - x}.$$

The inequality is simply reverted if g is locally convex. Hence, in either case,

$$\begin{aligned} & (x' - x)\min\left\{\frac{\Delta^{ATT}(x, \underline{x}_T(x'))}{\underline{x}_T(x') - x}, \frac{\Delta^{ATT}(x, \overline{x}_T(x'))}{\overline{x}_T(x') - x}\right\} \leq \Delta^{ATT}(x, x') \\ \leq \; & (x' - x)\max\left\{\frac{\Delta^{ATT}(x, \underline{x}_T(x'))}{\underline{x}_T(x') - x}, \frac{\Delta^{ATT}(x, \overline{x}_T(x'))}{\overline{x}_T(x') - x}\right\}. \end{aligned}$$

The reasoning is the same for marginal effects using, instead of Equation (A.1),

$$\frac{g(x_2, A_T) - g(x, A_T)}{x_2 - x} \leq \frac{\partial g}{\partial x}(x, A_T) \leq \frac{g(x_1, A_T) - g(x, A_T)}{x_1 - x}.$$

A.3 Proof of Theorem 4

m_t is identified by Theorem 1. We now show that we can apply Theorem 4.4 of D'Haultfoeuille & Février (2015). The idea for that is to observe that we have a triangular system

$$\left\{\begin{array}{rcl} \widetilde{Y}_t & = & g(X_t, A_t) \\ X_t & = & h(t, V_t) \end{array}\right.$$

where $h(t, v) = F^{-1}_{X_t}(v)$. This is a nonseparable triangular model where X_t can be seen as the potential endogenous variable corresponding to the value t of an instrument. The only difference between this model and the one considered by D'Haultfoeuille & Février (2015) is that we assume here rank similarity instead of rank invariance. Namely, A_t and V_t are allowed to vary with t here, while the potential error terms corresponding to each value of the instrument are identical in D'Haultfoeuille & Février (2015). But this does not affect the reasoning. In particular,

$$\begin{aligned}
F_{\widetilde{Y}_t|X_t}(g(x,a)|x) &= P(g(x,A_t) \leq g(x,a)|V_t = \mathbf{F}_t(x)) \\
&\overset{A.5}{=} P(A_t \leq a|V_t = \mathbf{F}_t(x)) \\
&\overset{A.1}{=} P(A_{t'} \leq a|V_{t'} = \mathbf{F}_t(x)) \\
&\overset{A.5}{=} P(g(q_{tt'}(x), A_{t'} \leq g(q_{tt'}(x),a)|X_{t'} = q_{tt'}(x)) \\
&= F_{\widetilde{Y}_{t'}|X_{t'}}(g(q_{tt'}(x),a)|q_{tt'}(x)).
\end{aligned}$$

These equalities were also derived by D'Haultfoeuille & Février (2015) (see p.6), the only difference being that they used an independence assumption (Assumption 1 in their paper) in place of our time invariance Assumption 1 here. But both lead to the same conclusion. Note also that the function $q_{tt'}$ plays the role of their function s_{ij}. The rest of the proof is identical, noting that their Assumption 2 holds by Assumption 5, and their Assumptions 3 and 4 are satisfied by Assumption 4.

B Proofs and Auxiliary Statements for Large Sample Properties

B.1 Preliminary Lemmas

Lemma 1 (Consistency of $\widehat{x}_t^*$)**.** *If Assumptions 1, 7 and 8-(i) hold,* $\widehat{x}_t^* - x_t^* = o_p(1)$.

Proof. Let $M_n(x) = -|\Psi_n(x)|$ and $M(x) = -|F_{X_T}(x) - F_{X_t}(x)|$ and let $I = [F_{X_T}^{-1}(\underline{p}) - \varepsilon, F_{X_T}^{-1}(\overline{p}) + \varepsilon]$ for some $\varepsilon > 0$. By Assumption 8-(i), x_t^* is the unique maximum of M on I. Besides, by Glivenko-Cantelli's theorem,

$$\begin{aligned}
\|M_n - M\|_\infty &\leq \|\Psi_n(x) - (F_{X_T}(x) - F_{X_t}(x))\|_\infty \\
&\leq \left\|\widehat{F}_{X_T} - F_{X_T}\right\|_\infty + \left\|\widehat{F}_{X_t} - F_{X_t}\right\|_\infty \\
&\xrightarrow{p} 0.
\end{aligned}$$

Fix $\eta > 0$ and let $B = \{x \in I : |x - x_t^*| \geq \eta\}$. Because B is compact and M is continuous, $\sup_{x\in B} M(x) = \max_{x\in B} M(x) < M(x_t^*)$. We have

$$\sup_{x\in B} M_n(x) \leq \|M_n - M\|_\infty + \sup_{x\in B} M(x) \xrightarrow{p} \sup_{x\in B} M(x) < M(x_t^*). \tag{B.1}$$

Suppose that $\widehat{x}_t^* \in B$ and $x_t^* \in [\widehat{F}_{X_T}^{-1}(\underline{p}), \widehat{F}_{X_T}^{-1}(\overline{p})]$. Then $\sup_{x\in B} M_n(x) = M_n(\widehat{x}_t^*) \geq M_n(x_t^*)$. Hence,

$$P\left(\widehat{x}_t^* \in B, x_t^* \in [\widehat{F}_{X_T}^{-1}(\underline{p}), \widehat{F}_{X_T}^{-1}(\overline{p})]\right) \leq P\left(\sup_{x\in B} M_n(x) - M_n(x_t^*) \geq 0\right),$$

but the latter probability tends to zero in view of (B.1). Now, remark that $x_t^* \in (F_{X_T}^{-1}(\underline{p}), F_{X_T}^{-1}(\overline{p}))$, so that with a probability approaching one, $x_t^* \in [\widehat{F}_{X_T}^{-1}(\underline{p}), \widehat{F}_{X_T}^{-1}(\overline{p})]$. With probability approaching one, we also have $[\widehat{F}_{X_T}^{-1}(\underline{p}), \widehat{F}_{X_T}^{-1}(\overline{p})] \subset I$, so that $\widehat{x}_t^* \in I$ with probability approaching one. Hence, $P(|\widehat{x}_t^* - x_t^*| < \eta) \xrightarrow{p} 0$. □

Lemma 2 (Convergence Rate of $\widehat{x}_t^*$)**.** *If Assumptions 1, 7 and 8 hold, then* $\sqrt{n}\,(\widehat{x}_t^* - x_t^*) = O_p(1)$.

Proof. Let $\psi_x(u,v) = \mathbb{1}\{u \leq x\} - \mathbb{1}\{v \leq x\}$ and $\Psi(x) = E(\psi_x(X_T, X_t))$. Because the set of functions $(\mathbb{1}\{. \leq x\})_x$ is Donsker and by the conservation properties of Donsker classes, $\mathcal{F}_\delta = \{\psi_x : |x - x_t^*| < \delta\}$ is Donsker for any $\delta > 0$. Moreover, by independence between X_t and X_T,

$$\begin{aligned} E\left(\psi_x(X_T, X_t) - \psi_{x_t^*}(X_T, X_t)\right)^2 &= F_{X_T}(x) + F_{X_t}(x) - 2F_{X_T}(x)F_{X_t}(x) + F_{X_T}(x_t^*) + F_{X_t}(x_t^*) \\ &\quad - 2F_{X_T}(x_t^*)F_{X_t}(x_t^*) \\ &\quad - 2\left(F_{X_T}(x \wedge x_t^*) + F_{X_t}(x \wedge x_t^*) - 2F_{X_T}(x \wedge x_t^*)F_{X_t}(x \wedge x_t^*)\right). \end{aligned}$$

Therefore, by continuity of F_{X_t} and F_{X_T},

$$E\left[\left(\psi_x(X_T, X_t) - \psi_{x_t^*}(X_T, X_t)\right)^2\right] \to 0 \text{ as } x \to x_t^*$$

This and Lemma 1 above imply (see, e.g., van der Vaart, 1998, Lemma 19.24) that

$$\sqrt{n}\left[(\Psi_n(\widehat{x}_t^*) - \Psi(\widehat{x}_t^*)) - (\Psi_n(x_t^*) - \Psi(x_t^*))\right] = o_P(1). \tag{B.2}$$

Besides, $\Psi(x_t^*) = 0$ and by the central limit theorem, $\Psi_n(x_t^*) = O_p(1/\sqrt{n})$. Moreover, with probability approaching one, $|\Psi_n(\widehat{x}_t^*)| \leq |\Psi_n(x_t^*)|$, implying $\Psi_n(\widehat{x}_t^*) = O_p(1/\sqrt{n})$. Combined with (B.2), this yields

$$\begin{aligned} \sqrt{n}\left[\Psi(\widehat{x}_t^*) - \Psi(x_t^*)\right] &= -\sqrt{n}\left[\Psi_n(\widehat{x}_t^*) - \Psi_n(x_t^*)\right] + o_p(1) \\ &= O_p(1). \end{aligned} \tag{B.3}$$

By Assumption 8-(ii) and because $\widehat{x}_t^*$ is consistent by Lemma 1, we have, with probability approaching one, $|\Psi(\widehat{x}_t^*) - \Psi(x_t^*)| \geq C^R\,|\widehat{x}_t^* - x_t^*|$. This and (B.3) yields the desired result. □

In the following, we let $\mathcal{D}$ denote the sets of càdlàg functions on $\mathcal{Y}$. We also let $\mathcal{C}^1$ denote the subset of $\mathcal{D}$ of continuously differentiable functions, with positive derivative.

Lemma 3 (Hadamard differentiability of two useful maps)**.** *The map* $Q : (F_1, F_2) \mapsto F_1^{-1} \circ F_2(x)$ *is Hadamard differentiable, tangentially to the set of continuous functions, at any* $(F_{10}, F_{20}) \in \mathcal{D}^2$ *such that* F_{10} *is differentiable at* $F_{10}^{-1} \circ F_{20}(x)$*, with positive derivative at this point. The map* $R : (F_1, F_2, F_3) \mapsto F_1 \circ F_2^{-1} \circ F_3$ *is also Hadamard differentiable at any* $(F_{10}, F_{20}, F_{30}) \in \mathcal{C}^1 \times \mathcal{C}^1 \times \mathcal{D}$ *continuously differentiable functions tangentially to the set of continuous functions.*

Proof. Let $Q_1 : (F_1, F_2) \mapsto (F_1, F_2(x))$ and $Q_2 : (F, p) \mapsto F^{-1}(p)$, so that $Q = Q_2 \circ Q_1$. The map Q_1 is linear and continuous, and therefore Hadamard differentiable at any $(F_{10}, F_{20}) \in \mathcal{D}^2$. Let us prove that Q_2 is Hadamard differentiable at any $(F_0, p) \in \mathcal{D} \times (0, 1)$ such that F_0 is differentiable at $F_0^{-1}(p)$, with a corresponding positive derivative. We have to show that for any h_t converging uniformly to h continuous and $p_t \to p$, $\lim_{t\to 0}[(F_0 + th_t)^{-1}(p_t) - F_0^{-1}(p)]$ exists. By differentiability of F_0^{-1} at p, this is the case if $\lim_{t\to 0}[(F_0 + th_t)^{-1}(p_t) - F_0^{-1}(p_t)]$ exists. Now, an inspection of the proof of Lemma 21.3 of van der Vaart (1998) reveals that it still applies if we replace p by p_t, with $p_t \to p$. Hence, Q_2 is Hadamard differentiable tangentially to the set of continuous functions at (F_0, p). By applying the chain rule (see van der Vaart, 1998, Theorem 20.9), Q is Hadamard differentiable at any $(F_{10}, F_{20}) \in \mathcal{D}^2$ such that F_{10} is differentiable at $F_{10}^{-1} \circ F_{20}(x)$, with positive derivative at this point. The result for R is proved in de Chaisemartin & D'Haultfoeuille (2014, see the proof of Lemma S5). □

Lemma 4 (Convergence rate of $\widehat{q}_t(x)$)**.** *Suppose that Assumption 7 holds and* F_{X_t} *is differentiable at* $q_t(x)$ *with* $F'_{X_t}(q_t(x)) > 0$*. Then* $\widehat{q}_t(x) - q_t(x) = O_P(1/\sqrt{n})$*.*

Proof. We have $q_t(x) = F_{X_t}^{-1} \circ F_{X_T}(x)$ and $\widehat{q}_t(x) = \widehat{F}_{X_t}^{-1} \circ \widehat{F}_{X_T}(x)$. By the standard Donsker's theorem (see, e.g., van der Vaart, 1998, Theorem 19.3),

$$\sqrt{n}\left(\widehat{F}_{X_t} - F_{X_t}, \widehat{F}_{X_T} - F_{X_T}\right) \xrightarrow{d} (G_1 \circ F_{X_t}, G_2 \circ F_{X_T}),$$

where G_1 and G_2 are two independent standard Brownian bridges. Because $F'_{X_t}(q_t(x)) > 0$, Lemma 3 and the functional delta method (see, e.g., van der Vaart & Wellner, 1996, Lemma 3.9.4) ensure that $\sqrt{n}\,(\widehat{q}_t(x) - q_t(x))$ is asymptotically normal. The result follows. □

In the following, we let $w_\tau(y, x) = F_{Y_\tau|X_\tau}(y|x) f_{X_\tau}(x)$ for $\tau \in \{t, T\}$. Let us also denote by $\widehat{f}_{X_\tau}$ the kernel density estimator of f_{X_τ} and $\widehat{w}_\tau(y, x) = \widehat{F}_{Y_\tau|X_\tau}(y|x)\widehat{f}_{X_\tau}(x)$.

Lemma 5 (Behavior of some nonparametric estimators)**.** *Suppose that Assumptions 7 and 9-10 hold. Then, for any closed and bounded interval* $V \subset \mathcal{X}$ *and* $\tau \in \{t, T\}$*,*

$$\sqrt{nh_n}\left\| E\left[\widehat{w}_\tau(., x)\right] / E\left[\widehat{f}_{X_\tau}(x)\right] - F_{Y_\tau|X_\tau}(., x)\right\|_\infty \longrightarrow 0,$$

$$\sup_{x\in V} \left\|\partial_x \widehat{w}_\tau(., x)\right\|_\infty = O_P(1).$$

Proof. First, because $K(y) \geq 0$, $E\left[\widehat{w}_\tau(y,x)\right]/E\left[\widehat{f}_{X_\tau}(x)\right] \leq 1$ for all y. Thus,

$$\begin{aligned}
&\left\| E\left[\widehat{w}_\tau(.,x)\right]/E\left[\widehat{f}_{X_\tau}(x)\right] - F_{Y_\tau|X_\tau}(.,x)\right\|_\infty \\
\leq & \frac{1}{f_{X_\tau}(x)}\left[\left\| E\left[\widehat{w}_\tau(.,x)\right] - w(.,x)\right\|_\infty + \left| E\left[\widehat{f}_{X_\tau}(x)\right] - f_{X_\tau}(x)\right|\right]. \qquad \text{(B.4)}
\end{aligned}$$

We have

$$E\left[\widehat{f}_{X_\tau}(x)\right] - f_{X_\tau}(x) = \int K(t)\left[f_{X_\tau}(x+th_n) - f_{X_\tau}(x)\right]dt.$$

Thus, because $|f'_{X_\tau}|$ is bounded,

$$\sqrt{nh_n}\left| E\left[\widehat{f}_{X_\tau}(x)\right] - f_{X_\tau}(x)\right| \leq C\sqrt{nh_n^5}\int |t|K(t)dt,$$

for some $C > 0$. Hence, the left-hand side tends to zero by Assumption 10-(i). Now consider the first term of (B.4). A change of variable yields

$$E\left[\widehat{w}_\tau(y,x)\right] - w(y,x) = \int K(t)\left[w(y, x - h_n t) - w(y,x)\right]dt.$$

By a second-order Taylor expansion, we obtain

$$E\left[\widehat{w}_\tau(y,x)\right] - w(y,x) = \int K(t)\left[-h_n t\partial_x w(y,x) + \frac{1}{2}(h_n t)^2 \partial_{xx} w(y, \widetilde{x}_t)\right]dt,$$

where $\widetilde{x}_t \in (x, x + h_n t)$. As a result, by Assumption 9-(ii) and 10-(ii),

$$\left\| E\left[\widehat{w}_\tau(.,x)\right] - w(.,x)\right\|_\infty \leq C' h_n^2,$$

for some $C' > 0$. By Assumption 10-(i) once more, the first term of (B.4) tends to zero, which yields the first result of the lemma.

To obtain the second result, first observe that by the triangular inequality,

$$\begin{aligned}
\sup_{x\in V}\left\|\partial_x \widehat{w}_\tau(.,x)\right\|_\infty \leq & \sup_{x\in V}\left\|\partial_x \widehat{w}_\tau(.,x) - E\left[\widehat{w}_\tau(.,x)\right]\right\|_\infty \\
& + \sup_{x\in V}\left\| E\left[\widehat{w}_\tau(.,x)\right] - \partial_x w(.,x)\right\|_\infty + \sup_{x\in V}\left\|\partial_x w(.,x)\right\|_\infty. \qquad \text{(B.5)}
\end{aligned}$$

By Assumption 9-(iii) $\sup_{x\in V}\left\|\partial_x w(.,x)\right\|_\infty < \infty$. Therefore, to show the result, it suffices to show that the two first terms of the right-hand side of (B.5) tend to zero in probability.

To analyse the first term, let us remark that

$$nh_n^2\left(\partial_x \widehat{w}_\tau(y,x) - E[\partial_x \widehat{w}_\tau(y,x)]\right) = \sum_{i=1}^n \mathbb{1}\{Y_{\tau i} \leq y\}K'\left(\frac{x - X_{\tau i}}{h}\right) - nE\left[\mathbb{1}\{Y_{\tau i} \leq y\}K'\left(\frac{x - X_{\tau i}}{h}\right)\right].$$

Thus, the left-hand side corresponds to $W(x, f)$ in Einmahl & Mason (2000), with $f(u) = \mathbb{1}\{y \leq u\}$ and K' in place of K. Moreover, f_{X_τ, Y_τ} is continuous, f_{X_τ} is continuous and $\inf_{x \in V} f_{X_\tau}(x) > 0$ and K' satisfies their (K)-(i) and (K)-(ii). Finally, remark that Proposition 1 of Einmahl & Mason (2000) does not rely on their condition (K)-(iii). Hence, with probability one,

$$\limsup_{n\to\infty} \sqrt{\frac{nh_n^3}{2|\log(h_n)|}} \sup_{x \in V} \|\partial_x \widehat{w}_\tau(., x) - E[\partial_x \widehat{w}_\tau(., x)]\|_\infty < \infty.$$

Because $nh_n^3/|\log(h_n)| \to \infty$ by Assumption 10, $\sup_{x\in V} \|\partial_x \widehat{w}_\tau(., x) - E[\partial_x \widehat{w}_\tau(., x)]\|_\infty \to 0$.

Now let us turn to the second term of (B.5). First, remark that

$$E\left[\partial_x \widehat{w}_\tau(y, x)\right] = \frac{1}{h_n^2} \int w(y, x) K'\left(\frac{x-u}{h_n}\right) du.$$

Integrating by part and using the facts that f_{X_t} is bounded above and $K(u) \to 0$ as $|u| \to \infty$, we obtain

$$E\left[\partial_x \widehat{w}_\tau(y, x)\right] = \int K(t) \partial_x w(y, x + h_n t) dt.$$

By Assumption 9-(iii), there exists a constant $C' > 0$ such that for all y and $x \in V$, $|\partial_x w(y, x + h_n t) - \partial_x w(y, x + h_n t)| \leq C' h_n |t|$. Hence,

$$\sup_{x\in V} \left\|E\left[\partial_x \widehat{w}_\tau(., x)\right] - \partial_x w(., x)\right\|_\infty \leq C' h_n \int |t| K(t) dt,$$

and the left-hand side tends to zero. □

Lemma 6 (Negligible effect of estimating covariates)**.** *Suppose that $x \in \mathcal{X}$ and $\widehat{x}$ satisfies $\widehat{x} - x = O_P(1/\sqrt{n})$. If Assumptions 7 and 9-10 hold, then, for $\tau \in \{t, T\}$,*

$$\sqrt{nh_n} \left\| \widehat{F}_{Y_\tau|X_\tau}(.|\widehat{x}) - \widehat{F}_{Y_\tau|X_\tau}(.|x) \right\|_\infty \xrightarrow{P} 0.$$

Proof. Let us denote by $\widehat{f}_{X_\tau}$ the kernel density estimator of f_{X_τ} and $\widehat{w}_\tau(y|x) = \widehat{F}_{Y_\tau|X_\tau}(y|x)\widehat{f}_{X_\tau}(x)$. With a large probability, $\widehat{x} \in V$. Then, using the fact that $\widehat{F}_{Y_\tau|X_\tau} \leq 1$,

$$\begin{aligned}
&\left\| \widehat{F}_{Y_\tau|X_\tau}(.|\widehat{x}) - \widehat{F}_{Y_\tau|X_\tau}(.|x) \right\|_\infty \\
\leq & \frac{1}{\inf_{x\in V} \widehat{f}_{X_\tau}(x)} \left[\|\widehat{w}_\tau(.|\widehat{x}) - \widehat{w}_\tau(.|x)\|_\infty + \left| \widehat{f}_{X_\tau}(\widehat{x}) - \widehat{f}_{X_\tau}(x) \right| \right] \\
\leq & \frac{1}{\inf_{x'\in V} \widehat{f}_{X_\tau}(x')} \left[\sup_{x'\in V} \|\partial_x \widehat{w}_\tau(.|x')\|_\infty | + \sup_{x'\in V} \left| \widehat{f}'_{X_\tau}(x') \right| \right] |\widehat{x} - x| .
\end{aligned}$$

Now, f_{X_τ} and f'_{X_τ} are uniformly continuous on V. By Assumption 10, $h_n \to 0$ and $nh_n^3/|\log(h_n)| \to \infty$. Moreover, K satisfies the conditions of Theorem A and C of Silverman (1978). K' may not satisfy condition (C2) of Silverman (1978), but this condition is not needed for the necessity part of his Theorem 3 that we use here. Therefore, $\widehat{f}_{X_\tau}$ and $\widehat{f}'_{X_\tau}$ are uniformly consistent on V. The result follows by $\widehat{x} - x = O_P(1/\sqrt{n})$, $h_n \to 0$ and Lemma 5. $\square$

Lemma 7 (Asymptotic distribution of $\widehat{F}_{Y_T|X_T}(.|x^*_t)$)**.** *If Assumptions 7 and 9-10 hold, then, for $\tau \in \{t, T\}$,*

$$\sqrt{nh_n}\left(\widehat{F}_{Y_T|X_T}(.|x) - F_{Y_T|X_T}(.|x), \widehat{F}_{Y_t|X_t}(.|q_t(x) - F_{Y_t|X_t}(.|q_t(x)), \widehat{F}_{Y_T|X_T}(.|x^*_T) - F_{Y_T|X_T}(.|x^*_t),\right.$$
$$\left.\widehat{F}_{Y_t|X_t}(.|x^*_t) - F_{Y_t|X_t}(.|x^*_t)\right) \xrightarrow{d} \mathbb{G},$$

where $\mathbb{G}$ is a continuous gaussian processes.

Proof. First, by Lemma 5, we have, for any $x \in \mathcal{X}$,

$$\left\| E\left[\widehat{F}_{Y_T|X_T}(.|x)\right] - \widehat{F}_{Y_T|X_T}(.|x)\right\|_\infty \leq C|h_n|,$$

for some $C > 0$. Hence, we may focus on the process $\mathbb{G}_n = \sqrt{nh_n}\left(\widehat{F}_{Y_T|X_T}(.|x) - E\left[\widehat{F}_{Y_T|X_T}(.|x)\right]\right)$. The proof readily extends to the multivariate process by the Cramér-Wold device. Note that convergence of the process follows if (i) for any $k \in \mathbb{N}$ and $(y_1, .., y_k) \in \mathcal{Y}^k$, $(\mathbb{G}_n(y_1), ..., \mathbb{G}_n(y_k))$ is asymptotically normal and (ii) $\mathbb{G}_n$ is asymptotically tight(see, e.g., van der Vaart, 1998, Theorem 18.14). (i) follows by the Cramér-Wold device, asymptotic normality of the Nadaraya-Watson estimator and Assumptions 9-10 (see, e.g., Bierens, 1987).

Now, let us prove (ii). By Theorem 1.1 of Einmahl & Mason (1997), the process $\sqrt{nh_n}\left(\widehat{w}_T(., x) - E[\widehat{w}_T(., x)]\right)$ is asymptotically tight. Now, remark that

$$\mathbb{G}_n = \frac{1}{f_{X_T(x)}}\left[\sqrt{nh_n}\left(\widehat{w}_T(., x) - w_T(., x)\right) + F_{Y_T|X_T}(.|x)\sqrt{nh_n}\left(\widehat{f}_{X_T}(x) - f_{X_T}(x)\right)\right.$$
$$\left. + \left(\widehat{F}_{Y_T|X_T}(.|x) - F_{Y_T|X_T}(.|x)\right)\sqrt{nh_n}\left(\widehat{f}_{X_T}(x) - f_{X_T}(x)\right)\right].$$

By Assumption 10, K is defined on a compact set and has bounded variation. Theorem 1 of Stute (1986, see also his remark p.893) then ensures that $\widehat{F}_{Y_T|X_T}(.|x)$ is a uniformly consistent estimator of $F_{Y_T|X_T}(.|x)$. Hence, the supremum norm of the third term in the brackets converges to zero in probability. The second term is asymptotically tight since $\sqrt{nh_n}\left(\widehat{f}_{X_T}(x) - f_{X_T}(x)\right) = O_P(1)$ and $F_{Y_T|X_T}(.|x)$ is uniformly continuous on $\mathcal{Y}$. Hence, $\mathbb{G}_n$ is asymptotically tight, and the result follows. $\square$

B.2 Proof of Theorem 6

Let $H(y) = F_{Y_t|X_t}\left(F^{-1}_{Y_T|X_T}(F_{Y_t|X_t}(y|x^*_t)|x^*_t)|q_t(x)\right)$ and $\widehat{H}(y) = \widehat{F}_{Y_t|X_t}\left(\widehat{F}^{-1}_{Y_T|X_T}(\widehat{F}_{Y_t|X_t}(y|\widehat{x}^*_t)|\widehat{x}^*_t)|\widehat{q}_t(x)\right)$. It is easy to see that H is the cumulative distribution function of $m_t(Y_t)$ conditional on

$X_t = q_t(x)$. Lemmas 6 and 7 imply that $\left(\widehat{F}_{Y_T|X_T}(.|x), \widehat{F}_{Y_t|X_t}(.|\widehat{q}_t(x)), \widehat{F}_{Y_T|X_T}(.|\widehat{x}_t^*), \widehat{F}_{Y_t|X_t}(.|\widehat{x}_t^*)\right)$ converges to a continuous gaussian process. By Lemma 3 and the functional delta method, $\left(\widehat{F}_{Y_T|X_T}(.|x), \widehat{H}\right)$ also converges to a continuous gaussian process at the rate $\sqrt{nh_n}$.

Now, by integration by parts for Lebesgue-Stieljes integrals,

$$\Delta^{ATT}(x, q_t(x)) = \int_{\underline{y}}^{\overline{y}} F_{Y_T|X_T}(y|x) - H(y)dy.$$

The map $\varphi : (F_1, F_2) \mapsto \int_{\underline{y}}^{\overline{y}} [F_1(y) - F_2(y)]dy$, defined on the set of bounded càdlàg functions, is linear and also continuous with respect to the supremum norm. It is therefore Hadamard differentiable. Because $\widehat{\Delta}^{ATT}(x, q_t(x)) = \varphi\left(\widehat{F}_{Y_T|X_T}(.|x), \widehat{H}\right)$, it is asymptotically normal at the rate $\sqrt{nh_n}$.

Regarding the quantile treatment effect, we have $\Delta^{QTT}(p, x, q_t(x)) = H^{-1}(p) - F^{-1}_{Y_T|X_T}(p|x)$ and $\widehat{\Delta}^{QTT}(p, x, q_t(x)) = \widehat{H}^{-1}(p) - \widehat{F}^{-1}_{Y_T|X_T}(p|x)$. Because the quantile function is Hadamard differentiable (see for instance Lemma 21.3 of van der Vaart, 1998), the map $(F_1, F_2) \mapsto F_1^{-1}(p) - F_2^{-1}(p)$ is Hadamard differentiable at any (F_{10}, F_{20}) such that F_{10} and F_{20} are differentiable at $F_{10}^{-1}(p)$ and $F_{20}^{-1}(p)$ respectively, with positive corresponding derivatives. The result follows by applying the functional delta method once more.

C Details on the application

The birth weight Y_t can be written as $Y_t = (1 - D_t)Y_{0t} + D_tY_{1t}$, where Y_{1t} (resp. Y_{0t}) is the potential birth weight if premature (resp. not premature) and D_t is the dummy of being premature. We have also $Y_{dt} = m_{dt} \circ g_d(X_t, A_t)$ and $D_t = h_t(X_t, B_t)$, where B_t denotes unobserved terms. We are then interested in the marginal effect of X_t on Y_t, $\text{ME}_t(x, x)$, with

$$\begin{aligned}\text{ME}_t(x, x') &= \frac{\partial}{\partial x} E[g_0(x, A_t) + h_t(x, B_t)(g_1 - g_0)(x, A_t)|X_t = x'] \\ &= \frac{\partial}{\partial x} E[g_0(x, A_t)|X_t = x'] + \frac{\partial}{\partial x} E[h_t(x, B_t)(g_1 - g_0)(x, A_t)|X_t = x'] \qquad \text{(C.1)}\end{aligned}$$

Let us suppose that $B_t \perp\!\!\!\perp (A_t, X_t)$. Then

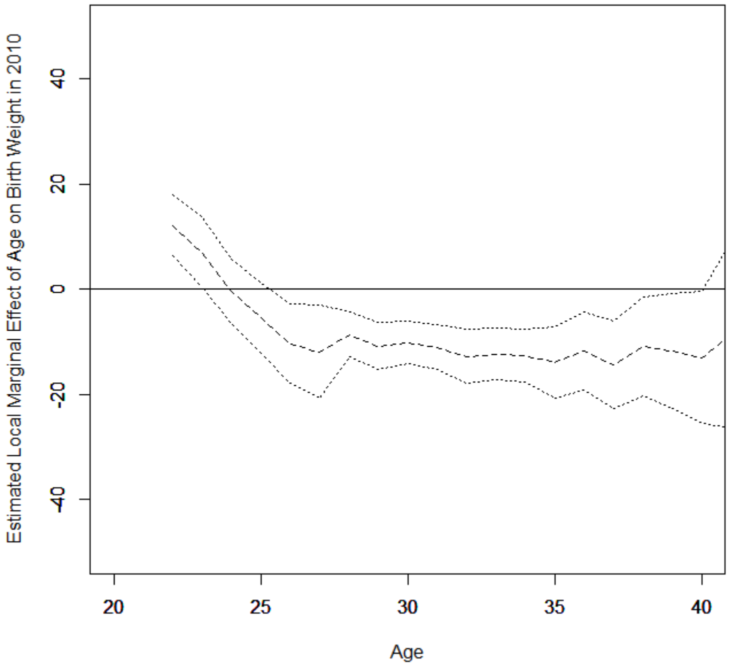
$$\frac{\partial}{\partial x} E[g_0(x, A_t)|X_t = x'] = \frac{\partial}{\partial x} E[g_0(x, A_t)|D_t = 0, X_t = x'],$$

which is already identified. Besides,

$$\begin{aligned}\frac{\partial}{\partial x} E[h_t(x, B_t)(g_1 - g_0)(x, A_t)|X_t = x'] &= \frac{\partial}{\partial x} E[h_t(x, B_t)|X_t = x']E[(g_1 - g_0)(x, A_t)|X_t = x'] \\ &= \frac{\partial}{\partial x} E[h_t(x, B_t)|X = x']E[(g_1 - g_0)(x, A_t)|X_t = x'] \\ &\quad + E[h_t(x, B_t)|X_t = x'] \frac{\partial}{\partial x} E[(g_1 - g_0)(x, A_t)|X_t = x']. \quad \text{(C.2)}\end{aligned}$$

Moreover, by independence between B_t and X_t, $\partial/\partial x E[h_t(x, B_t)|X_t = x')]$ taken at $x = x'$ is simply $\partial/\partial x E(D_t|X_t = x)$. Finally, suppose that $(g_1 - g_0)(x, A_t)$ does not depend on A_t. Then we can identify $E[(g_1 - g_0)(x, A_t)|X_t = x']$ taken at $x = x'$ by $E(Y_t|D_t = 1, X_t = x) - E(Y_t|D_t = 0, X_t = x)$. The derivative of this function with respect to x is identified similarly. We finally obtain $\mathrm{ME}_t(x, x)$ by using (C.1).

The results are displayed in Figure 9. The additional effect due to preterm birth is slightly negative, but the results are essentially unchanged. As expected, the standard errors are slightly larger, as we have to account for uncertainty stemming from preterm birth as well.



Notes: Effects for first birth in 2010, with bootstrap confidence intervals.

Figure 9: Marginal effects of mother's age on infant birth weights including preterm birth.

References

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

Abadie, A., Angrist, J. & Imbens, G. W. (2002), ‘Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings’, *Econometrica* **70**, 91–117.

Altonji, J. G. & Matzkin, R. L. (2005), ‘Cross section and panel data estimators for nonseparable models with endogenous regressors’, *Econometrica* **73**, 1053–1102.

Arellano, M. & Bonhomme, S. (2012), ‘Identifying distributional characteristics in random coefficients panel data models’, *Review of Economic Studies* **79**, 987–1020.

Ashenfelter, O. & Card, D. (1985), ‘Using the longitudinal structure of earnings to estimate the effect of training programs’, *Review of Economics and Statistics* **67**, 648–660.

Athey, S. & Imbens, G. W. (2006), ‘Identification and inference in nonlinear difference-in-differences models’, *Econometrica* **74**, 431–497.

Behrman, J. R. & Rosenzweig, M. R. (2004), ‘The returns to birth weight’, *Review of Economics and Statistics* **86**, 586–601.

Behrman, J. R., Rosenzweig, M. R. & Taubman, P. (1994), ‘Endowments and the allocation of schooling in the family and in the marriage market: The twins experiment’, *Journal of Political Economy* **102**, 1131–1174.

Bhattacharya, D. (2008), ‘Inference in panel data models under attrition caused by unobservables’, *Journal of Econometrics* **144**, 430–446.

Bierens, H. J. (1987), Kernel estimators of regression functions, *in* ‘Advances in econometrics: Fifth world congress’, Vol. 1, pp. 99–144.

Black, S. E., Devereux, P. J. & Salvanes, K. G. (2007), ‘From the cradle to the labor market? the effect of birth weight on adult outcomes’, *Quarterly Journal of Economics* **122**, 409–439.

Camacho, A. (2008), ‘Stress and birth weight: Evidence from terrorist attacks’, *American Economic Review, Papers and Proceedings* **98**, 511–515.

Chamberlain, G. (1982), ‘Multivariate regression models for panel data’, *Journal of Econometrics* **18**, 5–46.

Chamberlain, G. (1984), Panel data, *in* Z. Griliches & M. D. Intriligator, eds, ‘Handbook of econometrics’, Vol. 2, Elsevier, chapter 22, pp. 1247–1318.

Chernozhukov, V., Fernandez-Val, I., Hahn, J. & Newey, W. (2013), ‘Average and quantile effects in non separable panel data models’, *Econometrica* **81**, 535–580.

Collado, D. M. (1997), ‘Estimating dynamic models from time series of independent cross-sections’, *Journal of Econometrics* **82**, 37–62.

Corman, H., Joyce, T. J. & Grossman, M. (1987), ‘Birth outcome production functions in the u. s.’, *Journal of Human Resources* **22**, 339–360.

Currie, J. & Hyson, R. (1999), ‘Is the impact of health shocks cushioned by socioeconomic status? the case of low birth weight’, *American Economic Review, Papers and Proceedings* **89**, 245–250.

Currie, J. & Moretti, E. (2003), ‘Mother’s education and the intergenerational transmission of human capital: Evidence from college openings’, *Quarterly Journal of Economics* **118**, 1495–1532.

Das, M. (2004), ‘Simple estimators for nonparametric panel data models with sample attrition’, *Journal of Econometrics* **120**, 159–180.

de Chaisemartin, C. & D’Haultfoeuille, X. (2014), Fuzzy changes-in-changes. Working paper.

Deaton, A. (1985), ‘Panel data from time series of cross sections’, *Journal of Econometrics* **30**, 109–126.

Devereux, P. J. (2007), ‘Small-sample bias in synthetic cohort models of labor supply’, *Journal of Applied Econometrics* **22**, 839–848.

D’Haultfoeuille, X. & Février, P. (2015), ‘Identification of nonseparable triangular models with discrete instruments’, *Econometrica* **Forthcoming**.

Einmahl, U. & Mason, D. M. (1997), ‘Gaussian approximation of local empirical processes indexed by functions’, *Probability Theory and Related Fields* **107**, 283–311.

Einmahl, U. & Mason, D. M. (2000), ‘An empirical process approach to the uniform consistency of kernel-type function estimators’, *Journal of Theoretical Probability* **13**(1), 1–37.

Evans, W. N. & Ringel, J. S. (1999), ‘Can higher cigarette taxes improve birth outcomes?’, *Journal of Public Economics* **72**, 133–154.

Evdokimov, K. (2011), Nonparametric identification of a nonlinear panel model with application to duration analysis with multiple spells. Working paper.

Florens, J., Heckman, J. J., Meghir, C. & Vytlacil, E. (2008), 'Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects', *Econometrica* **76**, 1191–1206.

Geronimus, A. T. & Korenman, S. (1992), 'The socioeconomic consequences of teen childbearing reconsidered', *Quarterly Journal of Economics* **107**, 1187–1214.

Graham, B. S. & Powell, J. L. (2012), 'Identification and estimation of average partial effects in 'irregular' correlated random coefficient panel data models', *Econometrica* **80**, 2105–2152.

Grossman, M. & Joyce, T. J. (1990), 'Unobservables, pregnancy resolutions, and birth weight production functions in new york city', *Journal of Political Economy* **98**, 983–1007.

Hausman, J. A. & Wise, D. A. (1979), 'Attrition bias in experimental and panel data: the Gary income maintenance experiment', *Econometrica* **47**, 455–473.

Heckman, J. J., Ichimura, H. & Todd, P. E. (1997), 'Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme', *Review of Economic Studies* **64**, 605654.

Heckman, J. & Vytlacil, E. J. (1998), 'Instrumental variables methods for the correlated random coefficient model: Estimating the average return to schooling when the return is correlated with schooling', *Journal of Human Resources* **33**, 974–987.

Heffner, L. J. (2004), 'Advanced maternal age - how old is too old?', *The New England Journal of Medecine* **4**, 1927–1929.

Hirano, K., Imbens, G. W., Ridder, G. & Rubin, D. B. (2001), 'Combining panel data sets with attrition and refreshment samples', *Econometrica* **69**, 1645–1659.

Hoderlein, S. & Mammen, E. (2007), 'Identification of marginal effects in nonseparablle models without monotonicity', *Econometrica* **75**, 1513–1518.

Hoderlein, S. & Sasaki, Y. (2013), Outcome conditioned treatment effects. Working Paper.

Hoderlein, S. & White, H. (2012), 'Nonparametric identification in nonseparable panel data models with generalized fixed effects', *Journal of Econometrics* **168**, 300–314.

Hofferth, S. (1998), 'Long-term economic consequences for women of delayed childbearing and reduced family size', *Demography* **21**, 141–155.

Honore, B. (1992), 'Trimmed lad and least squares estimation of truncated and censored regression models with fixed effects', *Econometrica* **60**, 533–565.

Imbens, G. W. & Newey, W. K. (2009), 'Identification and estimation of triangular simultaneous equations models without additivity', *Econometrica* **77**, 1481–1512.

Kyriazidou, E. (1997), 'Estimation of a panel data sample selection model', *Econometrica* **65**, 1335–1364.

Manski, C. F. (1987), 'Semiparametric analysis of random effects linear models from binary panel data', *Econometrica* **55**, 357–362.

McKenzie, D. J. (2004), 'Asymptotic theory for heterogeneous dynamic pseudo-panels', *Journal of Econometrics* **120**, 235–262.

Moffitt, R. (1993), 'Identification and estimation of dynamic models with a time series of repeated cross sections', *Journal of Econometrics* **59**, 99–123.

Moffitt, R. & Ridder, G. (2007), Econometrics of data combination, *in* J. J. Heckman & E. E. Lleamer, eds, 'Hanbook of Econometrics', Elsevier.

Murtazashvili, I. & Wooldridge, J. (2008), 'Fixed effects instrumental variables estimation in correlated random coefficient panel data models', *Journal of Econometrics* **142**, 539–552.

Rosen, A. (2012), 'Set identification via quantile restrictions in short panels', *Journal of Econometrics* **166**, 127–137.

Rosenzweig, M. R. & Schultz, T. P. (1983), 'Estimating a household production function: Heterogeneity and the demand for health inputs, and their effects on birth weight', *Journal of Political Economy* **91**, 723–746.

Rosenzweig, M. R. & Wolpin, K. I. (1991), 'Inequality at birth: The scope for policy intervention', *Journal of Econometrics* **50**, 205–228.

Rosenzweig, M. R. & Wolpin, K. I. (1995), 'Sisters, siblings, and mothers: The effects of teen-age childbearing on birth outcomes', *Econometrica* **63**, 303–326.

Sasaki, Y. (2013), Heterogeneity and selection in dynamic panel data. Working paper.

Schennach, S., White, H. & Chalak, K. (2012), 'Local indirect least squares and average marginal effects in nonseparable structural systems', *Journal of Econometrics* **166**, 282–302.

Silverman, B. W. (1978), 'Weak and strong uniform consistency of the kernel estimate of a density and its derivatives', *The Annals of Statistics* **6**, 177–184.

Stotland, N., Canghey, A., Breed, E. & Escobar, G. (2004), 'Risk factors and obsteric complications associated with macrosomia', *International Journal of Gynecology and Obstetrics* **87**, 220–226.

Stute, W. (1986), 'On almost sure convergence of conditional empirical distribution functions', *The Annals of Probability* pp. 891–901.

Torgovitsky, A. (2015), Identification of nonseparable models with general instruments.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics.

van der Vaart, A. W. & Wellner, J. A. (1996), *Weak convergence and Empirical Processes*, Springer.

Verbeek, M. (1996), Pseudo panel data, *in* L. Matyas & P. Sevestre, eds, 'Econometrics of Panel Data', Kluwer.

Verbeek, M. & Nijman, T. (1992), 'Can cohort data be treated as genuine panel data?', *Empirical Economics* **17**, 9–23.

Verbeek, M. & Nijman, T. (1993), 'Minimum mse estimation of a regression model with fixed effects from a series of cross sections', *Journal of Econometrics* **59**, 125–136.

Wooldridge, J. (2003), 'Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model', *Economics Letters* **79**, 185–191.

Wooldridge, J. (2005), 'Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models.', *The Review of Economics and Statistics* **87**, 385–390.