Review of Economic Studies (2005) **72**, 1–19 0034-6527/05/00010000001$02.00

Semiparametric Difference-in-Differences Estimators

ALBERTO ABADIE
Harvard University and NBER

First version received June 2001; *final version accepted October* 2003 (*Eds.*)

The difference-in-differences (DID) estimator is one of the most popular tools for applied research in economics to evaluate the effects of public interventions and other treatments of interest on some relevant outcome variables. However, it is well known that the DID estimator is based on strong identifying assumptions. In particular, the conventional DID estimator requires that, in the absence of the treatment, the average outcomes for the treated and control groups would have followed parallel paths over time. This assumption may be implausible if pre-treatment characteristics that are thought to be associated with the dynamics of the outcome variable are unbalanced between the treated and the untreated. That would be the case, for example, if selection for treatment is influenced by individual-transitory shocks on past outcomes (Ashenfelter's dip). This article considers the case in which differences in observed characteristics create non-parallel outcome dynamics between treated and controls. It is shown that, in such a case, a simple two-step strategy can be used to estimate the average effect of the treatment for the treated. In addition, the estimation framework proposed in this article allows the use of covariates to describe how the average effect of the treatment varies with changes in observed characteristics.

A good way to do econometrics is to look for good natural experiments *and* use statistical methods that can tidy up the confounding factors that nature has not controlled for us. (Daniel McFadden, *Econometric Tools*)

1. INTRODUCTION

The use of natural experiments to evaluate treatment effects in the absence of truly experimental data has gained wide acceptance in empirical research in economics and other social sciences. Simple comparisons of pre-treatment and post-treatment outcomes for those individuals exposed to a treatment are likely to be contaminated by temporal trends in the outcome variable or by the effect of events, other than the treatment, that occurred between both periods. However, when only a fraction of the population is exposed to the treatment, an untreated comparison group can be used to identify temporal variation in the outcome that is not due to treatment exposure. The difference-in-differences (DID) estimator is based on this simple idea. Card and Krueger (1994) assess the employment effects of a raise in the minimum wage in New Jersey using a neighbouring state, Pennsylvania, to identify the variation in employment that New Jersey would have experienced in the absence of a raise in the minimum wage. Other applications of DID include studies of the effects of immigration on native wages and employment (Card, 1990), the effects of temporary disability benefits on time out of work after an injury (Meyer, Viscusi and Durbin, 1995), and the effect of anti-takeover laws on firms' leverage (Garvey and Hanka, 1999).

It is well known that the conventional DID estimator is based on strong assumptions. In particular, the conventional DID estimator requires that in absence of the treatment, the

average outcomes for treated and controls would have followed parallel paths over time. This assumption may be implausible if pre-treatment characteristics that are thought to be associated with the dynamics of the outcome variable are unbalanced between the treated and the untreated group.

This study considers the case in which differences in observed characteristics create non-parallel outcome dynamics for the treated and untreated groups. It is shown that, in such a case, a simple two-step strategy can be used to estimate the average effect of the treatment for the treated. In addition, the estimation framework proposed in this article allows the use of covariates to describe how the average effect of the treatment varies with changes in observed characteristics.

Despite the prolific literature on semiparametric and non-parametric methods, few articles have been devoted to studying and relaxing the DID identification restrictions. Some exceptions are Besley and Case (1994), Meyer (1995), Heckman, Ichimura and Todd (1997), Imbens, Liebman and Eissa (1997), Heckman, Ichimura, Smith and Todd (1998), Angrist and Krueger (1999), Blundell and MaCurdy (1999), Blundell, Costa Dias, Meghir and van Reenen (2001) and Athey and Imbens (2002).

The identification procedure used in this article originated with Heckman *et al.* (1997, 1998). However, the estimation procedure differs from the earlier literature in three ways. First, it does not require repeated observations for the same individuals. The proposed estimators are feasible under the data requirements for traditional DID estimators when applied to repeated cross-sections. Second, it allows the estimation of parsimonious parametric approximations to the average effect of the treatment on the treated conditional on selected covariates of interest. Finally, the framework can accommodate multilevel treatment variables (that is, different treatment intensities).

The rest of the article is organized as follows. Section 2 describes the conventional DID model and discusses some of its limitations. Section 3 presents the main identification results of the article, followed by some extensions. The estimation strategy along with asymptotic distribution theory is provided in Section 4. Section 5 concludes. Proofs are presented in the Appendix.

2. THE DID ESTIMATOR

The basic DID framework can be described as follows. Let $Y(i,t)$ be the outcome of interest for individual i at time t. The population is observed in a pre-treatment period $t = 0$, and in a post-treatment period $t = 1$. Between these two periods, some fraction of the population is exposed to the treatment. We denote $D(i,t) = 1$ if individual i has been exposed to the treatment previous to period t, $D(i,t) = 0$ otherwise. We call those individuals with $D(i,1) = 1$ *treated*, and those with $D(i,1) = 0$ *controls* (or *untreated*). Since individuals are only exposed to treatment after the first period, $D(i,0) = 0$ for all i.

The conventional DID estimator is often derived using a linear parametric model. It is useful to consider this formulation of the DID model first, to fix ideas, before studying non-parametric identification in Section 3. The following formulation of the DID model is based on that given in Ashenfelter and Card (1985). Suppose that the outcome variable is generated by a components of variance process

$$Y(i,t) = \delta(t) + \alpha \cdot D(i,t) + \eta(i) + \upsilon(i,t), \tag{1}$$

where $\delta(t)$ is a time-specific component, α represents the impact of the treatment, $\eta(i)$ is an individual-specific component, and $\upsilon(i,t)$ is an individual-transitory shock that has mean zero at each period, $t = 0, 1$, and is possibly correlated in time. Only $Y(i,t)$ and $D(i,t)$ are observed. The effect of the treatment, α, is not identified without further restrictions. A sufficient condition

for identification is that selection for treatment does not depend on the individual-transitory shocks, that is

$$P(D(i,1)=1 \mid \upsilon(i,t)) = P(D(i,1)=1) \tag{2}$$

for $t = 0, 1$. Adding and subtracting $E[\eta(i) \mid D(i,1)]$ in equation (1), we obtain

$$Y(i,t) = \delta(t) + \alpha \cdot D(i,t) + E[\eta(i) \mid D(i,1)] + \varepsilon(i,t), \tag{3}$$

where $\varepsilon(i,t) = \eta(i) - E[\eta(i) \mid D(i,1)] + \upsilon(i,t)$. Notice that $\delta(t) = \delta(0) + (\delta(1) - \delta(0))t$, and $E[\eta(i) \mid D(i,1)] = E[\eta(i) \mid D(i,1) = 0] + (E[\eta(i) \mid D(i,1) = 1] - E[\eta(i) \mid D(i,1) = 0])D(i,1)$. Let $\mu = E[\eta(i) \mid D(i,1) = 0] + \delta(0)$, $\tau = E[\eta(i) \mid D(i,1) = 1] - E[\eta(i) \mid D(i,1) = 0]$ and $\delta = \delta(1) - \delta(0)$. We obtain

$$Y(i,t) = \mu + \tau \cdot D(i,1) + \delta \cdot t + \alpha \cdot D(i,t) + \varepsilon(i,t). \tag{4}$$

The restriction in equation (2) for $t = 0, 1$ implies $E[(1, D(i,1), t, D(i,t)) \cdot \varepsilon(i,t)] = 0$, so all the parameters in equation (4), including the treatment impact α, are estimable by least squares. Notice that the model allows *any* kind of dependence between selection for treatment, $D(i,1) = 1$, and the individual-specific component, $\eta(i)$. This model is called "difference-in-differences" because under the identifying condition in equation (2) we have

$$\begin{aligned}\alpha = &\{E[Y(i,1) \mid D(i,1)=1] - E[Y(i,1) \mid D(i,1)=0]\} \\ &- \{E[Y(i,0) \mid D(i,1)=1] - E[Y(i,0) \mid D(i,1)=0]\},\end{aligned} \tag{5}$$

and the least squares estimator of α is the sample counterpart of equation (5).

This formulation of the problem is useful when repeated cross sections of $(Y(i,t), D(i,1))$ for $t = 0, 1$ are available. If a sample with repeated pre-treatment and post-treatment observations of the outcome variable, $Y(i,1)$ and $Y(i,0)$, is available, then α is estimable by least squares regression of $Y(i,1) - Y(i,0)$ on $D(i,1)$:

$$\alpha = E[Y(i,1) - Y(i,0) \mid D(i,1)=1] - E[Y(i,1) - Y(i,0) \mid D(i,1)=0].$$

Note that equation (2) for $t = 0, 1$ implies that $\upsilon(i,1) - \upsilon(i,0)$ is mean independent of $D(i,1)$, and therefore that, in absence of the treatment, the average outcome for the treated would have experienced the same variation as the average outcome for the untreated. This restriction, implied by the model, may be too stringent if treated and controls are unbalanced in covariates that are thought to be associated with the dynamics of the outcome variable.

For example, it has been documented that participants in training programmes experience a decline in earnings prior to the training period (Ashenfelter's dip, Ashenfelter (1978)). This fact suggests that selection for training may be affected by individual-transitory shocks in pre-training earnings. To accommodate this conjecture, Ashenfelter and Card (1985) propose the following model for the selection process:

$$D(i,1) = \begin{cases} 1 & \text{if } Y(i, 1-\kappa) + u(i) < \bar{Y} \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where κ is a positive integer, $\bar{Y}$ is a constant, and $u(i)$ is a random variable independent of any variance component. Under this formulation for the selection process, those individuals with low earning κ periods before training are more likely to participate in the training programme. The identification condition in equation (2) does not hold in general for this example. The reason is that the individual-specific components, $\upsilon(i,t)$, are allowed to be correlated in time. However, if the selection process can be represented by equation (6), then $P(D(i,1) = 1 \mid Y(i,1-\kappa), \upsilon(i,t)) = P(D(i,1) = 1 \mid Y(i,1-\kappa))$. The DID model holds conditional on

$Y(i, 1-\kappa)$, so the impact of the treatment is given by

$$\{E[Y(i,1) \mid X(i), D(i,1)=1] - E[Y(i,1) \mid X(i), D(i,1)=0]\}$$
$$-\{E[Y(i,0) \mid X(i), D(i,1)=1] - E[Y(i,0) \mid X(i), D(i,1)=0]\} \quad (7)$$

where $X(i) = Y(i, 1-\kappa)$.[1] More generally, in this article $X(i)$ is a vector of observed characteristics, such as demographic attributes, predetermined at $t = 0$. A conditional identification restriction is appealing in the DID framework when the variables in $X(i)$ are believed to be related to the outcome dynamics and their distributions differ between treated and controls.

The traditional way to accommodate covariates in the DID model is to introduce them linearly in equation (4):

$$Y(i,t) = \mu + X(i)'\pi(t) + \tau \cdot D(i,1) + \delta \cdot t + \alpha \cdot D(i,t) + \varepsilon(i,t), \quad (8)$$

where $X(i)$ is assumed uncorrelated with $\varepsilon(i,t)$. Because the coefficients on $X(i)$ change with t, this formulation of the DID model allows the use of covariates to represent heterogeneity in outcome dynamics. Differencing the last equation with respect to t, we obtain

$$Y(i,1) - Y(i,0) = \delta + X(i)'\pi + \alpha \cdot D(i,1) + (\varepsilon(i,1) - \varepsilon(i,0)),$$

where $\pi = \pi(1) - \pi(0)$. This alternative formulation is useful when a sample with repeated observations is available. However, as noticed by Meyer (1995), introducing covariates in this linear fashion may not be appropriate if the treatment has different effects for different groups in the population. Heterogeneity in treatment effects can be studied by specifying α in equation (8) as a function of $X(i)$ (*i.e.* by including interactions between $X(i)$ and $D(i,t)$ in equation (8)). Ideally, covariates should be treated non-parametrically, as in equation (7), so that any potential inconsistency created by functional form misspecification is avoided. However, when the number of covariates required to attain identification is large, some kind of integration over $X(i)$ is required in order to obtain interpretable results.

The next section proposes a flexible new procedure to control for the effect of covariates in the DID model which is based on conditional identification restrictions. As in the conventional DID model, the role of the covariates in this new approach is twofold. First, by using covariates we extend identification to those instances in which observed compositional differences between treated and controls cause non-parallel dynamics in the outcome variable. In addition, the effect of the treatment is allowed to differ among individuals, so covariates can be used to describe the effect of the treatment for different groups of the population. A distinctive feature of the methods proposed in this article is that, while covariates are treated non-parametrically for identification, the estimators provide parsimonious parametric approximations to the average effect of the treatment on the treated conditional on selected covariates of interest.[2]

A related way to accommodate covariates in a DID estimator has been explored by Heckman *et al*. (1997, 1998) who propose a DID estimator of the average treatment effect on the treated based also on conditional identification restrictions. Their estimator is constructed by matching differences in pre-treatment and post-treatment outcomes for the treated to weighted averages of differences in pre-treatment and post-treatment outcomes for the untreated. The differences are matched on the probability of treatment exposure conditional on the covariates

1. In fact, for Ashenfelter's dip model in equation (6), the second term in equation (7) is zero and estimation could be based solely on conditional averages of post-treatment outcomes.

2. In other words, while identification does not hinge on parametric restrictions, it allows the use of parametric functions to describe how the effect of the treatment varies with covariates. This approach follows the spirit of White (1981) and Roehrig (1988) among others who propose treating parametric models as convenient approximations to unknown functions of interest, so identification can be studied non-parametrically.

(the propensity score) and the weights are determined non-parametrically using local linear regression.[3] This article, however, proposes a direct weighting scheme on the propensity score that can be used to estimate the effect of the treatment on the treated without estimating weights non-parametrically in a previous step.

Estimators of treatment effects that weight on functions of the probability of treatment are based on the Horvitz–Thompson statistic (Horvitz and Thompson, 1952). Non-parametric generalizations of the Horvitz–Thompson statistic have been studied in Imbens, Hirano and Ridder (2003). Ichimura and Linton (2002) have developed higher-order asymptotic expansions for this class of estimators to guide bandwidth selection. Robins and co-authors have proposed related estimators in the context of parametric models for time-varying treatments (see, *e.g.* Hernán, Brumback and Robins, 2001). In these studies it is assumed that all other factors, aside from the treatment, that affect the outcome variable are either observed, or their distribution is the same for treated and untreated. Consequently, all factors which confound simple comparisons of the outcome distribution between treated and untreated are observed. In contrast, the identification conditions used in this article allow for the distribution of both observed and unobserved factors to differ between treated and untreated, as long as the effect of unobserved factors on the outcome does not vary with time (or, more generally, if it experiences the same variation, on average, for treated and untreated).

3. NON-PARAMETRIC IDENTIFICATION

In the previous section, we referred to α as the "impact of the treatment" or the "treatment effect", but the exact meaning of these terms was left undefined.

As in Rubin (1974) and Heckman (1990), the effect of the treatment will be defined in terms of potential outcomes. $Y^0(i,t)$ represents the outcome that individual i would attain at time t in absence of the treatment. In the same fashion, $Y^1(i,t)$ represents the outcome that individual i would attain at time t if exposed to the treatment. The effect of the treatment on the outcome for individual i at time t is then naturally defined as $Y^1(i,t) - Y^0(i,t)$.

The fundamental identification problem is that for any particular individual i and time period t, we do not observe both potential outcomes $Y^0(i,t)$ and $Y^1(i,t)$; so we cannot compute the individual treatment effect $Y^1(i,t) - Y^0(i,t)$. We only observe the realized outcome, $Y(i,t)$ that can be expressed as $Y(i,t) = Y^0(i,t) \cdot (1 - D(i,t)) + Y^1(i,t) \cdot D(i,t)$. Since, in the simple scenario considered here, the treatment is only administered after period $t = 0$, we can denote $D(i) = D(i,1)$, then we have that $Y(i,0) = Y^0(i,0)$ and $Y(i,1) = Y^0(i,1) \cdot (1 - D(i)) + Y^1(i,1) \cdot D(i)$.

Given the impossibility of computing individual treatment effects, researchers often focus on estimating some average effect, like the average effect of the treatment on the treated $E[Y^1(i,1) - Y^0(i,1) \mid D(i) = 1]$ (see, *e.g.* Heckman, 1990). Sometimes, when the desired level of aggregation is lower, researchers try to learn about some conditional version of the average effect on the treated $E[Y^1(i,1) - Y^0(i,1) \mid X(i), D(i) = 1]$.

For the rest of the article, the individual argument i will be dropped to reduce notation. I will take the next assumption to hold throughout this article.

Assumption 3.1. $E[Y^0(1) - Y^0(0) \mid X, D = 1] = E[Y^0(1) - Y^0(0) \mid X, D = 0]$.

Assumption 3.1 is the crucial identifying restriction in DID models. It states that, conditional on the covariates, the average outcomes for treated and controls would have followed parallel

3. Blundell *et al*. (2001) propose a related estimator which combines DID and matching on the propensity score.

paths in absence of the treatment.[4] Notice that when $E[Y^0(0) \mid X, D = 1] = E[Y^0(0) \mid X, D = 0]$, Assumption 3.1 collapses to a "selection on observables" restriction ($E[Y^0(1) \mid X, D = 1] = E[Y^0(1) \mid X, D = 0]$) which can be used in cross-sectional studies to identify the effect of the treatment on the treated.[5] Therefore, the results in this article also apply to this particular situation. This is the case for Ashenfelter's dip example, discussed in Section 2. I will come back to this point in Section 3.2.

Existence of the expectations is assumed throughout. Under Assumption 3.1, the effect of the treatment on the treated conditional on X can be expressed as (Heckman *et al*., 1997):[6]

$$E[Y^1(1) - Y^0(1) \mid X, D = 1] = \{E[Y(1) \mid X, D = 1] - E[Y(1) \mid X, D = 0]\} - \{E[Y(0) \mid X, D = 1] - E[Y(0) \mid X, D = 0]\}. \quad (9)$$

Even when Assumption 3.1 holds unconditionally, if it also holds conditional on some predetermined covariates of interest (*e.g.* gender), we may still use the conditional identification result to evaluate the effect of the treatment for different groups of the population (*e.g.* women vs. men).

In principle, the identification result in equation (9) can be used to estimate $E[Y^1(1) - Y^0(1) \mid X, D = 1]$ by producing non-parametric estimates of each one of the four expectations on the R.H.S. of equation (9). In practice, the number of observations required to attain an acceptable precision for this type of non-parametric estimator increases very rapidly with the dimension of X. This problem, often called the curse of dimensionality, may prevent us from using non-parametric estimators for $E[Y^1(1) - Y^0(1) \mid X, D = 1]$ in many practical instances. In addition, a simple non-parametric estimator of $E[Y^1(1) - Y^0(1) \mid X, D = 1]$, directly based on equation (9), may be difficult to interpret if the dimension of X is larger than two, so we cannot summarize the result in a graph. In those cases, some integration over the distribution of X for the treated is required in order to produce summary statistics. Even then, the estimation process is cumbersome. An estimator of average treatment effects for the treated directly based on equation (9) requires estimating four conditional expectations non-parametrically (or two if a sample with repeated outcomes, $Y(0)$ and $Y(1)$, for the same individuals is available) and then integrating the estimates to the desired level of aggregation.

This article proposes simple weighting schemes to produce estimators of the average effect on the treated $E[Y^1(1) - Y^0(1) \mid D = 1]$ and parsimonious parametric approximations to its conditional version $E[Y^1(1) - Y^0(1) \mid X_k, D = 1]$, where X_k is a function of X (for example,

4. Using experimental data, Heckman *et al*. (1997, 1998) have shown the plausibility of this identifying assumption in the context of the evaluation of a subsidized training programme.

5. In that case, Assumption 3.1 implies $E[Y^1(1) - Y^0(1) \mid X, D = 1] = E[Y(1) \mid X, D = 1] - E[Y(1) \mid X, D = 0]$, and pre-treatment data are not required to identify the average effect of the treatment on the treated. If, in addition, $E[Y^1(1) \mid X, D = 1] = E[Y^1(1) \mid X, D = 0]$, then $E[Y^1(1) - Y^0(1) \mid X] = E[Y(1) \mid X, D = 1] - E[Y(1) \mid X, D = 0]$. Selection on observables implies that all factors which confound simple comparisons of outcomes between treated and controls are observed. This is a too stringent assumption if the distribution of unobserved variables which affect the outcome is believed to differ between treated and controls. See, *e.g.* Rubin (1977) and Heckman *et al*. (1997).

6. A researcher could be interested in estimating $E[Y^1(1) \mid X, D = 1]$ and $E[Y^0(1) \mid X, D = 1]$ separately. Since $Y^1(1)$ is observed for the treated, both conditional expectations are identified. In fact, given that $Y^1(1)$ is observed for the treated, identification results on $E[Y^1(1) - Y^0(1) \mid X, D = 1]$ are equivalent to identification results on $E[Y^0(1) \mid X, D = 1]$. Here, I concentrate on the difference $E[Y^1(1) - Y^0(1) \mid X, D = 1]$ because it has been the object of interest in the difference-in-differences literature (see, *e.g.* Heckman *et al*. (1997, 1998)). Notice also that Assumption 3.1 by itself does not identify the average treatment effects conditional only on the covariates ($E[Y^1(1) - Y^0(1) \mid X]$), unless conditional average effects coincide for treated and untreated $E[Y^1(1) - Y^0(1) \mid X, D = 1] = E[Y^1(1) - Y^0(1) \mid X, D = 0]$. The reason is that Assumption 3.1 identifies $E[Y^0(1) \mid X, D = 1]$, and therefore, the effect of the treatment on the treated. However, Assumption 3.1 leaves $E[Y^1(1) \mid X, D = 0]$ totally unrestricted; so the effect of the treatment on the untreated is also unrestricted.

a subset of the variables in X). The weighting scheme is directly based on the propensity score, $P(D = 1 \mid X)$, which is the only function which needs to be estimated in a first step. As a result, the proposed method reduces the first step estimation burden and allows the researcher to use four or two times more observations for first step estimation, relative to direct estimation of equation (9). In practice, this feature may be an important advantage if non-parametric estimation is carried out in the first step. When the number of observations is too small for non-parametric estimation in the first step, the proposed method allows the researcher to circumvent the curse of dimensionality by placing parametric restrictions on the propensity score, which leaves $E[Y^1(1) - Y^0(1) \mid X_k, D = 1]$ unrestricted, rather than on each one of the conditional means of equation (9), which may impose unwanted restrictions on $E[Y^1(1) - Y^0(1) \mid X_k, D = 1]$.

The estimation of parametric approximations to $E[Y^1(1) - Y^0(1) \mid X_k, D = 1]$ has some benefits. First, it provides a simple method to produce estimation results at the level of aggregation desired by the analyst. In addition, the results are parsimoniously summarized by the estimates of the parameters that define the approximation to $E[Y^1(1) - Y^0(1) \mid X_k, D = 1]$. However, the quality of the information provided by our estimators will be low if the quality of the approximation to $E[Y^1(1) - Y^0(1) \mid X_k, D = 1]$ is poor.

Since the object of study is the effect of the treatment on the treated, the minimal requirement for the problem to be well defined is that some fraction of the population is exposed to the treatment. In addition, since identification is attained after controlling for the effect of some covariates X, it will be required that for each given value of the covariates there is some fraction of the population that remains untreated and can be used as controls.

Assumption 3.2. *$P(D = 1) > 0$ and with probability one $P(D = 1 \mid X) < 1$.*

Note that Assumption 3.2 implies that the support of the propensity score for the treated is a subset of the support of the propensity score for the untreated. This is a well-known condition for identification of the average impact on the treated under selection on covariates (see, *e.g.* Heckman *et al.*, 1997).

3.1. Random sample with repeated outcomes

In this section, I introduce the identification strategy proposed in this article by considering the situation in which we can observe both pre-treatment and post-treatment outcomes for a random sample of the population of interest. Examples of applications of DID estimators to data on repeated outcomes are Card and Krueger (1994), Heckman *et al*. (1997, 1998), Garvey and Hanka (1999) and Blundell *et al*. (2001). Under this sampling scheme, for each individual in our sample we observe $(Y(1), Y(0), D, X)$. Later in the article, the identification procedure is extended to repeated cross sections.

Lemma 3.1. *If Assumption* 3.1 *holds, and for values of X such that $0 < P(D = 1 \mid X) < 1$, we have $E[Y^1(1) - Y^0(1) \mid X, D = 1] = E[\rho_0 \cdot (Y(1) - Y(0)) \mid X]$, where*

$$\rho_0 = \frac{D - P(D = 1 \mid X)}{P(D = 1 \mid X) \cdot (1 - P(D = 1 \mid X))}.$$

For notational convenience, let $\rho_0 = -1$ if $P(D = 1 \mid X) = 0$ (this choice is inconsequential since the objects of interest will be integrals over the distribution of the X conditional on $D = 1$). The average effect of the treatment for the treated is given by

$$
\begin{aligned}
E[Y^1(1) - Y^0(1) \mid D = 1] &= \int E[Y^1(1) - Y^0(1) \mid X, D = 1] dP(X \mid D = 1) \\
&= \int E[\rho_0 \cdot (Y(1) - Y(0)) \mid X] dP(X \mid D = 1) \\
&= E\left[\rho_0 \cdot (Y(1) - Y(0)) \cdot \frac{P(D = 1 \mid X)}{P(D = 1)}\right] \\
&= E\left[\frac{Y(1) - Y(0)}{P(D = 1)} \cdot \frac{D - P(D = 1 \mid X)}{1 - P(D = 1 \mid X)}\right]. \qquad (10)
\end{aligned}
$$

In words, under Assumptions 3.1 and 3.2, a simple weighted average of temporal differences in the outcome variable recovers the average effect of the treatment for the treated. The weights depend on the propensity score. On an intuitive level, this scheme works by weighting-down the distribution of $Y(1) - Y(0)$ for the untreated for those values of the covariates which are over-represented among the untreated (that is, with low $P(D = 1 \mid X)/P(D = 0 \mid X)$), and weighting-up $Y(1) - Y(0)$ for those values of the covariates under-represented among the untreated (that is with high $P(D = 1 \mid X)/P(D = 0 \mid X)$). In this way the same distribution of the covariates is imposed for treated and untreated.[7]

Equation (10) suggests a simple two-step method to estimate the average effect of the treatment on the treated under Assumptions 3.1 and 3.2: (i) estimate the propensity score, $P(D = 1 \mid X)$, and compute the fitted values for the sample; (ii) plug the fitted values into the sample analogue of equation (10) to obtain an estimate of $E[Y^1(1) - Y^0(1) \mid D = 1]$.

In many practical instances, the desired level of aggregation is lower than the entire treated population and the analyst wants to study how the treatment affects the treated for different groups of the population. As explained above, a non-parametric estimator of $E[Y^1(1) - Y^0(1) \mid X, D = 1]$ may be difficult to interpret, especially if the dimension of X is large. Such a problem is circumvented here by focusing on parametric approximations to $E[Y^1(1) - Y^0(1) \mid X, D = 1]$. More generally, consider the situation in which we need to condition on some vector of random variables X to attain identification, but we are interested in $E[Y^1(1) - Y^0(1) \mid X_k, D = 1]$, where X_k is some deterministic function of X.[8] This situation is relevant when the number of covariates needed in order to attain identification is large, so the analyst may be willing to allow for a higher level of aggregation in the second step in order to obtain parsimonious results.

Consider a class of approximating functions $\mathcal{G} = \{g(X_k; \theta) : \theta \in \Theta \subset \mathbb{R}^k\}$, square-integrable with respect to $P(X_k \mid D = 1)$. Then, a least squares approximation from $\mathcal{G}$ to $E[Y^1(1) - Y^0(1) \mid X_k, D = 1]$ is given by $g(X_k; \theta_0)$ where

$$\theta_0 = \arg\min_{\theta \in \Theta} E[\{E[Y^1(1) - Y^0(1) \mid X_k, D = 1] - g(X_k; \theta)\}^2 \mid D = 1]. \qquad (11)$$

For example, if $\mathcal{G} = \{X_k'\theta : \theta \in \Theta \subset \mathbb{R}^k\}$, then θ_0 defines a linear least squares approximation to $E[Y^1(1) - Y^0(1) \mid X_k, D = 1]$. It is assumed that θ_0 exists uniquely.

Proposition 3.1. *If Assumptions* 3.1 *and* 3.2 *hold, then*

$$\theta_0 = \arg\min_{\theta \in \Theta} E[P(D = 1 \mid X) \cdot \{\rho_0 \cdot (Y(1) - Y(0)) - g(X_k; \theta)\}^2].$$

7. Similarly, Heckman *et al*. (1997) use matching on the propensity score to account for imbalances in the distribution of the covariates between treated and untreated. Matching on the propensity score works because it imposes the same distribution of the covariates for the two groups (see Rosenbaum and Rubin, 1983).

8. For example, X_k may contain a subset of the variables in X. In other instances, X may contain indicators for all different values of a discrete variable included in X_k.

As before, it is easy to construct a two-step estimator based on the sample analogue of the result in the last proposition. After some algebra, the result in equation (10) can be obtained applying Proposition 3.1 to the case when $g(X_k, \theta)$ is constant.

3.2. Some extensions and particular cases

This section contains extensions to repeated cross sections, and multilevel treatments. It also discusses the case of selection on observables.

3.2.1. Repeated cross sections. Often, a random sample with repeated outcomes is not available. In such a case, repeated cross-section data-sets (pre-treatment and post-treatment) may be used to construct DID estimators. However, the use of repeated cross sections for DID presents some issues of data availability. First, treatment status (in the post-treatment period) must be known for the individuals in the pre-treatment sample. This requirement is satisfied, for example, if treatment exposure can be determined from some individual characteristic observed in both periods.[9,10] In addition, covariates must be observed in the post-treatment sample. Since covariates are often pre-treatment variables, this second requirement may prove a problem when covariates are time-varying and there are not retrospective covariate data available. Examples of applications of DID estimators with covariates to repeated cross sections are Card (1990, 1992), Meyer *et al*. (1995), Eissa and Liebman (1996), Acemoglu and Angrist (2001), Corak (2001) and Finkelstein (2002). Here, I show how to apply the methods proposed in this article to repeated cross sections. The data requirements are the same as for traditional difference-in-differences estimators which use cross-sectional data and covariates.

Assume that random samples are available for the pre-treatment and the post-treatment periods. For each individual in the pooled sample (post-treatment and pre-treatment), we observe $Z = (Y, D, T, X)$ where T is a time indicator that takes value one if the observation belongs to the post-treatment sample.

Assumption 3.3. *Conditional on* $T = 0$*, the data are* i.i.d. *from the distribution of* $(Y(0), D, X)$*; conditional on* $T = 1$*, the data are* i.i.d. *from the distribution of* $(Y(1), D, X)$.

This sampling scheme produces the following mixture distribution:

$$P_M(Y = y, D = d, X = x, T = t) = \lambda \cdot t \cdot P(Y(1) = y, D = d, X = x) + (1 - \lambda) \cdot (1 - t) \cdot P(Y(0) = y, D = d, X = x),$$

where $\lambda \in (0, 1)$ reflects the proportion of the observations sampled in the post-treatment period.[11] Let $E_M[\cdot]$ denote expectations with respect to $P_M(\cdot)$.

Lemma 3.2. *If Assumptions* 3.1 *and* 3.3 *hold, and for values of* X *such that* $0 < P(D = 1 \mid X) < 1$*, we have* $E[Y^1(1) - Y^0(1) \mid X, D = 1] = E_M[\varphi_0 \cdot Y \mid X]$*, where*

$$\varphi_0 = \frac{T - \lambda}{\lambda \cdot (1 - \lambda)} \cdot \frac{D - P(D = 1 \mid X)}{P(D = 1 \mid X) \cdot P(D = 0 \mid X)}.$$

9. Note that identification requires in turn that such individual characteristic is excluded from X. Otherwise, the support condition in Assumption 3.2 would be violated. This exclusion restriction is problematic if the excluded variable influences the dynamics of the outcome variable, so Assumption 3.1 is not plausible.

10. The requirement may also be satisfied in other cases, for example when the pre-treatment sample can be linked to administrative data records on treatment participation.

11. For simplicity, I do not consider more complicated situations in which the data may be generated by stratified sampling (on X or D). In such a case, the results in this section apply for a suitably reweighted sample (see, *e.g.* Wooldridge, 2002).

Then, following a reasoning similar to that of the previous section we have that the average treatment effect on the treated is identified by

$$E_M\left[\frac{P(D=1 \mid X)}{P(D=1)} \cdot \varphi_0 \cdot Y\right] = E[Y^1(1) - Y^0(1) \mid D=1]. \tag{12}$$

The following proposition is analogous to Proposition 3.1.

Proposition 3.2. *If Assumptions* 3.1, 3.2, *and* 3.3 *hold, then for* θ_0 *defined in equation* (11) *we have*

$$\theta_0 = \arg\min_{\theta\in\Theta} E_M[P(D=1 \mid X) \cdot \{\varphi_0 \cdot Y - g(X_k;\theta)\}^2]. \tag{13}$$

The result in equation (12) can be obtained by considering a constant $g(X_k, \theta)$ in equation (13).

3.2.2. Selection on observables. In the previous section, it is shown how to approximate conditional average treatment effects by first weighting temporal differences in the outcome variable on the propensity score, and then projecting the weighted differences on a set of parametric functions of the covariates. The interpretation of the resulting functionals as approximations to conditional average treatment effects comes from the difference-in-differences condition in Assumption 3.1. However, it should be noticed that the same "first weight, then project" strategy can be applied in other contexts. In particular, the results in the previous section carry over naturally to "selection on observables":

$$E[Y^0(1) \mid X, D=1] = E[Y^0(1) \mid X, D=0]. \tag{14}$$

The reason is that "selection on observables" can be expressed as a particular case of Assumption 3.1 (when $E[Y^0(0) \mid X, D=1] = E[Y^0(0) \mid X, D=0]$). As a result, if equation (14) holds, then

$$\theta_0 = \arg\min_{\theta\in\Theta} E[P(D=1 \mid X) \cdot \{\rho_0 \cdot Y - g(X_k;\theta)\}^2],$$

for θ_0 defined in equation (11) and $Y = Y(1)$.[12,13] Similar estimators have been considered in Wooldridge (2001). Imbens *et al*. (2003) consider the situation in which the average treatment effect is estimated for a given distribution of the covariates. Abadie (2003) applies similar approximation methods in the context of instrumental variable models for treatment effects.

3.2.3. Multilevel treatments. So far, we have considered only the case of a binary treatment, which is the usual focus of DID estimators. However, the same ideas can be applied when individuals may be exposed to different levels (or doses) of the treatment. Let W represent the level of the treatment. For untreated individuals, let $W = 0$. For the treated, suppose that W takes on a finite number of positive values $w_1 < \cdots < w_J$, with positive probability.[14] Let

12. For this case, it is useful to define $Y = Y(1)$ because equation (14) may be adopted as an identification restriction in absence of measures on the outcome variable in a pretreatment period. However, as shown for Ashenfelter's dip example, it may be necessary to condition on the values of the outcome variable in a pre-treatment period in order for equation (14) to hold.

13. Alternatively, if the objects of interest are average treatment effects conditional only on the covariates, θ_0 can be redefined as $\theta_0 = \arg\min_{\theta\in\Theta} E[\{E[Y^1(1) - Y^0(1) \mid X_k] - g(X_k,\theta)\}^2]$. However, equation (14) does not identify the effect of the treatment on the untreated because it leaves $E[Y^1(1) \mid X, D=0]$ completely unrestricted. If we assume in addition that $E[Y^1(1) \mid X, D=1] = E[Y^1(1) \mid X, D=0]$, then θ_0 is identified by $\theta_0 = \arg\min_{\theta\in\Theta} E[(\rho_0 \cdot Y - g(X_k;\theta))^2]$.

14. Here, I assume that treatment levels are ordered (*e.g.* number of weeks in a training programme). For expositional simplicity and since it is often the case in applications, I consider only a finite number of treatment levels. However, the analysis presented in this section can be generalized to continuous treatments by substituting densities for W for probabilities for W, and integrals over those densities for sums.

$\mathcal{W}^+ = \{w_1, \ldots, w_J\}$, then $D = 1_{\mathcal{W}^+}(W)$, where $1_A(\cdot)$ is the indicator function for the set A (that is, $1_A(w) = 1$ if $w \in A$, zero otherwise). For $w \in \{0\} \cup \mathcal{W}^+$ and $t \in \{0, 1\}$, let $Y^w(t)$ be the potential outcome for treatment level w and period t. Suppose that Assumption 3.1 holds for each treatment level: $E[Y^0(1) - Y^0(0) \mid X, W = w] = E[Y^0(1) - Y^0(0) \mid X, W = 0]$, for $w \in \mathcal{W}^+$. In words, this assumption requires that, in absence of the treatment, the average outcomes for all treatment groups would have followed parallel trends, conditional on the covariates. As in the usual DID case with a binary treatment variable, the assumption allows the levels of average potential outcomes without the treatment to differ arbitrarily between treatment groups.

For $w \in \mathcal{W}^+$, let $\rho_0^w = (1_{\{w\}}(W)/P(W = w \mid X)) - (1_{\{0\}}(W)/P(W = 0 \mid X))$. (Note that, for ρ_0 defined in Lemma 3.1, $\rho_0 = \rho_0^1$.) Then, following the same reasoning as for Lemma 3.1, we obtain $E[\rho_0^w(Y(1) - Y(0)) \mid X] = E[Y^w(1) - Y^0(1) \mid X, W = w]$. In addition, for some class of square-integrable approximating functions $\mathcal{G} = \{g(W, X_k; \theta) : \theta \in \Theta\}$, redefine

$$\theta_0 = \arg\min_{\theta \in \Theta} E[(E[Y^W(1) - Y^0(1) \mid X_k, W] - g(W, X_k; \theta))^2 \mid D = 1]. \tag{15}$$

The parameters θ_0 define a least squares approximation to a function describing average effects for the treated $E[Y^w(1) - Y^0(1) \mid X_k, W = w]$. Then, these parameters are identified by

$$\theta_0 = \arg\min_{\theta \in \Theta} E\left[\sum_{j=1}^{J} P(W = w_j \mid X)(\rho_0^{w_j}(Y(1) - Y(0)) - g(w_j, X_k; \theta))^2\right].$$

This result collapses to the result in Proposition 3.1 if W is binary, and can be proven using the same argument as for Proposition 3.1.

4. ESTIMATION AND ASYMPTOTIC DISTRIBUTION

For concreteness, I will concentrate here on linear approximations to $E[Y^1(1) - Y^0(1) \mid X_k, D = 1]$ where X_k is a deterministic function of X. In addition, only the case of repeated cross sections is explicitly considered here. However, the analysis is also valid for the case of repeated observations if $Y(1) - Y(0)$ is substituted for $((T - \lambda)/\lambda(1 - \lambda)) \cdot Y$, and expectations are taken with respect to the distribution of $(Y(1), Y(0), D, X)$. Consider

$$\beta_0 = \arg\min_{\beta \in \Theta} E_M[\pi_0 \cdot \{\varphi_0 Y - X_k'\beta\}^2]$$

where $\pi_0(X) = P(D = 1 \mid X)$, and

$$\varphi_0 = \frac{T - \lambda}{\lambda \cdot (1 - \lambda)} \cdot \frac{D - \pi_0(X)}{\pi_0(X) \cdot (1 - \pi_0(X))}.$$

Consider also the following estimator of β_0:

$$\widehat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_{ki}\widehat{\pi}(X_i)X_{ki}'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_{ki}\widehat{\pi}(X_i)\widehat{\varphi}_i Y_i,$$

where $\widehat{\pi}(X_i)$ is an estimator of $\pi_0(X_i)$, and

$$\widehat{\varphi}_i = \frac{T_i - \lambda}{\lambda \cdot (1 - \lambda)} \cdot \frac{D_i - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i) \cdot (1 - \widehat{\pi}(X_i))},$$

for $\lambda = n_1/(n_0 + n_1)$. Under the conditions of the theorems stated below, $\widehat{\beta}$ is well defined with probability approaching one.

4.1. Non-parametric first step estimation of the propensity score

Here, I consider the case in which non-parametric (power series) regression is used in a first step to estimate π_0. Let $\zeta = (\zeta_1, \ldots, \zeta_r)'$ be a vector of non-negative integers where r is

the dimension of X. Also let $X^{\zeta} = \prod_{j=1}^{r} X_j^{\zeta_j}$ and $|\zeta| = \sum_{j=1}^{r} \zeta_j$. Let $\{\zeta(k)\}_{k=1}^{\infty}$ be a sequence containing all distinct vectors ζ, with $|\zeta|$ non-decreasing. For a positive integer K, let $p^K(X) = (p_{1K}(X), \ldots, p_{KK}(X))'$ where $p_{kK}(X) = X^{\zeta(k)}$. Then, for $K = K(n) \to \infty$ a power series non-parametric estimator of π_0 is given by

$$\widehat{\pi}(X) = p^K(X)'\widehat{\gamma} \tag{16}$$

where $\widehat{\gamma} = (\sum_{i=1}^{n} p^K(X_i)p^K(X_i)')^{-}(\sum_{i=1}^{n} p^K(X_i)D_i)$ and A^{-} denotes any symmetric generalized inverse of the matrix A.

Assumption 4.1. (i) *$K^6/n = o(1)$, $\pi_0(X)$ is continuously differentiable of order s, and $nK^{-2s/r} = O(1)$;* (ii) *the support of X is a Cartesian product of compact intervals on which X has density that is bounded away from zero;* (iii) *$\pi_0(X)$ is bounded away from zero and one;* (iv) *β_0 is an interior point of a compact set $\Theta \subset \mathbb{R}^k$;* (v) *$E_M Y^2 < \infty$, $\|X_k\|$ is bounded, and $E[X_k X_k' \mid D = 1]$ is non-singular.*

Let

$$\delta(X) = E_M\left[X_k\left(\frac{T-\lambda}{\lambda\cdot(1-\lambda)}\cdot\frac{D-1}{(1-\pi_0)^2}\cdot Y - X_k'\beta_0\right)\middle| X\right].$$

Theorem 4.1. *If $n_0, n_1 \to \infty$, $n_1/(n_0+n_1) = \lambda \in (0,1)$ and Assumptions* 3.1, 3.3 *and* 4.1 *hold, then $n^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{d} N(0, V)$, where $V = Q^{-1}\Sigma Q^{-1}$, $Q = E[X_k D X_k']$, $\Sigma = E_M[\psi\psi']$, $\psi = X_k\pi_0(X)(\varphi_0 Y - X_k'\beta_0) + \delta(X)\cdot(D - \pi_0(X))$.*

To construct an estimator of the asymptotic variance, let $\widehat{V} = \widehat{Q}^{-1}\widehat{\Sigma}\widehat{Q}^{-1}$, where

$$\widehat{Q} = \frac{1}{n}\sum_{i=1}^{n} X_{ki} D_i X_{ki}', \qquad \widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\psi}_i\widehat{\psi}_i',$$

$$\widehat{\delta}(X_i) = \left(\sum_{i=1}^{n} X_{ki}\left(\frac{T_i-\lambda}{\lambda\cdot(1-\lambda)}\cdot\frac{D_i-1}{(1-\widehat{\pi}(X_i))^2}\cdot Y_i - X_{ki}'\widehat{\beta}\right)p^K(X_i)'\right) \times \left(\sum_{i=1}^{n} p^K(X_i)p^K(X_i)'\right)^{-} p^K(X_i)$$

and $\widehat{\psi}_i = X_{ki}\widehat{\pi}(X_i)(\widehat{\varphi}_i Y_i - X_{ki}'\widehat{\beta}) + \widehat{\delta}(X_i)\cdot(D_i - \widehat{\pi}(X_i))$. The following theorem establishes the consistency of $\widehat{V}$.

Theorem 4.2. *If the assumptions of Theorem* 4.1 *hold and $K^7/n \to 0$, then $\widehat{V} \xrightarrow{p} V$.*

4.2. Parametric first step estimation of the propensity score

Often, samples are too small to use a non-parametric estimator in the first step. This is particularly likely when the analyst deals with longitudinal data-sets. In such cases, it may be convenient to use a parametric restriction in the first step and estimate the propensity score by maximum likelihood. This section provides distribution theory for that case. The results include probit, logit and linear probability first step estimation of π_0 as special cases.

Assumption 4.2. (i) *γ_0 is an interior point of a compact set $\Gamma \subset \mathbb{R}^r$;* (ii) *the support of X is a subset of a compact set S; $E[XX']$ is non-singular;* (iii) *there is a (known) function $\pi : \mathbb{R} \mapsto [0,1]$ such that $\pi_0(X) = \pi(X'\gamma_0)$;* (iv) *let $\mathcal{V} = \{x'\gamma : x \in S, \gamma \in \Gamma\}$; for $v \in \mathcal{V}$, $\pi(v)$ is bounded away from zero and one, strictly increasing and continuously differentiable with*

derivative bounded away from zero and infinity; (v) β_0 *is an interior point of a compact set* $\Theta \subset \mathbb{R}^k$*;* (vi) $E_M Y^2 < \infty$, $\|X_k\|$ *is bounded, and* $E[X_k X_k' \mid D = 1]$ *is non-singular.*

Under this assumption, γ_0 can be estimated by maximum likelihood:

$$\widehat{\gamma} = \arg\max_{\gamma \in \Gamma} \frac{1}{n} \sum_{i=1}^{n} D_i \log \pi(X_i'\gamma) + (1 - D_i) \log(1 - \pi(X_i'\gamma)).$$

Then, $\widehat{\pi}(X_i) = \pi(X_i'\widehat{\gamma})$. Boundedness of X is not a necessary restriction but allows Assumption 4.2 to encompass maximum likelihood estimation of the logit, probit and linear probability models. For simplicity, and since in most cases X_k is some subset of the variables in X, X_k is also assumed to be bounded. If the covariates are discrete and the vector X is saturated with indicators of all the possible values of the covariates, then the functional form of π is not restrictive and the estimation of $\widehat{\pi}(X)$ is completely non-parametric.

Let $\dot{\pi} = \partial \pi(v)/\partial v$ and $\dot{\pi}_0 = \dot{\pi}(X'\gamma_0)$. Under standard regularity conditions (*e.g.* Assumption 4.2(i)–(iv)), $\widehat{\gamma}$ is asymptotically linear, that is $n^{1/2}(\widehat{\gamma} - \gamma_0) = n^{-1/2} \sum_{i=1}^{n} \psi_{\gamma_0}(Z_i) + o_p(1)$, where

$$\psi_{\gamma_0}(Z) = E\left[\frac{\dot{\pi}_0^2}{\pi_0(1-\pi_0)} XX'\right]^{-1} X \frac{\dot{\pi}_0}{\pi_0(1-\pi_0)} (D - \pi_0). \tag{17}$$

Let

$$M_{\gamma_0} = E_M\left[X_k \left(\frac{T - \lambda}{\lambda(1-\lambda)} \frac{D-1}{(1-\pi_0)^2} Y - X_k'\beta_0\right) \dot{\pi}_0 X'\right].$$

Theorem 4.3. *If* $n_0, n_1 \to \infty$, $n_1/(n_0 + n_1) = \lambda \in (0,1)$ *and Assumptions* 3.1, 3.3 *and* 4.2 *hold, then* $\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} N(0, V)$*, where* $V = Q^{-1}\Sigma Q^{-1}$, $Q = E[X_k D X_k']$, $\Sigma = E_M[\psi\psi']$, $\psi = m(Z, \beta_0, \gamma_0) + M_{\gamma_0}\psi_{\gamma_0}$, $m(Z, \beta_0, \gamma_0) = X_k \pi_0[\varphi_0 \cdot Y - X_k'\beta_0]$.

Let $\widehat{V} = \widehat{Q}^{-1}\widehat{\Sigma}\widehat{Q}^{-1}$, where

$$\widehat{Q} = \frac{1}{n}\sum_{i=1}^{n} X_{ki} D_i X_{ki}', \qquad \widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\psi}_i \widehat{\psi}_i',$$

$$\widehat{M}_{\widehat{\gamma}} = \frac{1}{n}\sum_{i=1}^{n} X_{ki}\left(\frac{T_i - \lambda}{\lambda(1-\lambda)} \frac{D_i - 1}{(1 - \widehat{\pi}(X_i))^2} Y_i - X_{ki}'\widehat{\beta}\right) \dot{\pi}(X_i'\widehat{\gamma}) X_i',$$

$$\widehat{\psi}_{\widehat{\gamma}}(Z_i) = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\dot{\pi}(X_i'\widehat{\gamma})^2}{\widehat{\pi}(X_i)(1-\widehat{\pi}(X_i))} X_i X_i'\right)^{-1} X_i \frac{\dot{\pi}(X_i'\widehat{\gamma})}{\widehat{\pi}(X_i)(1-\widehat{\pi}(X_i))}(D_i - \widehat{\pi}(X_i)),$$

and $\widehat{\psi}_i = X_{ki}\widehat{\pi}(X_i)(\widehat{\varphi}_i Y_i - X_{ki}'\widehat{\beta}) + \widehat{M}_{\widehat{\gamma}}\widehat{\psi}_{\widehat{\gamma}}(Z_i)$. The following theorem establishes consistency of $\widehat{V}$.

Theorem 4.4. *If the assumptions of Theorem* 4.3 *hold and* $\pi(v)$ *is twice differentiable with bounded second derivative in* $\mathcal{V}$*, then* $\widehat{V} \xrightarrow{p} V$.

5. CONCLUSIONS

In this article, I have introduced a family of semiparametric difference-in-differences estimators of treatment effects based on conditional identification restrictions. These estimators may be particularly appropriate when the distribution of observed characteristics that are thought to be related to the dynamics of the outcome variable differs between treated and untreated.

Identification does not entail parametric restrictions. However, the methods presented here can be used to estimate parsimonious parametric approximations to conditional versions of the average treatment effect for the treated.

APPENDIX. PROOFS

Proof of Lemma 3.1. For $0 < P(D = 1 \mid X) < 1$, we have that

$$\begin{aligned} E[\rho_0 \cdot (Y(1) - Y(0)) \mid X] &= E[\rho_0 \cdot (Y(1) - Y(0)) \mid X, D = 1] \cdot P(D = 1 \mid X) \\ &\quad + E[\rho_0 \cdot (Y(1) - Y(0)) \mid X, D = 0] \cdot P(D = 0 \mid X) \\ &= E[Y(1) - Y(0) \mid X, D = 1] - E[Y(1) - Y(0) \mid X, D = 0]. \end{aligned}$$

Applying equation (9) we obtain the result. ‖

Proof of Proposition 3.1. Let $G(\theta) = E[P(D = 1 \mid X) \cdot \{\rho_0 \cdot (Y(1) - Y(0)) - g(X_k; \theta)\}^2]$. Adding and subtracting $E[Y^1(1) - Y^0(1) \mid X_k, D = 1]$, we obtain

$$\begin{aligned} G(\theta) = {} & E[P(D = 1 \mid X) \cdot \{\rho_0 \cdot (Y(1) - Y(0)) - E[Y^1(1) - Y^0(1) \mid X_k, D = 1]\}^2] \\ & + E[P(D = 1 \mid X) \cdot \{E[Y^1(1) - Y^0(1) \mid X_k, D = 1] - g(X_k; \theta)\}^2] \\ & + 2E[P(D = 1 \mid X) \cdot (\rho_0 \cdot (Y(1) - Y(0)) - E[Y^1(1) - Y^0(1) \mid X_k, D = 1]) \\ & \times (E[Y^1(1) - Y^0(1) \mid X_k, D = 1] - g(X_k; \theta))]. \end{aligned} \tag{A.1}$$

The first term on the R.H.S. of equation (A.1) does not depend on θ.

By the law of iterated expectations, the second term on the R.H.S. of equation (A.1) is equal to

$$\begin{aligned} & E[D \cdot \{E[Y^1(1) - Y^0(1) \mid X_k, D = 1] - g(X_k; \theta)\}^2] \\ & \quad = E[\{E[Y^1(1) - Y^0(1) \mid X_k, D = 1] - g(X_k; \theta)\}^2 \mid D = 1]P(D = 1). \end{aligned}$$

Therefore, by Assumption 3.2 and equation (11), the second term on the R.H.S. of equation (A.1) is minimized at θ_0.

The expectation in the third term on the R.H.S. of equation (A.1) is equal to

$$\begin{aligned} & E[E[P(D = 1 \mid X) \cdot (\rho_0 \cdot (Y(1) - Y(0)) - E[Y^1(1) - Y^0(1) \mid X_k, D = 1]) \mid X_k] \\ & \quad \times (E[Y^1(1) - Y^0(1) \mid X_k, D = 1] - g(X_k; \theta))]. \end{aligned}$$

Applying the law of iterated expectations and Lemma 3.1:

$$\begin{aligned} & E[P(D = 1 \mid X) \cdot \rho_0 \cdot (Y(1) - Y(0)) \mid X_k] = E[P(D = 1 \mid X) E[\rho_0 \cdot (Y(1) - Y(0)) \mid X] \mid X_k] \\ & \quad = E[P(D = 1 \mid X) E[Y^1(1) - Y^0(1) \mid X, D = 1] \mid X_k] = E[E[D(Y^1(1) - Y^0(1)) \mid X] \mid X_k] \\ & \quad = E[D(Y^1(1) - Y^0(1)) \mid X_k] = P(D = 1 \mid X_k) \cdot E[Y^1(1) - Y^0(1) \mid X_k, D = 1]. \end{aligned}$$

Therefore,

$$\begin{aligned} & E[P(D = 1 \mid X) \cdot (\rho_0 \cdot (Y(1) - Y(0)) - E[Y^1(1) - Y^0(1) \mid X_k, D = 1]) \mid X_k] \\ & \quad = E[P(D = 1 \mid X) \cdot \rho_0 \cdot (Y(1) - Y(0)) \mid X_k] - P(D = 1 \mid X_k) E[Y^1(1) - Y^0(1) \mid X_k, D = 1] = 0. \end{aligned}$$

Consequently, the third term on the R.H.S. of equation (A.1) is equal to zero, and the result of the proposition holds. ‖

Proof of Lemma 3.2. Notice that

$$\begin{aligned} E_M[\varphi_0 \cdot Y \mid X] &= E_M[E_M[\varphi_0 \cdot Y \mid X, T] \mid X] = E_M[E[\varphi_0 \cdot Y \mid X, T] \mid X] \\ &= E[\rho_0 \cdot Y \mid X, T = 1] - E[\rho_0 \cdot Y \mid X, T = 0] = E[\rho_0 \cdot Y(1) \mid X] - E[\rho_0 \cdot Y(0) \mid X] \\ &= \{E[Y(1) \mid X, D = 1] - E[Y(1) \mid X, D = 0]\} - \{E[Y(0) \mid X, D = 1] - E[Y(0) \mid X, D = 0]\}, \end{aligned}$$

and the result follows from equation (9). ‖

Proof of Proposition 3.2. The proof follows the same steps as the proof of Proposition 3.1.

Proof of Theorem 4.1. Let $|\cdot|_\infty$ denote the supremum norm. By Assumption 4.1(i), (ii) and Theorem 4 in Newey (1997), it follows that $|\widehat{\pi} - \pi_0|_\infty = O_p(K \cdot [(K/n)^{1/2} + K^{-s/r}]) = o_p(1)$. Let $m(Z, \beta, \pi) = X_k \pi(\varphi(\pi) \cdot Y - X_k' \beta)$.

Assumption 4.1(v) guarantees that $E_M[m(Z,\beta,\pi_0)]$ has a unique zero at $\beta_0 = (E[X_k\pi_0 X_k'])^{-1}E_M[X_k\pi_0\varphi_0 Y]$. In addition,

$$\sup_{\beta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^n m(Z_i,\beta,\widehat{\pi}) - E_M[m(Z,\beta,\pi_0)]\right\| \leq \sup_{\beta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^n m(Z_i,\beta,\widehat{\pi}) - m(Z_i,\beta,\pi_0)\right\| + \sup_{\beta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^n m(Z_i,\beta,\pi_0) - E_M[m(Z,\beta,\pi_0)]\right\|. \tag{A.2}$$

Let

$$\varphi_{0i} = \frac{T_i-\lambda}{\lambda\cdot(1-\lambda)}\cdot\frac{D_i-\pi_0(X_i)}{\pi_0(X_i)\cdot(1-\pi_0(X_i))}.$$

Notice that, $m(Z_i,\beta,\widehat{\pi}) - m(Z_i,\beta,\pi_0) = X_{ki}\{(\widehat{\pi}(X_i)\,\widehat{\varphi}_i - \pi_0(X_i)\,\varphi_{0i})\,Y_i - (\widehat{\pi}(X_i)-\pi_0(X_i))\,X_{ki}'\beta\}$, and

$$\widehat{\pi}(X_i)\,\widehat{\varphi}_i - \pi_0(X_i)\,\varphi_{0i} = \frac{T_i-\lambda}{\lambda(1-\lambda)}\cdot\frac{D_i-1}{(1-\widehat{\pi}(X_i))(1-\pi_0(X_i))}\cdot(\widehat{\pi}(X_i)-\pi_0(X_i)).$$

By Assumption 4.1(iii) and uniform convergence of $\widehat{\pi}$, with probability approaching one $\widehat{\pi}$ is bounded away from one and there is a constant C such that

$$\sup_{\beta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^n m(Z_i,\beta,\widehat{\pi}) - m(Z_i,\beta,\pi_0)\right\| \leq C\cdot\frac{1}{n}\sum_{i=1}^n\{\|X_{ki}Y_i\| + \|X_{ki}\|^2\|\beta\|\}|\widehat{\pi}-\pi_0|_\infty.$$

Therefore, the first term of the R.H.S. of equation (A.2) is $o_p(1)$. Assumption 4.1 also implies that $\|m(Z,\beta,\pi_0)\|$ is dominated by an integrable function. Then, since $m(Z,\beta,\pi_0)$ is continuous at each $\beta\in\Theta$ compact, Lemma 2.4 in Newey and McFadden (1994) implies that the second term of the R.H.S. of equation (A.2) is also $o_p(1)$ and that $E[m(Z,\beta,\pi_0)]$ is continuous. Now, the usual consistency argument for estimators based on moment equations (*e.g.* van der Vaart (1998, Theorem 5.9)) proves that $\widehat{\beta}\xrightarrow{p}\beta_0$.

Since $|\widehat{\pi}-\pi_0|_\infty$ converges in probability to zero, with probability approaching one $n^{-1}\sum_{i=1}^n X_{ki}\widehat{\pi}(X_i)X_{ki}'$ is non-singular and

$$n^{1/2}(\widehat{\beta}-\beta_0) = \left(\frac{1}{n}\sum_{i=1}^n X_{ki}\widehat{\pi}(X_i)X_{ki}'\right)^{-1}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ki}\widehat{\pi}(X_i)(\widehat{\varphi}_i Y_i - X_{ki}'\beta_0).$$

Since X_k has second moments,

$$\frac{1}{n}\sum_{i=1}^n X_{ki}\widehat{\pi}(X_i)X_{ki}' = E[X_k D X_k'] + o_p(1).$$

Now, let us prove that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n m(Z_i,\beta_0,\widehat{\pi}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ki}\widehat{\pi}(X_i)(\widehat{\varphi}_i Y_i - X_{ki}'\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(Z_i) + o_p(1).$$

Consider

$$\Lambda(Z,\pi,\beta,\widetilde{\pi}) = \left(\frac{\partial}{\partial\pi}m(Z,\beta,\pi)_{|\pi=\widetilde{\pi}}\right)\cdot\pi = X_k\left[\frac{T-\lambda}{\lambda\cdot(1-\lambda)}\cdot\frac{D-1}{(1-\widetilde{\pi})^2}\cdot Y - X_k'\beta\right]\cdot\pi.$$

It can be easily seen that for $\pi,\widetilde{\pi}$ bounded away from one we have

$$m(Z,\beta,\pi) - m(Z,\beta,\widetilde{\pi}) - \Lambda(Z,\pi-\widetilde{\pi},\beta,\widetilde{\pi}) = X_k\cdot\frac{T-\lambda}{\lambda\cdot(1-\lambda)}\cdot\frac{D-1}{(1-\pi)(1-\widetilde{\pi})^2}\cdot Y\cdot(\pi-\widetilde{\pi})^2.$$

Therefore, since, with probability approaching one, $(1-\widehat{\pi})^{-1}$ is bounded by some constant C, we have

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n m(Z_i,\beta_0,\widehat{\pi}) - m(Z_i,\beta_0,\pi_0) - \Lambda(Z_i,\widehat{\pi}-\pi_0,\beta_0,\pi_0)\right\| \leq n^{1/2}|\widehat{\pi}-\pi_0|_\infty^2 C\cdot\frac{1}{n}\sum_{i=1}^n\left\|X_{ki}\cdot\frac{T_i-\lambda}{\lambda\cdot(1-\lambda)}\cdot\frac{D_i-1}{(1-\pi_0)^2}\cdot Y_i\right\| = o_p(1).$$

This result holds because Assumption 4.1(i) implies that $n^{1/2}K^2\cdot(K/n+K^{-2s/r})\to 0$, and therefore $n^{1/2}|\widehat{\pi}-\pi_0|_\infty^2 = o_p(1)$. The assumptions $E_M|Y|<\infty$, $\|X_k\|$ bounded and π_0 bounded away from one take care of the sample average term. Therefore

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n m(Z_i,\beta_0,\widehat{\pi}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n m(Z_i,\beta_0,\pi_0) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \Lambda(Z_i,\widehat{\pi}-\pi_0,\beta_0,\pi_0) + o_p(1).$$

Note that $\|\Lambda(Z,\pi,\beta_0,\pi_0)\| \leq b(Z)\cdot|\pi|_\infty$ with $b(Z) = \|X_k[(T-\lambda)/(\lambda\cdot(1-\lambda))]\cdot[(D-1)/(1-\pi_0)^2]\cdot Y - X_k'\beta_0\|$. By $E_M Y^2 < \infty$ and $\|X_k\|$ bounded, we have that $E_M b(Z)^2 < \infty$. Then, it follows from the proof of Theorem 6.1 in Newey (1994) that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\{\Lambda(Z_i,\widehat{\pi}-\pi_0,\beta_0,\pi_0) - E_M[\Lambda(Z,\widehat{\pi}-\pi_0,\beta_0,\pi_0)]\}$$
$$= O_p(K^{-s/r}) + O_p(K\cdot[(K/n)^{1/2} + K^{-s/r}]) = o_p(1).$$

Note that $E_M[\Lambda(Z,\pi,\beta_0,\pi_0)] = E[\delta(X)\pi(X)]$. If:

(a) For each K, there exists ξ_K such that $E[\|\delta(X)-\xi_K p^K(X)\|^2]\to 0$ and $nK^{-2s/r}E[\|\delta(X)-\xi_K p^K(X)\|^2]\to 0$,
(b) $K^5/n\to 0$,
(c) $K^2\cdot K^{-2s/r}\to 0$,

then, Assumption 6.6 in Newey (1994) holds for $\Lambda(Z,\pi,\beta_0,\pi_0)$. Condition (a) holds by Assumption 4.1(i) and by square-integrability of $\delta(X)$. Conditions (b) and (c) follow from Assumption 4.1(i). Under such conditions, Newey (1994) shows that

$$n^{1/2}E_M[\Lambda(Z,\widehat{\pi}-\pi_0,\beta_0,\pi_0)] = \frac{1}{\sqrt{n}}\sum_{i=1}^n\delta(X_i)\cdot(D_i-\pi_0(X_i)) + o_p(1).$$

So the result of the theorem follows from existence of second moments of ψ. ‖

Proof of Theorem 4.2. By the law of large numbers $\widehat{Q}$ converges in probability to Q, which is non-singular. Notice that

$$X_{ki}\widehat{\pi}(X_i)(\widehat{\varphi}_i Y_i - X_{ki}'\widehat{\beta}) - X_{ki}\pi_0(X_i)(\varphi_{0i}Y_i - X_{ki}'\beta_0)$$
$$= X_{ki}\left(\frac{T_i-\lambda}{\lambda(1-\lambda)}\cdot\frac{D_i-1}{(1-\widehat{\pi}(X_i))(1-\pi_0(X_i))}\right)Y_i(\widehat{\pi}(X_i)-\pi_0(X_i)) - X_{ki}X_{ki}'(\widehat{\beta}-\beta_0).$$

Now, since $\|X_{ki}\|$ is bounded, $E_M Y^2<\infty$, π_0 is bounded away from zero, and $|\widehat{\pi}-\pi_0|_\infty$ and $\|\widehat{\beta}-\beta_0\|$ are $o_p(1)$, we obtain

$$\frac{1}{n}\sum_{i=1}^n\|X_{ki}\widehat{\pi}(X_i)(\widehat{\varphi}_i Y_i - X_{ki}'\widehat{\beta}) - X_{ki}\pi_0(X_i)(\varphi_{0i}Y_i - X_{ki}'\beta_0)\|^2 = o_p(1). \tag{A.3}$$

For π bounded away from one, we have that

$$\Lambda(Z,\tilde{\pi},\beta,\pi) - \Lambda(Z,\tilde{\pi},\beta_0,\pi_0) = \left\{\left(X_k\frac{T-\lambda}{\lambda(1-\lambda)}\cdot\frac{(1-\pi_0)+(1-\pi)}{(1-\pi)^2(1-\pi_0)^2}(D-1)Y\right)(\pi-\pi_0) - X_kX_k'(\beta-\beta_0)\right\}\tilde{\pi}\cdot$$

Therefore, there exists some function, $b(\cdot)$, such that $\|\Lambda(Z,\tilde{\pi},\beta,\pi)-\Lambda(Z,\tilde{\pi},\beta_0,\pi_0)\| \leq b(Z)(|\pi-\pi_0|_\infty + \|\beta-\beta_0\|)|\tilde{\pi}|_\infty$ with $E_M b(Z)<\infty$. This result, along with $K^7/n\to 0$ guarantees that

$$\frac{1}{n}\sum_{i=1}^n\|\widehat{\delta}(X_i)(D_i-\widehat{\pi}(X_i)) - \delta(X_i)(D_i-\pi_0(X_i))\|^2 = o_p(1) \tag{A.4}$$

(see Newey, 1994). Equations (A.3) and (A.4), along with the Triangle and Hölder's inequalities, imply $\widehat{\Sigma}\xrightarrow{p}\Sigma$. As a result, $\widehat{V}\xrightarrow{p}V$. ‖

Proof of Theorem 4.3. First, let us prove that under Assumption 4.2(i)–(iv), $\widehat{\gamma}$ is asymptotically linear with influence function given by equation (17). By Assumption 4.2(ii), $E[XX']$ is non-singular, therefore $(\gamma-\gamma_0)'E[XX'](\gamma-\gamma_0) > 0$ for $\gamma\neq\gamma_0$. As a result, for $\gamma\neq\gamma_0$, $X'\gamma\neq X'\gamma_0$ in a set of positive probability. Since $\pi(v)$ is strictly increasing in $v\in V$, we have that $\pi(X'\gamma)\neq\pi(X'\gamma_0)$ in a set of positive probability. Let $m_\gamma = D\log(\pi(X'\gamma)) + (1-D)\log(1-\pi(X'\gamma))$.

$$E[m_\gamma - m_{\gamma_0}\mid X] = E[D\log(\pi(X'\gamma)/\pi(X'\gamma_0)) + (1-D)\log((1-\pi(X'\gamma))/(1-\pi(X'\gamma_0)))\mid X]$$
$$= \pi(X'\gamma_0)\log(\pi(X'\gamma)/\pi(X'\gamma_0)) + (1-\pi(X'\gamma_0))\log((1-\pi(X'\gamma))/(1-\pi(X'\gamma_0)))$$
$$\leq 0. \tag{A.5}$$

The last inequality follows from $\log\lambda\leq\lambda-1$, for $\lambda>0$, which is strict for $\lambda\neq 1$. Since $\pi(X'\gamma)\neq\pi(X'\gamma_0)$ with positive probability, then the inequality in equation (A.5) is strict with positive probability and $E[m_\gamma - m_{\gamma_0}]<0$ for $\gamma\neq\gamma_0$. Since $\pi(v)$ is bounded away from zero and one in $\mathcal{V}$, then the absolute value of m_γ is bounded by a constant

for any $\gamma \in \Gamma$. This last fact, along with compactness of Γ and S, continuity of $\pi(\cdot)$, and Lemma 2.4 in Newey and McFadden (1994) guarantees that $E[m_\gamma]$ is a continuous function of γ and that

$$\sup_{\gamma\in\Gamma} \left\| \frac{1}{n}\sum_{i=1}^n m_\gamma(Z_i) - E[m_\gamma(Z)] \right\| \xrightarrow{p} 0.$$

These conditions are sufficient for $\widehat{\gamma} \xrightarrow{p} \gamma_0$ (see, *e.g.* Newey and McFadden (1994, Theorem 2.1)).

Now, let us study the asymptotic distribution of $\sqrt{n}(\widehat{\gamma} - \gamma_0)$. Assumption 4.2 along with Lemma 7.6 in van der Vaart (1998) guarantees that, for $p_\gamma = \pi(X'\gamma)^D(1-\pi(X'\gamma))^{(1-D)}$, the map $\gamma \mapsto p_\gamma^{1/2}$ is differentiable in quadratic mean with derivative

$$\dot{l}_\gamma = X\frac{\dot{\pi}(X'\gamma)}{\pi(X'\gamma)(1-\pi(X'\gamma))}(D - \pi(X'\gamma)).$$

Therefore, the map $\gamma \mapsto p_\gamma^{1/2}$ is differentiable in quadratic mean at γ_0 with derivative $\dot{l}_{\gamma_0}$. Take a convex open neighbourhood N_{γ_0} of γ_0 contained in Γ. By Assumption 4.2(ii) and (iv), $\|\partial m_\gamma/\partial\gamma\|$ is bounded by a constant on N_{γ_0}. Therefore, by Theorem 9.19 in Rudin (1976), $|m_{\gamma_1} - m_{\gamma_2}| \leq M\|\gamma_1 - \gamma_2\|$ for any $\gamma_1, \gamma_2 \in N_{\gamma_0}$ and for some constant M. Finally notice that, by Assumption 4.2(ii)–(iv), $E[\dot{l}_{\gamma_0}\dot{l}'_{\gamma_0}] = E[\{\dot{\pi}_0^2/(\pi_0(1-\pi_0))\}XX']$ is non-singular. Therefore, by Theorem 5.39 in van der Vaart (1998), we obtain $n^{1/2}(\widehat{\gamma} - \gamma_0) = n^{-1/2}\sum_{i=1}^n \psi_{\gamma_0}(Z_i) + o_p(1)$.

Let $m(Z,\beta,\gamma) = X_k\pi(X'\gamma)[\varphi(Z,\gamma)Y - X_k'\beta]$. By Assumption 4.2(vi), $E_M[m(Z,\beta,\gamma_0)]$ has a unique zero at $\beta_0 = (E[X_k\pi_0X_k'])^{-1}E_M[X_k\pi_0\varphi_0Y]$. Under the assumptions of the theorem, the function $m(Z,\beta,\gamma)$ is continuous at each $(\beta,\gamma) \in \Theta\times\Gamma$. Since $\pi(X'\gamma)$ is bounded away from zero and one, and both Y and $\|X_k\|$ have finite first moments, then $\|m(Z,\beta,\gamma)\|$ is dominated by a variable with finite first moment. Therefore, by Lemma 2.3 in Newey and McFadden (1994)

$$\sup_{(\beta,\gamma)\in\Theta\times\Gamma} \left\| \frac{1}{n}\sum_{i=1}^n m(Z_i,\beta,\gamma) - E_M[m(Z,\beta,\gamma)] \right\| \xrightarrow{p} 0, \tag{A.6}$$

and $E[m(Z,\beta,\gamma)]$ is continuous at each $(\beta,\gamma) \in \Theta\times\Gamma$. By the Triangle inequality,

$$\begin{aligned}\sup_{\beta\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^n m(Z_i,\beta,\widehat{\gamma}) - E[m(Z,\beta,\gamma_0)] \right\| &\leq \sup_{\beta\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^n m(Z_i,\beta,\widehat{\gamma}) - E[m(Z,\beta,\widehat{\gamma})] \right\| \\ &\quad + \sup_{\beta\in\Theta} \|E[m(Z,\beta,\widehat{\gamma})] - E[m(Z,\beta,\gamma_0)]\|. \end{aligned}\tag{A.7}$$

Equation (A.6) implies that the first term of the R.H.S. of equation (A.7) is $o_p(1)$. Continuity of $E[m(Z,\beta,\gamma)]$ and consistency of $\widehat{\gamma}$ give pointwise convergence for the second term of the R.H.S. of equation (A.7), uniform convergence holds by Θ being compact. These conditions guarantee $\widehat{\beta} \xrightarrow{p} \beta_0$.

It can be shown that $\|m(Z,\beta_1,\gamma_1) - m(Z,\beta_2,\gamma_2)\| \leq \|X_k\|^2\cdot\|\beta_1-\beta_2\| + (C\cdot|Y|\cdot\|X_k\| + \|X_k\|^2\|\beta_2\|)\cdot\|\pi(X'\gamma_1) - \pi(X'\gamma_2)\|$, for some $C > 0$ and $\beta_1, \beta_2 \in \Theta$. Using Theorem 9.19 in Rudin (1976), it is easy to show that for γ_1, γ_2 in some open ball containing γ_0, and $X \in S$, $\|\pi(X'\gamma_1) - \pi(X'\gamma_2)\| \leq M\|\gamma_1 - \gamma_2\|$, for some $M > 0$. This Lipschitz property, along with the existence of finite second moments for Y and X_k, implies that the class of functions $\{m(Z,\beta,\gamma) : \|\beta-\beta_0\| < c, \|\gamma-\gamma_0\| < c\}$ is Donsker for some $c > 0$ (see, *e.g.* van der Vaart (1998, p. 271)). In addition, $E_M\|m(Z,\beta,\gamma) - m(Z,\beta_0,\gamma_0)\|^2 \to 0$ as $(\beta,\gamma) \to (\beta_0,\gamma_0)$. Existence of finite second moments for $\|X_k\|$ also implies that $\partial E_M[m(Z,\beta,\gamma)]/\partial\beta = -E[X_k\pi(X'\gamma)X_k']$ which is non-singular in a neighbourhood of γ_0 by continuity. Notice that M_{γ_0} is the derivative of $E_M[m(Z,\beta_0,\gamma)]$ at γ_0. Applying the delta method: $n^{1/2}E_M[m(Z,\beta_0,\widehat{\gamma})] = n^{1/2}M_{\gamma_0}(\widehat{\gamma} - \gamma_0) + o_p(1)$. Now, since $\widehat{\gamma}$ is asymptotically linear, apply Theorem 5.31 in van der Vaart (1998) to get

$$n^{1/2}(\widehat{\beta} - \beta_0) = Q^{-1}\frac{1}{n^{1/2}}\sum_{i=1}^n m(Z_i,\beta_0,\gamma_0) + M_{\gamma_0}\psi_{\gamma_0}(Z_i) + o_p(1).$$

Then, existence of second moments of ψ implies the result of the theorem. ‖

Proof of Theorem 4.4. Like in the proof of Theorem 4.2, notice that

$$\begin{aligned} &X_{ki}\widehat{\pi}(X_i)(\widehat{\varphi}_iY_i - X_{ki}'\widehat{\beta}) - X_{ki}\pi_0(X_i)(\varphi_{0i}Y_i - X_{ki}'\beta_0) \\ &\quad = X_{ki}\left(\frac{T_i-\lambda}{\lambda(1-\lambda)}\cdot\frac{D_i-1}{(1-\widehat{\pi}(X_i))(1-\pi_0(X_i))}\right)Y_i(\widehat{\pi}(X_i) - \pi_0(X_i)) - X_{ki}X_{ki}'(\widehat{\beta} - \beta_0). \end{aligned}$$

By the Lipschitz property shown above for $\pi(\cdot)$, and $E_MY^2 < \infty$, we obtain

$$\frac{1}{n}\sum_{i=1}^n \|m(Z_i,\widehat{\beta},\widehat{\gamma}) - m(Z_i,\beta_0,\gamma_0)\|^2 = o_p(1). \tag{A.8}$$

Continuity of $\dot{\pi}$ and Lemma 4.3 in Newey and McFadden (1994) imply that $\|\widehat{M}_{\widehat{\gamma}} - M_{\gamma_0}\| = o_p(1)$. Using similar arguments, it can be easily seen that boundedness of $\ddot{\pi}$ implies that $n^{-1}\sum_{i=1}^{n}\|\widehat{\psi}_{\widehat{\gamma}}(Z_i) - \psi_{\gamma_0}\|^2 = o_p(1)$. Apply the Triangle and Hölder's inequalities to obtain the desired result. ‖

Acknowledgements. I thank Joshua Angrist, Jinyong Hahn, Guido Imbens, José Machado, seminar participants at Harvard/MIT, ITAM, the 2000 American Statistical Association Meetings, the 2003 Econometric Society North American Winter Meetings, and the 2003 Bank of Portugal Conference on Labour Market Reform, the referees, and two editors (Orazio Attanasio and Bernard Salanié) for helpful comments.

REFERENCES

ABADIE, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models", *Journal of Econometrics*, **113**, 231–263.
ACEMOGLU, D. and ANGRIST, J. D. (2001), "Consequences of Employment Protection? The Case of the Americans with Disabilities Act", *Journal of Political Economy*, **109**, 915–957.
ANGRIST, J. D. and KRUEGER, A. B. (1999), "Empirical Strategies in Labor Economics", in O. Ashenfelter and D. Card (eds.) *Handbook of Labor Economics*, Vol. 3A (Amsterdam: Elsevier Science) 1277–1366.
ASHENFELTER, O. (1978), "Estimating the Effect of Training Programs on Earnings", *Review of Economics and Statistics*, **60**, 47–57.
ASHENFELTER, O. and CARD, D. (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effects of Training Programs", *Review of Economics and Statistics*, **67**, 648–660.
ATHEY, S. and IMBENS, G. (2002), "Identification and Inference in Nonlinear Difference-in-Differences Models" (Mimeo, Stanford University).
BESLEY, T. and CASE, A. (1994), "Unnatural Experiments? Estimating the Incidence of Endogenous Policies" (Working Paper No. 4956, National Bureau of Economic Research).
BLUNDELL, R., COSTA DIAS, M., MEGHIR, C. and VAN REENEN, J. (2001), "Evaluating the Employment Impact of a Mandatory Job Search Assistance Program" (Mimeo, UCL).
BLUNDELL, R. and MACURDY, T. (1999), "Labor Supply: A Review of Alternative Approaches", in O. Ashenfelter and D. Card (eds.) *Handbook of Labor Economics*, Vol. 3A (Amsterdam: Elsevier Science) 1559–1695.
CARD, D. (1990), "The Impact of the Mariel Boatlift on the Miami Labour Market", *Industrial and Labor Relations Review*, **44**, 245–257.
CARD, D. (1992), "Do Minimum Wages Reduce Employment? A Case Study of California, 1987–1989", *Industrial and Labor Relations Review*, **46**, 38–54.
CARD, D. and KRUEGER, A. B. (1994), "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania", *American Economic Review*, **84**, 772–793.
CORAK, M. (2001), "Death and Divorce: The Long Term Consequences of Parental Loss on Adolescents", *Journal of Labor Economics*, **19**, 682–715.
EISSA, N. and LIEBMAN, J. B. (1996), "Labor Supply Response to the Earned Income Tax Credit", *Quarterly Journal of Economics*, **111**, 605–637.
FINKELSTEIN, A. (2002), "The Effect of Tax Subsidies to Employer-Provided Supplementary Health Insurance: Evidence from Canada", *Journal of Public Economics*, **84**, 305–339.
GARVEY, G. T. and HANKA, G. (1999), "Capital Structure and Corporate Control: The Effect of Antitakeover Statutes on Firm Leverage", *Journal of Finance*, **54**, 519–546.
HECKMAN, J. J. (1990), "Varieties of Selection Bias", *American Economic Review*, **80**, 313–318.
HECKMAN, J. J., ICHIMURA, H., SMITH, J. and TODD, P. E. (1998), "Characterizing Selection Bias using Experimental Data", *Econometrica*, **66**, 1017–1098.
HECKMAN, J. J., ICHIMURA, H. and TODD, P. E. (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies*, **64**, 605–654.
HERNÁN, M. A., BRUMBACK, B. and ROBINS, J. M. (2001), "Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments", *Journal of the American Statistical Association*, **96**, 440–448.
HORVITZ, D. G. and THOMPSON, D. J. (1952), "A Generalization of Sampling without Replacement from a Finite Universe", *Journal of the American Statistical Association*, **47**, 663–685.
ICHIMURA, H. and LINTON, O. (2002), "Asymptotic Expansions for some Semiparametric Program Evaluation Estimators" (Mimeo, London School of Economics).
IMBENS, G., LIEBMAN, J. B. and EISSA, N. (1997), "The Econometrics of Difference in Differences" (Mimeo, Harvard University).
IMBENS, G. W., HIRANO, K. and RIDDER, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", *Econometrica*, **71** (4), 1161–1189.
MEYER, B. D. (1995), "Natural and Quasi-Experiments in Economics", *Journal of Business & Economic Statistics*, **13**, 151–161.
MEYER, B. D., VISCUSI, W. K. and DURBIN, D. L. (1995), "Worker's Compensation and Injury Duration: Evidence from a Natural Experiment", *American Economic Review*, **85**, 322–340.
NEWEY, W. K. (1994), "The Asymptotic Variance of Semiparametric Estimators", *Econometrica*, **62**, 1349–1382.
NEWEY, W. K. (1997), "Convergence Rates and Asymptotic Normality for Series Estimators", *Journal of Econometrics*, **79**, 147–168.

NEWEY, W. K. and MCFADDEN, D. (1994), "Large Sample Estimation and Hypothesis Testing", in R. F. Engle and D. McFadden (eds.) *Handbook of Econometrics*, Vol. 4 (Amsterdam: Elsevier Science).
ROEHRIG, C. S. (1988), "Conditions for Identification in Nonparametric and Parametric Models", *Econometrica*, **56**, 433–447.
ROSENBAUM, P. and RUBIN, D. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, **70**, 41–55.
RUBIN, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, **66**, 688–701.
RUBIN, D. B. (1977), "Assignment to Treatment of the Basis of a Covariate", *Journal of Educational Statistics*, **2**, 1–26.
RUDIN, W. (1976) *Principles of Mathematical Analysis* (New York: McGraw-Hill).
VAN DER VAART, A. W. (1998) *Asymptotic Statistics* (Cambridge: Cambridge University Press).
WHITE, H. (1981), "Consequences and Detection of Misspecified Nonlinear Regression Models", *Journal of the American Statistical Association*, **76**, 419–433.
WOOLDRIDGE, J. M. (2001), "Estimating Average Partial Effects under Conditional Moment Independence Restrictions" (Mimeo, Michigan State University).
WOOLDRIDGE, J. M. (2002) *Econometric Analysis of Cross Section and Panel Data* (Cambridge: MIT Press).