Universität St.Gallen

The Estimation of Causal Effects by Difference-in-Difference Methods

Michael Lechner

October 2011 Discussion Paper no. 2010-28

Department of Economics University of St. Gallen

Editor:	Martina Flockerzi University of St. Gallen Department of Economics Varnbüelstrasse 19 CH-9000 St. Gallen	
	Phone	+41 71 224 23 25
	Fax	+41 71 224 31 35
	Email	vwaabtass@unisg.ch
Publisher:	Department of Economics University of St. Gallen Varnbüelstrasse 19 CH-9000 St. Gallen	
	Phone	+41 71 224 23 25
	Fax	+41 71 224 31 35
Electronic Publication:	http://www.vwa.unisg.ch	

The Estimation of Causal Effects by Difference-in-Difference Methods [1]

Michael Lechner[2]

Author's address: Prof. Dr. Michael Lechner
Swiss Institute for Empirical Economic Research (SEW-HSG)
Varnbüelstrasse 14
CH-9000 St. Gallen
Phone +41 71 224 2320
Fax +41 71 224 2302
Email Michael.lechner@unisg.ch
Website www.sew.unisg.ch/lechner

[1] This project received financial support from the St. Gallen Research Centre of Aging, Welfare, and Labour Market Analysis (SCALA).The paper was presented at the econometrics breakfast at the University of St. Gallen and benefited from comments by seminar participants. I thank Christina Felfe, Alex Lefter, and Conny Wunsch for carefully reading a previous draft of this manuscript, Patrick Puhani for an interesting discussion on some issues related to limited dependent outcome variables, and Jeff Smith for additional useful references. The usual disclaimer applies.

[2] Research Associate of ZEW, Mannheim, and a Research Fellow of CEPR, London, CESifo, Munich, IAB, Nuremberg, IZA, Bonn, and PSI, London.

Abstract

This survey gives a brief overview of the literature on the difference-in-difference (DiD) estimation strategy and discusses major issues using a treatment effect perspective. In this sense, this survey gives a somewhat different view on DiD than the standard textbook discussion of the difference-in-difference model, but it will also not be as complete as the latter. This survey contains also a couple of extensions to the literature, for example, a discussion of and suggestions for non-linear DiD as well as DiD based on propensity-score type matching methods.

Keywords

Causal inference, counterfactual analysis, before-after-treatment-control design, control group design with pretest and posttest.

JEL Classification

C21, C23, C31, C33.

1. Introduction

The Difference-in-Difference (DiD) approach is a research design for estimating causal effects. It is popular in empirical economics, for example, to estimate the effects of certain policy interventions and policy changes that do not affect everybody at the same time and in the same way. It is used in other social sciences as well.[1] DiD could be an attractive choice when using research designs based on controlling for confounding variables or using instrumental variables is deemed unsuitable, and at the same time, pre-treatment information is available.[2] The DiD design is usually based on comparing de facto four different groups of objects. Three of these groups are not affected by the treatment. In many applications, 'time' is an important variable to distinguish the groups.[3] Besides the group which already received the treatment (post-treatment treated), these groups are the treated prior to their treatment (pre-treatment treated), the nontreated in the period before the treatment occurs to the treated (pre-treatment nontreated), and the nontreated in the current period (post-treatment nontreated).[4] The idea of this empirical strategy is that if the two treated and the two nontreated groups are subject to the same time trends, and if the treatment has had no effect in the pre-treatment

[1] In other social sciences the DiD approach is also denoted as "untreated control group design with independent pretest and posttest samples" or "control group design with pretest and posttest". See, for example, Cook and Campbell (1979), Rosenbaum (2001), and Shadish, Cook, and Campbell (2002) for further references.

[2] Following the literature, the event for which we want to estimate the causal effect is called the *treatment*. The *outcome* denotes the variable that will be used to measure the effect of the treatment. Outcomes that would be realised if a specific treatment has, or would have been applied, are called *potential outcomes*. A variable is called *confounding* if it is related to the treatment and the potential outcomes. A variable is called an *instrument* if it influences the treatment but not the potential outcomes.

[3] As the concept of time is only used to define a group that is similar to the treated group with respect to relevant unobservable variables and whose members have not (yet) participated, any other characteristic may be used instead of time as well, as long as the formal conditions given below are fulfilled.

[4] When a data set is available in which everybody is observed in all periods, there will be just two groups with outcomes measured before and after the treatment.

period, then an estimate of the 'effect' of the treatment in a period in which it is known to have none, can be used to remove the effect of confounding factors to which a comparison of post-treatment outcomes of treated and nontreated may be subject to. This is to say that we use the mean changes of the outcome variables for the nontreated over time and add them to the mean level of the outcome variable for the treated prior to treatment to obtain the mean outcome the treated would have experienced if they had not been subject to the treatment.

This survey presents a brief overview of the literature on the difference-in-difference estimation strategy and discusses major issues mainly using a treatment effect perspective (and language) that allows, in our opinion, more general considerations than the classical regression formulation that still dominates the applied work. In this sense, this survey might give a somewhat different perspective than the standard text book discussion of the difference-in-difference design, but it will not be as complete as the latter. Thus, this paper is more of a complement than a substitute to the excellent text type discussions of the difference-in-difference approach that are already available in the literature (e.g. Angrist and Pischke, 2009, Blundell and Costa Dias, 2009, and Imbens and Wooldridge, 2009).

This paper focuses on the case of only two differences although the basic ideas of difference-in-difference (DiD) estimation could be extended to more than two dimensions to create difference-in-difference-in-difference-in-… estimators.[5] However, the basic ideas of the approach of taking multiple differences are already apparent with two dimensions. Thus, we refrain from addressing these higher dimensions to keep the discussion as focused as possible.

[5] For example, Yelowitz (1995) analyses the effects of losing public health insurance on labour market decisions in the US by using Medicaid eligibility that varies over time, state and age (of the child in the household). Another example for a triple difference is the paper by Ravallion, Galasso, Lazo, and Philipp (2005) who analyse the effects of a social programme based on a comparison of participants with nonparticipants and ex-participants.

The outline of this survey is as follows: The next section gives a historical perspective and discusses some interesting applications. Section 3, which is the main part of this survey, discusses identification issues at length. Section 4 concerns DiD specific issues related to estimation, including a discussion of propensity score matching estimation of DiD models. Section 5 discusses some specific issues related to inference, and section 6 considers important practical extensions to the basic approach. Section 7 concludes. Some short proofs are relegated to a technical appendix.

2. The History of DiD

The method of difference-in-difference estimation is a well established econometric tool and, although there are a couple of open issues, the main components of this approach are well understood.[6] The first scientific study using explicitly a difference-in-difference approach known to the author of this survey is the study by Snow (1855).[7] Snow (1855) was interested in the question whether cholera was transmitted by (bad) air or (bad) water. He used a change in the water supply in one district of London, namely the switch from polluted water taken from the Themes in the centre of London to a supply of cleaner water taken upriver. The basic idea of his study is described by a quote from the introduction to (the reprint of) his book by Ralph R. Frerichs: "… Another investigation cited in his book which drew praise for Snow was his recognition and analysis of a natural experiment involving two London water companies, one polluted with cholera and the other not. He demonstrated that

[6] Expositions of this approach at an advanced textbook level are provided for example by Meyer (1995), Angrist and Krueger (1999), Heckman, LaLonde, and Smith (1999), Angrist and Pischke (2009), Blundell and Costa Dias (2009), and Imbens and Wooldridge (2009). For one of the rather rare treatments of this topic in the statistics literature see Rosenbaum (2001).

[7] It is also mentioned by Angrist and Pischke (2009). Besides this approach, Snow also used 'before-after' type of methods when analysing a cholera outbreak related to one particular water street-pump in Broad Street, London. See the short article by Snow (1854) himself or the interesting account given by Lai (2011).

persons who received contaminated water from the main river in London had much higher death rates due to cholera. Most clever was his study of persons living in certain neighbourhoods supplied by both water companies, but who did not know the source of their water. He used a simple salt test to identify the water company supplying each home. This reduced misclassification of exposure, and provided him with convincing evidence of the link between impure water and disease." Obviously, using close neighbourhoods is clever, as they are probably exposed to similar air quality. It is worth adding that Snow also had data on death rates in those neighbourhoods prior to the switch of water supply. He used them to correct his estimates for other features of these neighbourhoods that could have also lead to differential death rates. In that way, the first difference-in-difference estimate had an important impact for public health in a scientific as well as in a very practical, life-saving way. Probably the most important reason for this impact was the high credibility of Snow's (1855) clever research design.

Later on, the DiD type of approach became relevant also for other fields, like psychology, for example. Rose (1952) investigated the effects of a regime of 'mandatory mediation' on work stoppages by a difference-in-difference design. The following quote from his article reveals the key issues: "To test the effectiveness of mandatory mediation in preventing work stoppages, it is necessary to make two simultaneous comparisons: (1) comparisons of states with the law to states without the law; (2) comparisons of the former states before and after the law is put into operation. The first comparison can be achieved by taking percentages of each of the three states to the total United States, for the measures used. The second comparison can be achieved by setting the date of the passage of the law at zero for each of the states. Figure 1 indicates both comparisons simultaneously. …" (Rose, 1952, p. 191).

In economics, the basic idea of the difference-in-difference approach appeared early. Obenauer and von der Nienburg (1915) analysed the impact of a minimum wage by using an introduction of the minimum wage (in the retail industry) in the state of Oregon that, for a particular group of employees, led to higher wage rates in Portland, the largest city, compared to the rest of the state. Therefore, the authors documented the levels of various outcome variables for the different groups of employees in Portland before and after the introduction of the minimum wage and compared the respective changes to those computed for Salem, which is also located in Oregon and thought to be comparable to Portland.

Another early application in economics has been conducted by Lester (1946). He was concerned with the effects of wages on employment. He based his analysis on a survey of firms that had operations in both the northern and the southern US states. His idea was to compare employment levels, before and after various minimum wage rises, of groups of firms with low average wages to groups of firms with higher wage levels. The wage bills of the latter were naturally only mildly affected, if at all, by the rise in the minimum wage.

An important aspect of DiD estimation highlighted by these early applications is that it does not require high powered computational effort to compute the basic DiD estimates, at least as long as further covariates are not needed and no complicated inference methods are used. This simplicity certainly makes some of the intuitive appeal of DiD (and is also responsible for some of its weaknesses that will be discussed below).

Over time the field of economics developed a literature that, like Rose (1952), uses changes in state laws and regulations to define pre-treatment periods (prior to the introduction of the policy) and unaffected comparison groups (states having a different policy than the one of interest). One early example is the analysis of the price elasticity of liquor sales which has been conducted by Simon (1966): "The essence of the method is to examine the "before" and "after" sales of a given state, sandwiched around a price change and standardized with the sales figures of states that did not have a price change. The standardizing removes year-to-year fluctuations and the trend. We then pool the results of as many quasi-experimental "trial" events as are available." (Simon, 1966, p. 196).[8] The question of the price reaction of liquor sales has also been addressed in another important study by Cook and Tauchen (1982) exploiting the state variation of the exercise tax for liquor with a DiD approach.

Later on, DiD designs have been used to address many other important policy issues, like the effects of minimum wages on employment (e.g. Card and Krueger, 1994), the effects of training and other active labour market programmes for unemployed on labour market outcomes (e.g. Ashenfelter, 1978, Ashenfelter and Card, 1985, Heckman and Robb, 1986, Heckman and Hotz, 1989, Heckman, Ichimura, Smith, and Todd, 1998, Blundell, Meghir, Costa Dias, and van Reenen, 2004), the effect of immigration on the local labour market (e.g. Card, 1990), or the analysis of labour supply (e.g. Blundell, Duncan, and Meghir, 1998).

There is also a considerable literature analysing various types of labour market regulations with DiD designs. Meyer, Viscusi, and Durbin (1995) consider the effect of workers injury compensation on injury related absenteeism. Waldfogel (1998) looks at maternity leave regulation. Acemoglu and Angrist (2001) investigate the effects of the American with Disabilities Act, and Besley and Burgess (2004) consider the impact of more

[8] Note that in those times liquor prices were fixed by the states. It is also interesting that Simon (1966) relates this type of approach to the experimental literature, a relation that is still frequently used: "This investigation uses a method that has features of both the cross section and the time series. Though it has not been used by economists, to my knowledge, it is very similar to designs used for experiments in psychology and the natural sciences and to sociological paradigms. Because this method is not an experiment, though similar to one, we call it the "quasi-experimental" method." (Simon, 1966, p. 195). Economists also use the term of a natural experiment (e.g., Meyer, 1995).

or less worker friendly labour regulations on growth in a developing country (India), to mention only some important examples.[9]

3. Models, effects, and identification

3.1 Notation and effects

We start with a simple set-up to show the main ideas and problems of DiD. The treatment variable, denoted by D, is binary, i.e. $d \in \{0,1\}$.[10] We have measurements of the various variables at most in two time periods, *T*, $t \in \{0,1\}$. Period zero indicates a time period before the treatment (pre-treatment period) and period one indicates a time period after the treatment took place (post-treatment period). Assuming that the treatment happens between the two periods means that every member of the population is untreated in the pre-treatment period. We are interested in discovering the mean effect of switching *D* from zero to one on some outcome variables. Therefore, we define 'potential' outcome variables indexed by the potential states of the treatment, so that Y_t^d denotes the outcome that would be realized for a specific value of *d* in period *t*. The outcome that is realized (and thus observable) is denoted by Y (not indexed by *d*). Finally, denote some further observable variables by *X*. They are assumed not to vary over time. Later on, several of the restrictions implied by this framework, like time constant *X* and observing only two periods, will be relaxed. However, imposing them initially helps exposing the main ideas without unnecessary complications.

[9] DiD estimates may also be the starting point (first stage) of instrumental variable estimation strategies, like in the analysis of an Indonesian school construction programme by Duflo (2001). These types of extensions of the DiD approach are however not the focus of this partial survey.

[10] We use the convention that capital letters denote random variables and small letters denote specific values or realisations.

Having defined the notation, the mean treatment effects can be derived. In line with the literature on causal inference (see Imbens and Wooldridge, 2009, for a recent and comprehensive survey) we would like to consider effects for the treated, the nontreated, and the population at large, separately. However, it will be shown below that the latter two are only identified under considerably stronger DiD assumptions that are not attractive in typical DiD applications. Hence, their identification is not an important topic in this survey. The average treatment effect on the treated in period *t* is defined in the usual way:

$$\begin{aligned} ATET_t &= E\left(Y_t^1 - Y_t^0 \mid D=1\right) = \\ &= E\left[\underbrace{E\left(Y_t^1 - Y_t^0 \mid X = x, D=1\right)}_{\theta_t(x)} \mid D=1\right] = \\ &= E_{X|D=1}\theta_t(x). \end{aligned} \qquad (1)$$

ATET_t denotes the so-called average treatment effect on the treated and $\theta_t(x)$ are the corresponding effects in the respective subpopulations defined by the value of *X* being *x*.[11]

3.2 Identification

Although the DiD approach is frequently used within the linear regression model, we start by studying the properties of this approach in a nonparametric framework, which is common in the econometric literature on causal inference. Among many other virtues, it has the important advantage that the treatment effects are naturally allowed to be heterogeneous across the members of the population.

[11] Recently, Bonhomme and Sauder (2011) extended the DiD logic to the distribution of the outcome variables using characteristic functions. Hence, their approach can be used to recover quantile treatment effects as well. For the sake of simplicity, the rest of this survey sticks to mean effects, however.

3.2.1 *Identifying assumptions in the standard nonparametric model*

As mentioned before, the main idea of the difference-in-difference identification strategy is to compute the difference of the mean outcomes of treated and controls after the treatment and subtract the outcome difference that had been there already before the treatment had any effect (conditional on a given value of *x*). If the assumptions formulated below hold, this strategy will indeed identify a mean causal effect.

To motivate the assumptions in this section we use an empirical example from the literature on active labour market policy evaluation. For the sake of this example, suppose that we are interested in estimating the earnings effect of participation in a training programme for the unemployed based on micro data containing information on the periods before and after training as well as on programme participants and nonparticipants.

The first assumption implies that one, and only one, of the potential outcomes is indeed observable for every member of the population. This assumption, sometimes called the observation rule, follows from the so-called Stable Unit Treatment Value assumption (SUTVA, Rubin, 1977). Importantly, it implies that the treatments are completely represented and, in particular, that there are no relevant interactions between the members of the population.

$$Y_t = dY_t^1 + (1-d)Y_t^0, \quad \forall t \in \{0,1\}. \qquad (SUTVA)$$

If SUTVA is violated, we observe neither of the two potential outcomes and conventional microeconometric evaluation strategies break down.[12]

[12] Manski (2011) calls this assumption *individualistic treatment response* and analyses identification if it does not hold. Miguel and Kremer (2004) is an early study analyzing this phenomenon of spill-over effects or treatment externalities. They point out that in such cases conducting experiments based on group randomization may be more adequate than randomizing individuals.

Example for a violation of SUTVA: If the training programme is very large, it may change the equilibrium wages in the labour market by influencing skill-specific demand and supply relations. For example, while some unemployed are participating in the training courses, it will be easier for the nonparticipants to find a job than without the existence of the programme. However, after the programme, nonparticipants with skills comparable to those obtained in the training programme will have more difficulties in finding a job because the supply in this skill group is now larger compared to the hypothetical world without the training programme. Thus, the nonparticipants' outcome is not the same outcome as the one they would have experienced in a world without the programme. Therefore, SUTVA is violated. Clearly, SUTVA is more relevant for period one than for period zero, but even in period zero it may play a role. An example for this is that individuals anticipating their future programme participation reduce their job search efforts. Thus, it is easier for the searching future nonparticipants to find a job.

The next assumption concerns the conditioning variables *X* because the main behavioural assumptions are supposed to hold conditional on some covariates *X*. To make sure that this conditioning does not destroy identification, it is assumed that the components of *X* are not influenced by the treatment. This assumption is called EXOG (exogeneity) and formalized applying the potential outcome notation to the control variables, X^d:[13]

$$X^1 = X^0 = X, \qquad \forall x \in \chi. \qquad \textit{(EXOG)}$$

χ denotes the subspace of *X* that is of interest to the researcher.

Examples for violations of EXOG: It is particularly likely that variables that are measured after the treatment is known (like post-treatment job satisfaction) may be influenced by the treatment. Measuring variables prior to the treatment, however, does not automatically ensure exogeneity: individuals may anticipate the treatment and change behaviour accordingly. If they do this in any way that impacts also the outcome variables, endogeneity of such control variables is a problem. Note that given our overall set of assumptions, variables that cannot change over time are exogenous by construction, because we consider

[13] Lechner (2008) shows that this assumption is too strong as one needs only to rule out that any influence of *D* on *X* does not affect the potential outcomes. Nevertheless, we keep it here for convenience.

a time varying treatment. The problem with conditioning on a variable influenced by the treatment can easily be seen with an extreme example. Suppose the observed outcome variable *Y_1* would be included among the control variables, then by construction $\theta_1(y_1)=0$. In other words, conditioning on an endogenous variable is like estimating only that part of the causal effect that is not already captured by the particular endogenous variable.

While SUTVA and EXOG are standard assumptions in microeconometric causal studies, the assumption that in the pre-treatment period the treatment had no effect on the pre-treatment population (NEPT) is specific to the DiD.

$$\theta_0(x)=0; \qquad \forall x \in \chi. \qquad \textit{(NEPT)}$$

NEPT also rules out behavioural changes of the treated that influence their pre-treatment outcome in anticipation of a future treatment.[14]

Example: This is very similar to the exogeneity condition but now applied to the pre-treatment outcomes instead of the covariates. In the training example when the outcome of interest is unemployment, NEPT would be violated if individuals decided not to search for a job because they know (or plausibly anticipate in a way not captured by *X*) that they will participate in an attractive training programme.

Next, we state the defining assumptions for the DiD approach, namely the 'common trend' (CT) and 'bias stability' (BS) assumptions. The common trend assumption is given by the following expression:

[14] Note that the observation rule (SUTVA) does not exclude the possibility that the treatment has an effect before it starts (anticipation), because we observe the treatment outcome of the treated before and after the treatment. Hence, in this paper we separate SUTVA from the assumption that the treatment has no effect in period zero. Some papers combine these assumptions by defining the observation rule in a way such that in period zero we always observe the nonparticipation outcome. In that case, explicitly assuming NEPT is redundant as NEPT is already implied by this type of observation rule.

$$\begin{aligned}
&E\left(Y_1^0 \mid X = x, D = 1\right) && -E\left(Y_0^0 \mid X = x, D = 1\right) = \\
&E\left(Y_1^0 \mid X = x, D = 0\right) && -E\left(Y_0^0 \mid X = x, D = 0\right) = \qquad\qquad (CT)\\
&E\left(Y_1^0 \mid X = x\right) && -E\left(Y_0^0 \mid X = x\right); \qquad \forall x \in \chi.
\end{aligned}$$

This assumption states that the differences in the expected potential nontreatment outcomes over time (conditional on *X*) are unrelated to belonging to the treated or control group in the post-treatment period. This is the key assumption of the DiD approach. It implies that if the treated had not been subjected to the treatment, both subpopulations defined by *D=1* and *D=0* would have experienced the same time trends conditional on *X*. Thus, this also implies that the covariates *X* should be selected such that they capture all variables that would lead to differential time trends (in other words, select control variables for which the time trends of the nonparticipation outcome differ for different values of *X*, and at the same time for which the distribution of *X* differs between treated and controls). The common trend assumption already gives the intuition of the identification proof below. As the nontreatment potential outcomes share the same trend for treated and nontreated, any deviation of the trend of the observed outcomes of the treated from the trend of the observed outcomes of the nontreated will be directly attributed to the effect of the treatment and not to differences in other characteristics of the treatment and control group.

Example for a violation of the common trend assumption: Suppose that unemployed individuals from shrinking sectors are particularly likely to be admitted into the training programme. Thus, unemployed who worked in such declining sectors are overrepresented in the group of programme participants. As these workers possess sector specific skills, the reemployment chances of unemployed from declining sectors (sectors that continuously reduce their demand for labour) are likely to deteriorate faster than the reemployment chances of unemployed searching jobs in sectors in which the demand for labour increases over time. Since the respective shares of these types of unemployed differ in the treated and control groups, the common trend assumption is violated unconditionally. However, it may hold if the sector of the last employment is used as a control variable.

As another example, it has been observed by Ashenfelter (1978) that trainees from public training programmes suffer a larger drop in earnings prior to training than nontrainees. Suppose these drops are due to negative 'idiosyncratic temporary shocks'. The temporary nature of these shocks implies that individuals who received the shock will recover faster than other individuals once the effect of the shock disappears. This reaction directly violates the common trend assumption (see, for example, the exposition of this problem in Heckman, LaLonde, and Smith, 1999, or in Blundell and Costa Dias, 2009).

Alternatively one can see the intuition behind the DiD approach by considering the possibility of estimating the effects of D in both periods while (falsely) pretending that a selection-on-observables assumption would be correct conditional on X. If NEPT is true, then a nonzero effect in the estimation of the effects of D on Y_0 (in the pre-treatment period) implies that the estimator is biased and inconsistent and the selection-on-observables assumption is implausible. If (and only if) this bias is constant over time, it can be used to correct the estimate of the effect of D on Y_1, i.e. in the post-treatment period, which is the effect we are interested in (e.g., Heckman, Ichimura, Todd, and Smith, 1998). Therefore, the assumption corresponding to this intuition may be called 'constant bias' assumption (CB) and is formalized by:

$$\begin{aligned}
&E\left[Y_0^0 \mid X=x, D=1\right]-E\left[Y_0^0 \mid X=x, D=0\right] \quad [=Bias_0(x)]= \\
&E\left[Y_1^0 \mid X=x, D=1\right]-E\left[Y_1^0 \mid X=x, D=0\right] \quad [=Bias_1(x)], \qquad \forall x \in \chi.
\end{aligned} \tag{2}$$

By simple rewriting of the CT and CB assumptions we see that they are identical:

$$\begin{aligned}
Bias_1(x)-Bias_0(x) &= \left[E\left(Y_1^0 \mid X=x, D=1\right)-E\left(Y_1^0 \mid X=x, D=0\right)\right] \\
&\quad -\left[E\left(Y_0^0 \mid X=x, D=1\right)-E\left(Y_0^0 \mid X=x, D=0\right)\right] \\
&= \left[E\left(Y_1^0 \mid X=x, D=1\right)-E\left(Y_0^0 \mid X=x, D=1\right)\right] \\
&\quad -\left[E\left(Y_1^0 \mid X=x, D=0\right)-E\left(Y_0^0 \mid X=x, D=0\right)\right].
\end{aligned} \tag{3}$$

From these assumptions it is obvious that identification relies on the counterfactual difference $E\left(Y_1^0 \mid X = x, D = 1\right)$ - $E\left(Y_0^0 \mid X = x, D = 1\right)$ being identical to the observable difference $E\left(Y_1 \mid X = x, D = 0\right)$ - $E\left(Y_0 \mid X = x, D = 0\right)$. Therefore, it is necessary that observations with characteristics *x* exist in all four subsamples. This is guaranteed by the so-called common support assumption:

$$P\left[TD = 1 \mid X = x, (T,D) \in \left\{(t,d),(1,1)\right\}\right] < 1;\ \ \forall (t,d) \in \left\{(0,1),(0,0),(1,0)\right\};\ \ \forall x \in \chi. \quad \textit{(COSU)}$$

Example of a violation of COSU: If participation in a training programme was compulsory for unemployed below 25, and unemployed below 25 years were subject to a different trend than unemployed above 25 years (so that this agecut-off is required as conditioning variable to make the common trend assumption plausible), then the common support assumption would be violated because there would not be any nonparticipants of age 25 or younger.

This assumption is formulated in terms of observable quantities and is thus testable. All the other (identifying) assumptions mentioned above are formulated in terms of unobservable random variables and are thus not testable. If common support did not hold for all values of *X*, a common practice would be to redefine the population for which we estimate the average treatment effects of interest to those treated types, defined by the values of *X* (χ), that are observable in all four subpopulations. Alternatively, one may has to be satisfied with partial identification of the original parameter.[15]

3.2.2 Proof of identification of the average effect on the treated

Although the proof of identification is straightforward and available in the literature, because of its instructive nature it is repeated below.

[15] See Lechner (2008b) for such a bounding strategy in the case of matching. This strategy could be directly transferred to the context of DiD estimation.

First, note that once the conditional-on-*X* effects, $\theta_1(x)$, are identified for all relevant values of *x*, $ATET_1$ is identified as well ($ATET_0$ is zero because of the NEPT assumption). This property holds because of the common support assumption implying that *X* has support in all four subpopulations defined by the different values of *D* and *T*. Hence, the identification proof shows identification of $\theta_1(x)$ only.

$$\begin{aligned}\theta_1(x) &= E\left(Y_1^1 - Y_1^0 \mid X = x, D = 1\right)\\ &\overset{SUTVA}{=} \underbrace{E\left(Y_1 \mid X = x, D = 1\right)}_{\text{ìdentified}} - E\left(Y_1^0 \mid X = x, D = 1\right);\end{aligned}$$

$$\begin{aligned}E\left(Y_1^0 \mid X = x, D = 1\right) &\overset{CT}{=} E\left(Y_1^0 \mid X = x, D = 0\right) - E\left(Y_0^0 \mid X = x, D = 0\right) + E\left(Y_0^0 \mid X = x, D = 1\right)\\ &\overset{SUTVA}{=} \underbrace{E\left(Y_1 \mid X = x, D = 0\right) - E\left(Y_0 \mid X = x, D = 0\right)}_{\text{identified}} + E\left(Y_0^0 \mid X = x, D = 1\right);\end{aligned}$$

$$\begin{aligned}E\left(Y_0^0 \mid X = x, D = 1\right) &\overset{NEPT}{=} E\left(Y_0^1 \mid X = x, D = 1\right) =\\ &\overset{SUTVA}{=} \underbrace{E\left(Y_0 \mid X = x, D = 1\right)}_{\text{identified}}.\end{aligned}$$

Putting all pieces together, we get:

$$\begin{aligned}\theta_1(x) = &\left[E\left(Y_1 \mid X = x, D = 1\right) - E\left(Y_0 \mid X = x, D = 1\right)\right]\\ &-\left[E\left(Y_1 \mid X = x, D = 0\right) - E\left(Y_0 \mid X = x, D = 0\right)\right].\end{aligned}$$

Since $\theta_1(x)$ is a function of random variables for which realisations are observable, it is identified. Aggregating the conditional effects with respect to the appropriate distribution of *X* in the group of the treated in the post-treatment period leads to the desired average treatment effect on the treated.

The interpretation of the identification of the counterfactual nontreatment outcome is obvious: We use the pre-treatment outcome of the participants to infer the level of the nontreatment outcome and then infer the change of that potential outcome that would occur from period zero to period one from the change we actually observe for the nonparticipants.

Note that the assumptions imposed above rule out that the composition of the group of nontreated is affected by the treatment outcomes (this is mentioned, for example, by Angrist and Pischke, 2009, as one of the common pitfalls with DiD estimation in practice).

3.2.3 Identification of the average effects on the population in general and the nontreated

To be able to identify the average effect for the nontreated as well, $ATENT_t = E\left(Y_t^1 - Y_t^0 \mid D=0\right)$, and thus also to be able to identify the mean effect for the population, $ATE_t = E\left(Y_t^1 - Y_t^0\right) = ATENT_t \times P\left(D=0\right) + ATET_t \times P\left(D=1\right)$, it is required to express another counterfactual, namely $E\left(Y_1^1 \mid X = x, D=0\right)$, in terms of observables using DiD type of assumptions. In this case, the common trend assumption would have to involve the potential outcomes when treated and could be formalized in the following form:

$$\begin{aligned}&E\left(Y_1^1 \mid X = x, D=1\right) - E\left(Y_0^1 \mid X = x, D=1\right) = \\ &E\left(Y_1^1 \mid X = x, D=0\right) - E\left(Y_0^1 \mid X = x, D=0\right); \qquad \forall x \in \chi.\end{aligned}$$

This (technical) condition could of course be assumed (together with some further generalisations of the assumptions made in section 3.2.1). However, to use the same ideas as for the average treatment effect on the treated, we would need a subpopulation that is treated in period 0 and somehow become untreated later on. Correctly speaking, 'becoming untreated' means that the *effect* of the treatment vanishes in period one. Such a scenario is unlikely to be plausible in most economic applications. Thus, empirical papers using DiD almost always do neither attempt to identify the ATE nor to identify ATET and ATENT together.

3.2.4 The scale dependence of the identifying assumptions

As mentioned before, it has been observed by several authors, e.g. Meyer, Viscusi, and Durbin (1995), that the identifying assumptions in the difference-in-difference framework are scale dependent, i.e. if they hold for the level of *Y*, they may not hold for monotone transformations of *Y*. In other words, the way how we measure and transform the outcome variable is relevant for the plausibility of the identifying assumptions, even without postulating any parametric model for the relation of confounders and treatment to the outcomes. This is a feature that is not shared by other (nonparametric) identification strategies like instrumental variables or matching. Thus, we should call the DiD design a semiparametric identification strategy in contrast to the nonparametric identification strategies.

This distinction can easily be seen by considering the following example. Suppose that the potential nontreatment outcomes are log-normally distributed and that covariates play no role. Parameterize the log-normal distribution in one of the following ways: (i) $\ln Y_t^0 \mid X = x, D = d \sim N(0, 2d + 2t)$; (ii) $\ln Y_t^0 \mid X = x, D = d, T = t \sim N(d + t, 2)$; or (iii) $\ln Y_t^0 \mid X = x, D = d, T = t \sim N(d, d+1)$, $\forall d, t \in \{0,1\}$. In the first case, the log of the potential outcome has mean zero and is heteroscedastic. In the second case, it is homoscedastic, but its mean depends on group membership and time period. In the third example we shut down the trend and consider the stationary case. Consider two choices of scale of *Y (for example earnings)* that are popular for continuous variables: the level of the outcome variable (Y^0) as well as a log transformation ($\ln Y^0$). Next, we check whether the *CT* assumption holds in these settings:[16]

[16] Note that $\ln Y \sim N(\mu, \sigma^2) \quad \Rightarrow \quad Y \sim N(e^{\mu + 0.5\sigma^2}, (e^{\sigma^2} - 1)(e^{2\mu + \sigma^2}))$.

Case 1: $\ln Y_t^0 \mid D = d \sim N(0, 2d + 2t)$

$$\underbrace{E(\ln Y_1^0 \mid D=1)}_{0} - \underbrace{E(\ln Y_0^0 \mid D=1)}_{0} = 0,$$
$$\underbrace{E(\ln Y_1^0 \mid D=0)}_{0} + \underbrace{E(\ln Y_0^0 \mid D=0)}_{0} = 0;$$

$$\underbrace{E(Y_1^0 \mid D=1)}_{e^2} - \underbrace{E(Y_0^0 \mid D=1)}_{e^1} = e(e-1),$$
$$\underbrace{E(Y_1^0 \mid D=0)}_{e^1} + \underbrace{E(Y_0^0 \mid D=0)}_{e^0=1} = e-1.$$

For case I, the common trend assumption holds for the logs but not for the levels. Next, we consider case II:

Case 2: $\ln Y_t^0 \mid D = d \sim N(d + t, 2)$

$$\underbrace{E(\ln Y_1^0 \mid D=1)}_{2} - \underbrace{E(\ln Y_0^0 \mid D=1)}_{1} = 1,$$
$$\underbrace{E(\ln Y_1^0 \mid D=0)}_{1} + \underbrace{E(\ln Y_0^0 \mid D=0)}_{0} = 1;$$

$$\underbrace{E(Y_1^0 \mid D=1)}_{e^3} - \underbrace{E(Y_0^0 \mid D=1)}_{e^2} = e^2(e-1),$$
$$\underbrace{E(Y_1^0 \mid D=0)}_{e^2} + \underbrace{E(Y_0^0 \mid D=0)}_{e^1} = e(e-1).$$

Again, the common trend assumption holds for the log-specification, but not for the level-specification. Another feature of the functional form dependence is also apparent in the second example: if the conditional mean would be *t* times *d* instead of *t* + *d*, then the common trend assumption would be violated for the log as well as for the level specification.

Case 3: $\ln Y_t^0 \mid D = d \sim N(2d, 2d + 2)$.

This case is indeed trivial. As neither the mean nor the variance changes over time, the common trend assumption is automatically fulfilled for all transformations of *Y*.

This dependence of the validity of the identifying assumption on the scale of measurement of the outcome variable is a disadvantage of DiD, because the credibility of the 'common trend' or 'constant bias' assumptions becomes functional form dependent. Identification by functional form is less attractive as the researcher very seldom has access to credible information about the appropriate functional forms. There is, however, another way of viewing this problem of functional form (or scale of measurement) dependence: The appropriate functional form of the outcome variable should follow from the parameter the researcher is interested in. However, even when a sensible functional form can be derived from the parameters of interest, it remains a credibility issue. Why should the CT or CB assumptions be plausible for that particular choice of scaling, while they may be violated for other choices, which might be equally plausible a priori?

3.2.5 The changes-in-changes model by Athey and Imbens (2006)

The functional form dependence explored in the previous section is the starting point for the so-called 'changes-in-changes model' that has been proposed by Athey and Imbens (2006). The goal of their paper is to state a set of DiD-like assumptions that are not scale dependent.

The CiC ('**C**hanges-in-**C**hanges') model proposed by Athey and Imbens (2006) assigns the idea of the DiD model to the distribution of the potential outcomes. The idea is to compare the cumulative distribution functions (cdf) of the outcomes in the four groups. Then the difference of the cumulative distribution functions of the treated and the non-treated group in the pre-programme period is used to predict the hypothetical non-treatment cumulative distribution function of the treated group in the post-treatment period if they were not treated. Comparing this predicted cumulative distribution functions to the observed cumulative distribution functions of the treated in the post-treatment period gives the desired effect.

The key difference to the standard DiD approach is that the assumptions made as well as the information exploited for identification and estimation comes from the whole outcome distribution and not just from the first moments. Estimation is based on estimating cumulative distribution functions as well as their inverses for each group defined by treatment and time conditional on *X* and predicting the counterfactual outcomes based on those functionals.

Although the estimation problem is straightforward in principle (the authors provide $\sqrt{N}-$consistent and asymptotically normal estimators), considerable problems may appear in practice either if the outcome variables are not continuous or the types of individuals are too heterogeneous (based on the different relevant values of *x*). In the first case, the problem is that the inversion of a cumulative distribution function of a discrete random variable is not unique. Therefore, only bounds of the true effects are identified (they are given in the paper). In the second case, when control variables are present, estimation either has to be performed within cells defined by the discrete values of those variables, or appropriate smoothing techniques have to be used in the case of continuous variables (or discrete variables with many different possible values). Both issues are only of limited concern asymptotically. However, given the usual dimensions of the control variables and sample sizes in applications, the curse of dimensionality could be a major concern for applied researchers who may need to control for many variables to make the common trend assumption sufficiently credible. Perhaps for this reason the number of empirical applications of this approach is limited so far. The only published application known to the author (other than the one provided by Athey and Imbens in their seminal paper) is the one by Havnes and Mogstad (2010) who analyse the effects of child care in Norway. They are not only interested in mean impacts but also in the effects on the quantiles of the earnings distribution.

3.2.6 The role of covariates

As mentioned above, we need to control for precisely those exogenous variables that lead to differential trends and that are not influenced by the treatment. Including further control variables has positive and negative aspects, even when these additional variables do not lead to a violation of the DiD assumptions postulated above. On the positive side, it could help to detect effect heterogeneity that may be of substantive interest to the researcher (e.g. estimating the effects for men and women separately although men and women experience the same trends for their potential outcomes). On the negative side, every additional variable makes the common support assumption more difficult to fulfil.

So far we discussed the case of time constant covariates. In many studies, though, measurements in different time periods may be available. In particular, in a repeated cross-section setting, it is likely that the only available measurement of some covariates was taken at the same time as the measurement of the outcome variables. The particular concern here is the post-treatment period because time varying trend-confounding variables measured after the treatment are more likely to be influenced by it. Thus, the exogeneity condition might be violated. In this case, controlling for such time varying covariates leads to biased estimates. Generally, time varying covariates are no problem if they are exogenous. Indeed, if they are exogenous they may even be better suited to remove trend-confounding than covariates that do not vary over time (for example using age instead of birth year may be a superior choice to remove trend-confounding in some applications). If these variables are endogenous and if anticipation effects play no role, then using pre-treatment measurements whenever available may be the best empirical strategy.

Now, we change the perspective somewhat and look at the type of circumstances under which an unobservable variable that influences the potential outcomes can be safely ignored when estimating a DiD model. To see this, we consider the impact of the excluded (perhaps unobservable) time constant variable *U* in the simplest case without covariates. Obviously, we need to require CT to hold unconditionally, because *U* is unobserved. To see its role on the unconditional common trend assumption, we use iterated expectations:

$$
\begin{aligned}
CT:\ & E_{U|D=1}\left[E\left(Y_1^0 \mid D=1, U=u\right)-E\left(Y_0^0 \mid D=1, U=u\right)\right]= \\
&= E_{U|D=0}\left[E\left(Y_1^0 \mid D=0, U=u\right)-E\left(Y_0^0 \mid D=0, U=u\right)\right].
\end{aligned}
\tag{4}
$$

Clearly, if the common trend assumption holds conditional on *U* (if not, then there is no reason to consider using *U* as an additional control variable), and if the distribution of *U* does not depend on the treatment status, then *U* can be safely ignored in the estimation without violating CT.

Next, consider the case in which CT again holds conditional on *U*, but the distribution of *U* differs for different values of *D*. In this case, if *U* cannot be used as covariate, further assumptions are required. One such assumption is that the effect of the unobservable variable on the potential outcomes is constant over time (but may vary across treatment status). The following separable structure has such a property:

$$
E_{U|D=d} E\left(Y_t^0 \mid D=d, U=u\right)=E\left(Y_t^0 \mid D=d\right)+E_{U|D=d} f^d(U), \tag{5}
$$

where $f^d(U)$ denotes some arbitrary function of *U* that may depend on the selection into group *d* but not on time. By taking differences over time, the term $E_{U|D=d} f^d(U)$ disappears. Thus, we can allow for selection into treatment based on *unobservable* variables that also influence potential outcomes as long as their impact is constant over time. This feature is the reason why in a linear parametric setting with panel data difference-in-difference estimation and fixed-effects estimation is indeed very similar and sometimes identical (see for example the discussion in Angrist and Pischke, 2009).

3.2.7 The relation of the DiD assumptions to the matching assumptions

Returning to the case of time constant covariates, next we consider the relation of DiD to another closely related identification and estimation approach, namely matching. An assumption that identifies *ATET* is that the expectation of the respective potential outcome does not depend on the treatment status conditional on the covariates:

$$E\left(Y_1^0 \mid X = x, D = 1\right) = E\left(Y_1^0 \mid X = x, D = 0\right). \qquad (6)$$

This assumption is implied by the set of assumptions that characterises the matching approach to the identification of causal effects (these assumptions are labelled as conditional independence, selection on observables, and unconfoundedness assumptions). It is in fact related to the DiD assumption that presumes that the difference of the expectations of the potential outcomes over time does not depend on the treatment status. Despite their similarity, neither of the assumption nests the other: the DiD assumption allows for some selection on unobservables, which is ruled out in matching, while matching makes no assumptions about the pre-treatment periods, which is required in DiD. In applications based on matching estimation the conditional independence assumption is frequently strengthened by assuming that not only the mean, but the distribution of the potential outcomes conditional on covariates is independent of the treatment status. This stronger assumption has the virtue that it identifies not only the counterfactual expectations but the full counterfactual distribution. Furthermore, it makes the identification invariant to the chosen scaling of *Y*. Although the same approach can be chosen in a difference-in-difference framework, namely assuming that the difference of the potential outcomes over time is independent of the treatment, Appendix A.2 shows that there are no comparable gains for DiD estimation as there are for matching. Hence, the 'independence of the differences of the potential outcomes over time' assumption is not that attractive because compared to the common trend in means assumption it is more restrictive without identifying further interesting quantities.[17]

3.2.8 Panel data

So far in this exposition, it was not discussed whether the same individuals are observed in the pre- and post-treatment periods, because all identification results that are valid for repeated cross-sections also hold for panel data. So even though *individual* pre-treatment-post-treatment differences of outcome variables can be computed with panel data, the nature of the estimator is the same but its precision may change.

One consequence of basing the estimator on individual differences over time is that all influences of time constant confounding factors that are additively separable from the remaining part of the conditional expectations of the potential outcomes are removed by the DiD-type of differencing, as shown in the previous section. Therefore, it is not surprising that adding fixed individual effects instead of the treatment group dummy d in the regression formulation below (and all time constant covariates X), will lead to the same estimand (e.g. Angrist and Pischke, 2009).

Furthermore, from the point of view of identification, a substantial advantage of panel data is that matching estimation based on conditioning on pre-treatment outcomes is feasible as well. This is an important issue because it appears to be a natural requirement for a 'good' comparison group to have similar pre-treatment means of the outcome variables. This is not possible with repeated cross-sections since we do not observe pre-treatment outcomes from the same individuals but only from some group that is similar to the individuals obtaining the treatment in terms of other observable characteristics X.

[17] See, however, the alternative assumptions imposed by Athey and Imbens (2006) and Bonhomme and Sauder (2011) that identify distributional effects.

The corresponding matching-type assumption when lagged outcome variables are available can be expressed as follows:

$$E\left(Y_1^0 \mid Y_0 = y_0, X = x, D = 1\right) = E\left(Y_1^0 \mid Y_0 = y_0, X = x, D = 0\right). \tag{7}$$

Imbens and Wooldridge (2009) observe that the common trend assumption and this matching-type assumption impose different identifying restrictions on the data which are not nested and must be rationalized based on substantive knowledge about the selection process, i.e. only one of them can be true. Angrist and Krueger (1999) elaborate on this issue on the basis of regression models and come to the same conclusions. The advantage of the DiD method is that it allows for time constant confounding unobservables while requiring common trends, whereas matching does not require common trends but assumes that conditional on pre-treatment outcomes confounding unobservables are irrelevant. Of course, one may argue that conditioning on the past outcome variables already controls for the part of the unobservables that manifested itself in the lagged outcome variables.

One may try to combine the good features of both methods by including pre-treatment outcomes among the covariates in a DiD framework. This is however identical to matching: Taking the difference while keeping the pre-treatment part of that difference constant at the individual level in any comparison (i.e. the treated and matched control observations have the same pre-treatment level) is equivalent to just ignoring the difference in DiD and to focus on the post-treatment variables alone. Thus, such a procedure implicitly requires the matching assumptions.[18] In other words, assuming common trends conditional on the start of the trend (which means it has to be the same starting point for treated and controls) is practically

[18] Although the confounding individual effect has been removed by taking differences, conditioning on the pre-treatment levels is like conditioning on it again and thus may induce a correlation with *D*. In other words, if the DiD assumptions hold unconditionally on the pre-treatment outcome, they are likely to be violated conditional on pre-treatment outcomes.

identical to assuming no confounding (i.e. that the matching assumptions hold) conditional on past outcomes.

Thus, Imbens and Wooldridge's (2009, p. 70) conclusion about the usefulness of DiD in panel data compared to matching is negative: "As a practical matter, the DiD approach appears less attractive than the unconfoundedness-based approach in the context of panel data. It is difficult to see how making treated and control units comparable on lagged outcomes will make the causal interpretation of their difference less credible, as suggested by the DID assumptions." However, a recent paper by Chabé-Ferret (2010) gives several examples in which a difference-in-difference strategy leads to a consistent estimator while matching conditional on past outcomes may be biased. He also shows calibrations based on real data suggesting that this bias may not be small.

3.2.9 The regression formulation

Derivation of the linear specification

Most of the empirical applications employing the DiD identification strategy so far used the linear model. The linear regression formulation can be motivated by the following assumptions about the conditional expectations for the potential outcomes:

$$\begin{aligned}&E\left(Y_t^1 \mid X = x, D = d\right) = \alpha + t\delta^1 + d\gamma + x\beta + tx\lambda^1 + dx\pi;\\ &E\left(Y_t^0 \mid X = x, D = d\right) = \alpha + t\delta^0 + d\gamma + x\beta + tx\lambda^0 + dx\pi; \quad \forall d \in \{0,1\}, \forall x \in \chi.\end{aligned} \tag{8}$$

The specification is flexible to some degree as it includes several interaction terms between the control variables and group membership. However, it does not include interactions between time and treatment status as these interactions would violate the common trend assumption. This exclusion restriction gives rise to an interpretation of difference-in-difference estimation as conditional (on *X*, *D*, and *T*) IV estimation with the interaction term *D T* acting as an instrument (with perfect compliance). This interpretation again points to the essentially parametric nature of this approach as there cannot be independent variation of this instrument given its components *D* and *T*. This is only possible if a (semi-) parametric model will be specified in which the effects of *D₁* and *T* are separable.

Note that some of the coefficients do not vary across potential outcomes as a simple way to ensure that the treatment has no effect in period zero.[19]

The next step is to show that this specification indeed fulfils the common trend assumption:

$$
\begin{aligned}
&E\left(Y_1^0 \mid X=x, D=1\right)-E\left(Y_0^0 \mid X=x, D=1\right)= \\
&\quad =\alpha+\delta^0+\gamma+x\beta+x\lambda^0+x\pi-\alpha-\gamma-x\beta-x\pi=\delta^0+x\lambda^0;
\end{aligned}
$$

$$
\begin{aligned}
E\left(Y_1^0 \mid X=x, D=0\right)-E\left(Y_0^0 \mid X=x, D=0\right)&=\alpha+\delta^0+x\beta+\ x\lambda^0-\alpha-x\beta \\
&=\delta^0+x\lambda^0.
\end{aligned}
$$

Again, these derivations clarify that having differential trends for the different potential outcomes is no problem as long as the trends do not depend on the treatment status.

Starting from the specifications for the potential outcomes, the effects can be expressed in terms of their regression coefficients:

$$
\begin{aligned}
\theta_1(x)= \quad & E\left(Y_1 \mid X=x, D=1\right)-E\left(Y_0 \mid X=x, D=1\right) \\
& -\left[E\left(Y_1 \mid X=x, D=0\right)-E\left(Y_0 \mid X=x, D=0\right)\right] \\
& =\underbrace{(\alpha+\delta^1+\gamma+x\beta+x\lambda^1+x\pi)-(\alpha+\gamma+x\beta+x\pi)}_{\delta^1+x\lambda^1}-\underbrace{[(\alpha+\delta^0+x\beta+x\lambda^0)-(\alpha+x\beta)]}_{\delta^0+x\lambda^0} \\
& =\underbrace{(\delta^1-\delta^0)}_{\delta}+x\underbrace{(\lambda^1-\lambda^0)}_{\lambda}=\delta+x\lambda.
\end{aligned}
$$

[19] In a more general model, we would have $E\left[Y_t^{\tilde d} \mid X=x, D=d\right]=\alpha^{\tilde d}+t\delta^{\tilde d}+d\gamma^{\tilde d}+x\beta^{\tilde d}+tx\lambda^{\tilde d}+dx\pi^{\tilde d}$ and $E\left[Y_0^1-Y_0^0 \mid X=x, D=d\right]=(\alpha^1-\alpha^0)+d(\gamma^1-\gamma^0)+x[(\beta^1-\beta^0)+d(\pi^1-\pi^0)]\overset{!}{=}0.$

superscript 0 missing for lambda on third line.Therefore, the task of regression estimation is to obtain consistent estimates of δ and λ. To do so, we derive the regression model for the *observed* outcome variable.

The observation rule (SUTVA), $Y_t = dY_t^1 + (1-d)Y_t^0$, leads to the following specification for the observed outcome:

$$\begin{aligned}
E\left(Y_t \mid X = x, D = d\right) &= E\left[dY_t^1 + (1-d)Y_t^0 \mid X = x, D = d\right] \\
&= dE\left(Y_t^1 \mid X = x, D = d\right) + (1-d)E\left(Y_t^0 \mid X = x, D = d\right) \\
&= d(\alpha + t\delta^1 + \gamma + x\beta + tx\lambda^1 + x\pi) + (1-d)(\alpha + t\delta^0 + x\beta + tx\lambda^0) \\
&= \alpha + t\delta^0 + x\beta + tx\lambda^0 + d(\alpha + t\delta^1 + \gamma + x\beta + tx\lambda^1 + x\pi - \alpha - t\delta^0 - x\beta - tx\lambda^0) \\
&= \alpha + t\delta^0 + x\beta + tx\lambda^0 + d\left[\gamma + x\pi + t\underbrace{(\delta^1 - \delta^0)}_{\delta} + tx\underbrace{(\lambda^1 - \lambda^0)}_{\lambda}\right] \\
&= \alpha + t\delta^0 + d\gamma + x\beta + tx\lambda^0 + dx\pi + dt\delta + dtx\lambda; \qquad \forall d \in \{0,1\}, x \in \chi.
\end{aligned}$$

From these derivations, we see that a regression with group and time dummies (so-called main effects) plus the various interaction terms identify the causal effects. In such a regression, the coefficients of the interaction terms between time and treatment group capture the effects. It is rather common practice to assume that the coefficient λ is zero, implying that the interaction of group and time with the control variables disappears.

Advantages of the regression formulation

The advantage of the regression formulation of the DiD identification and estimation problem is the easiness of obtaining the final estimates and their standard errors (although even in this simple case there are some DiD specific inference problems that were discovered recently; they are mentioned below). Furthermore, we can easily extend the model to cover more periods and more treatments, including continuous treatments, and add additional covariates without much further computational effort.

Disadvantages of the regression formulation

There are also disadvantages of this regression-based approach to DiD. They concern (i) the effect heterogeneity that is allowed for, (ii) the way how control variables are included, and (iii) the possibility of arriving at estimates that are not plausible. It is important to note that those issues only appear if covariates are included. If covariates are not included, then estimation of the effects in DiD designs using the linear regression framework described above is fully nonparametric (in the sense that it does not impose any further assumptions than the ones needed for identification, which were discussed in the previous section).

Let us consider the potential issues (restrictions implied by the regression framework) in turn. First, for the issue of effect heterogeneity consider the simpler case in which λ equals 0. If there is indeed any effect heterogeneity, it means that the true coefficient δ is random instead of being a constant. Since the regression estimation essentially assumes a nonstochastic coefficient, the stochastic deviation that cannot be captured in the regression becomes part of the error term. This may not be harmful, if the heterogeneity is pure random noise uncorrelated with all variables included in the regression or if the model is fully saturated in the controls variables (see Angrist and Pischke, 2009). The regression coefficient still captures the average effect. However, if the heterogeneity is related to those variables and the model is not fully saturated in the covariates, then for example δ estimated by OLS is inconsistent (and asymptotically biased) for the ATET.

Second, including the control variables in a linear fashion implies the assumption of common trends conditional on the linear index, $X\beta$, which is more restrictive than assuming common trends conditional on *X*. Any deviation from the linear index is again absorbed in the regression error term and may invalidate the estimates.

Finally, if the outcomes have limited support, such as a binary variable, there is no guarantee that the predicted expected potential outcomes will respect this support condition.

The latter is one of the reasons why in practice linear models are rarely used in these cases and why nonlinear models, like logit or probit models for binary outcome variables, are usually preferred. However, these non-linear models come with their particular problems in a DiD setup, as will be explained below.

Nonlinear models with the standard common trend assumption

There are many types of outcome variables for which it is common practice to use nonlinear models instead of linear ones, because they provide a better approximation of the statistical properties of such random variables. Popular examples are probit, logit, tobit, count data models, and duration models. The general arguments in favour of such models do not, however, carry over to effects identified by the DiD assumptions. When applying nonlinear models in a DiD framework, researchers typically use a linear index structure together with a nonlinear link function (e.g. Hunt, 1995, Gruber and Poterba, 1994). The linear index structure is specified as if it would be a specification for the linear regression model. Then, the model is estimated and the estimated coefficients, or (average) marginal effects, are interpreted causally. However, while the linear regression can be derived from the DiD assumptions together with some restrictions on functional forms, such rationalisation is generally not possible for nonlinear models. To see this more clearly, let us consider a nonlinear regression model in the same fashion as we have analysed the corresponding linear model.

We start with a 'natural' nonlinear model with a linear index structure which is transformed by a link function, $G(\cdot)$, to yield the conditional expectation of the potential outcome. In the case of the probit model, this link function would, for example, be the cumulative distribution function of the standard normal distribution:

$$E\left(Y_t^1 \mid X=x, D=d\right)=G(\alpha+t\delta^1+d\gamma+x\beta+tx\lambda^1+dx\pi);$$
$$E\left(Y_t^0 \mid X=x, D=d\right)=G(\alpha+t\delta^0+d\gamma+x\beta+tx\lambda^0+dx\pi); \quad \forall d\in\{0,1\}, \forall x\in\chi. \tag{9}$$

This specification resembles the linear model with the exception of the addition of the link function. It fulfils the NEPT assumption. Using the observation rule and performing the same transformation as for the linear model (inside the $G(\cdot)$ -function), we obtain the following specification for the observable outcomes that can be used for the econometric estimation of the model parameters (again, using the notation as introduced for the linear model above):

$$\begin{aligned}
E\left(Y_t \mid X=x, D=d\right) &= E\left[dY_t^1+(1-d)Y_t^0 \mid X=x, D=d\right] \\
&= dE\left(Y_t^1 \mid X=x, D=d\right)+(1-d)E\left(Y_t^0 \mid X=x, D=d\right) \\
&= G\left[d(\alpha+t\delta^1+\gamma+x\beta+tx\lambda^1+x\pi)+(1-d)(\alpha+t\delta^0+x\beta+tx\lambda^0)\right] \\
&= \ldots \\
&= G\left(\alpha+t\delta^0+d\gamma+x\beta+tx\lambda^0+dx\pi+dt\delta+dtx\lambda\right); \forall d\in\{0,1\}, \forall x\in\chi.
\end{aligned}$$

(10)

This equation, sometimes specified with fewer interaction terms, is the basis for the empirical analysis in papers using (standard) nonlinear difference-in-differences.[20]

Of course, as for the linear model, we need to check whether such a specification indeed fulfils the common trend assumption:

$$\begin{aligned}
&E\left(Y_1^{\tilde d} \mid X=x, D=1\right) \qquad - E\left(Y_0^{\tilde d} \mid X=x, D=1\right)= \\
&\quad = G(\alpha+\delta^{\tilde d}+\gamma+x(\beta+\lambda^{\tilde d}+\pi))-G(\alpha+\gamma+x(\beta+\pi)); \\
&E\left(Y_1^{\tilde d} \mid X=x, D=0\right) \qquad - E\left(Y_0^{\tilde d} \mid X=x, D=0\right)= \\
&\quad = G(\alpha+\delta^{\tilde d}+x(\beta+\lambda^{\tilde d})) \qquad - G(\alpha+x\beta); \qquad \forall \tilde d\in\{0,1\}.
\end{aligned}$$

[20] For different ways to estimate 'causal' parameters from these models, see the papers by Ai and Norton (2003) and Puhani (2008).

It may or may not come as a surprise, but the intuitive specification does *not* fulfil the common trend assumption. The common trend assumption relies on differencing out specific terms of the unobservable potential outcome, which does not happen in this nonlinear specification. Indeed, the common trend assumption would only be respected if γ and π would be zero. However, these are exactly those coefficients that capture the group specific differences. Whereas the linear specification requires the group specific differences to be time constant, the nonlinear specification requires them to be absent. Of course, this property of this nonlinear specification removes the attractive feature that DiD allows for some selection on unobservable group and individual specific differences. Thus, we conclude that estimating a difference-in-difference model with the standard specification of a nonlinear model would usually lead to an inconsistent estimator if the standard common trend assumption is upheld.

In other words, if the standard DiD assumptions hold, this nonlinear model does not exploit them (it will usually violate them). Therefore, estimation based on this model does not identify the causal effect of *D* on *Y*. Let us generalize the above model by introducing group specific coefficients:

$$\begin{aligned}
&E\left(Y_t^1 \mid X = x, D = d\right) = G(\alpha^d + t\delta^{1,d} + x\beta^d + tx\lambda^{1,d});\\
&E\left(Y_t^0 \mid X = x, D = d\right) = G(\alpha^d + t\delta^{0,d} + x\beta^d + tx\lambda^{0,d}); \qquad \forall d \in \{0,1\}.
\end{aligned} \tag{11}$$

Note that the terms $d\gamma^d$ and $dx\pi^d$ are absorbed by the group specific constant and slope (α^d and $x\beta^d$). For the common trend assumption, we then obtain:

$$\begin{aligned}
&E\left(Y_1^{\tilde{d}} \mid X = x, D = 1\right) - E\left(Y_0^{\tilde{d}} \mid X = x, D = 1\right) =\\
&\qquad = G\left[\alpha^1 + \delta^{\tilde{d},1} + x(\beta^1 + \lambda^{\tilde{d},1})\right] - G(\alpha^1 + x\beta^1);\\
&E\left(Y_1^{\tilde{d}} \mid X = x, D = 0\right) - E\left(Y_0^{\tilde{d}} \mid X = x, D = 0\right) =\\
&\qquad = G\left[\alpha^0 + \delta^{\tilde{d},0} + x(\beta^0 + \lambda^{\tilde{d},0})\right] - G(\alpha^0 + x\beta^0); \quad \forall \tilde{d} \in \{0,1\}.
\end{aligned}$$

This model fulfils the common trend assumption under a set of restrictions on the coefficients (e.g., $\alpha^1 + \delta^{\tilde{d},1} = \alpha^0 + \delta^{\tilde{d},0}$, $\alpha^1 = \alpha^0$ $\beta^1 + \lambda^{\tilde{d},1} = \beta^0 + \lambda^{\tilde{d},0}$, $\beta^1 = \beta^0$, so that $\lambda^{\tilde{d},1} = \lambda^{\tilde{d},0} = \lambda^{\tilde{d}}$ and $\delta^{\tilde{d},1} = \delta^{\tilde{d},0} = \delta^{\tilde{d}}$). However, those restrictions imply, as before, that there are no group effects and thus this parameterization for the potential outcomes is not attractive either. To summarize, these 'standard' nonlinear parametric specifications of the potential outcomes and the derived observed outcomes are not attractive, because they fulfil the DiD assumptions only under additional constraints which are usually not attractive in typical applications.[21]

Since the simple way of using standard parametric models does not work in the nonlinear case when identification is achieved by the difference-in-difference assumptions, what are the alternatives? One alternative is to use a linear specification despite its problematic features for outcome variables with bounded support. A second alternative is to use nonlinear parametric approximations to predict the four components of the conditional-on-*X* effects, $E\left(Y_t \mid X = x, D = d\right)$, $t,d \in \{0,1\}$ in a parsimonious way, and then average those conditional-on-*X* effects according to the desired distribution of confounders to obtain estimates for the treated population. For example, with a binary outcome variable we may want to estimate a probit model in all three subsamples and obtain the following estimator for the average treatment effect on the treated:

$$\widehat{ATET_1} = \sum_{i=1}^{N} d_i t_i \left\{ \left[y_{1i} - \Phi(x_i \hat{\varphi}_1^0) \right] - \left[\Phi(x_i \hat{\varphi}_0^1) - \Phi(x_i \hat{\varphi}_0^0) \right] \right\}, \qquad (12)$$

[21] This problem has already been observed by Meyer (1995, p. 155), who explains it in the following way: "… if the mean of the outcome variable is very different in the treatment and control group, then [comparing expectations of means in the four groups] could not be an appropriate model both in levels and in means … This problem occurs because nonlinear transformations of the dependent variable imply different marginal effects on the dependent variable at different levels of the dependent variable. Thus, time could not have an effect of the exact same magnitude in both treatment and control groups in both a linear and a logarithmic specification." A similar observation has been made by Heckman (1996) in his discussion of a paper by Eissa (1996).

where $\hat{\varphi}_t^d$ denotes a vector of coefficients (including a constant) estimated using a probit model with dependent variable, Y_t, in the subsample defined by group *d* in period *t.* $\Phi(a)$ denotes the cumulative distribution function of the standard normal distribution evaluated at *a.* Note that one could predict the outcome for the treated in the post-treatment periods as well, but the average of such a prediction should be close to the mean of the outcome (it would be identical if a logit model estimated by maximum likelihood is used).

Estimating three (or four) probit models is of course similar to estimating a model in the overall sample (or the three or four subsamples) which is fully interacted with respect to *t* and *d.* In practice, one may wish to estimate a more parsimonious specification by omitting some of those interaction terms.

Although this approach may work well in practice (however, there seems to be no applied literature using such a specification in this way), a drawback of using these approximations is that we cannot recover the exact functional specifications of the mean *potential* outcomes, which makes it harder to understand the underlying restrictions that come from the required functional form assumptions.

Nonlinear models with a modified common trend assumption

In the previous section, it was shown that commonly used nonlinear models violate the common trend assumption. Here, we show that, indeed, for certain types of outcome variables the nature of these outcome variables makes it hard to believe from the outset that common trends may prevail at all. To see this problem using an example, assume that a binary variable for a particular group of nontreated in the post-treatment period has a mean of 0.9. Suppose further that the gap between the treated and nontreated groups prior to treatment is 0.2 in favour of the treated. In this case, adjusting for common trends would lead to an expected nontreatment outcome for the treated of 1.1, which would be outside the support of the outcome variable. Thus, in this example the common trend assumption must be violated.[22]

Following ideas similar to those of Blundell and Costa Dias (2009), we now explore the potential of a different way to specify identifying assumptions that resemble the key ideas of difference-in-difference estimation, but may appear to be more plausible than the common trend assumption for variables with bounded support and other cases in which the original DiD assumption appears implausible.[23] The following exposition is based on the concept of a latent dependent variable. Such variables figure very prominently in microeconometrics to link standard econometric linear models to discrete, censored or truncated dependent variables.

Concretely, assume that the conditional expectation of the observable outcome variable, *Y*, is related to the conditional expectation of a latent outcome variable, Y^*, in the following way:

$$E\left(Y_t^0 \mid X = x, D_1 = d\right) = H\left[E(Y_t^{0*} \mid X = x, D = d)\right]; \quad \forall d, t \in \{0,1\}, \forall x \in \chi. \qquad (13)$$

The function $H(\cdot)$ is assumed to be strictly monotonously increasing and invertible. The inverted function is denoted by $H(\cdot)^{-1}$. Therefore, we get:

$$E\left(Y_t^{0*} \mid X = x, D = d\right) = H^{-1}\left[E(Y_t^0 \mid X = x, D = d)\right]; \quad \forall d, t \in \{0,1\}, \forall x \in \chi.$$

The function $H(\cdot)$ plays the role of a typical link function that appears in probit, logit, tobit, and many other nonlinear models. For example, in the probit model $H(\cdot)$ is the

[22] I thank Patrick Puhani for a very interesting discussion on this subject.

[23] Blundell and Costa Dias (2009) use specifications based on various error terms that lead to the same results as the more direct approach followed here.

cumulative distribution function of the standard normal distribution. Now, let us assume common trends at the level of the expectations of the latent non-treatment outcome variables:

$$\begin{aligned} &E\left(Y_1^{0*} \mid X = x, D = 1\right) &&- E\left(Y_0^{0*} \mid X = x, D = 1\right) = \\ &E\left(Y_1^{0*} \mid X = x, D = 0\right) &&- E\left(Y_0^{0*} \mid X = x, D = 0\right) = && (CT*) \\ &E\left(Y_1^{0*} \mid X = x\right) &&- E\left(Y_0^{0*} \mid X = x\right), && \forall x \in \chi. \end{aligned}$$

Clearly, whether this assumption is plausible or not depends on the particular parameterisation of the model which critically involves the $H(\cdot)$ function. Even more important is that sometimes a substantive meaning can be given to the latent outcome variable, like an utility or an earnings potential, for example, which can then be used as the basis for judging the credibility of this assumption.

Using (NEPT) and the modified common trend assumption, we can show that the usual effects are identified:

$$\begin{aligned} \underbrace{E\left(Y_1^{0*} \mid X = x, D = 1\right)}_{H^{-1}\left[E\left(Y_1^0 \mid X = x, D=1\right)\right]} = \underbrace{E\left(Y_1^{0*} \mid X = x, D = 0\right)}_{H^{-1}\left[E\left(Y_1^0 \mid X = x, D=0\right)\right]} - \underbrace{E\left(Y_0^{0*} \mid X = x, D = 0\right)}_{H^{-1}\left[E\left(Y_0^0 \mid X = x, D=0\right)\right]} \\ + \underbrace{E\left(Y_0^{0*} \mid X = x, D = 1\right)}_{H^{-1}\left[E\left(Y_0^0 \mid X = x, D=1\right)\right]} = \\ = \underbrace{H^{-1}\left[E\left(Y_1^0 \mid X = x, D = 0\right)\right]}_{H^{-1}\left[E\left(Y_1 \mid X = x, D=0\right)\right]} - \underbrace{H^{-1}\left[E\left(Y_0^0 \mid X = x, D = 0\right)\right]}_{H^{-1}\left[E\left(Y_0 \mid X = x, D=0\right)\right]} \\ + \underbrace{H^{-1}\left[E\left(Y_0^0 \mid X = x, D = 1\right)\right]}_{H^{-1}\left[E\left(Y_0 \mid X = x, D=1\right)\right]} = \\ = H^{-1}\left[E\left(Y_1 \mid X = x, D = 0\right)\right] - H^{-1}\left[E\left(Y_0 \mid X = x, D = 0\right)\right] \\ + H^{-1}\left[E\left(Y_0 \mid X = x, D = 1\right)\right]. \end{aligned}$$

Therefore, we can express the missing counterfactual $E\left(Y_1^0 \mid X = x, D = 1\right)$ as

$$E\left(Y_1^0 \mid X = x, D = 1\right) = H\left[E\left(Y_1^{*0} \mid X = x, D = 1\right)\right] =$$
$$= H\left\{H^{-1}\left[E\left(Y_1 \mid X = x, D = 0\right)\right] -\right.$$
$$\left. - H^{-1}\left[E\left(Y_0 \mid X = x, D = 0\right)\right] + H^{-1}\left[E\left(Y_0 \mid X = x, D = 1\right)\right]\right\}.$$

Thus, *ATET₁* is identified.[24]

As an example consider a binary outcome variable analysed with a probit model in its linear index form with subsample specific coefficients. In this case, $H(\cdot)$ will be the cumulative distribution function of the standard normal distribution, $\Phi\left(\cdot\right)$, and $H^{-1}\left(\cdot\right)$ will be the respective inverse, $\Phi^{-1}\left(\cdot\right)$, which exists and is unique in this case. Such a probit model is defined as:

$$E\left(Y_t \mid X = x, D = d\right) = \Phi(x\varphi_t^d). \qquad (14)$$

This expressions leads to the final result:

$$E\left(Y_1^0 \mid X = x, D = 1\right) = \Phi\left\{\Phi^{-1}\left[\Phi(x\varphi_1^0)\right] - \Phi^{-1}\left[\Phi(x\varphi_0^0)\right] + \Phi^{-1}\left[\Phi(x\varphi_0^1)\right]\right\} =$$
$$= \Phi(x\varphi_1^0 - x\varphi_0^0 + x\varphi_0^1).$$

Letting $\hat{\varphi}_1^0$, $\hat{\varphi}_0^0$, and $\hat{\varphi}_0^1$ be consistent estimates of the unknown coefficients leads to the following expression for the missing mean potential outcome and thus the average treatment effect on the treated can be consistently estimated by:

$$\widehat{ATET_1} = \frac{1}{N^1}\sum_{i=1}^{N} d_i t_i \left[y_{1i} - \Phi(x\hat{\varphi}_1^0 - x\hat{\varphi}_0^0 + x\hat{\varphi}_0^1)\right], \quad N^1 = \sum_{i=1}^{N} d_i t_i.$$

[24] Note that when leaving out the covariates, this is exactly the expression derived by Blundell and Costa Dias (2009, p. 586).

It is obvious that this expression is different from the one given above that was also based on the probit model. Although both examples are based on a probit model, the former assumes common trends for the expected potential outcomes, whereas the latter assumes common trends for a nonlinear transformation of the expected outcomes. As already seen before, DiD is functional form dependent. Therefore, such transformations matter and lead to different results.

4. Some issues concerning estimation

4.1 Parametric models

So far, most empirical studies relying on a difference-in-difference approach are using parametric models. In this case, estimation is usually simple, at least as long as the model that is estimated is either a linear or a standard nonlinear model. Of course, if the model is based on a DiD assumption imposed on some complex transformation as discussed in the previous section on nonlinear models, it may be that even estimating a parametric model may be demanding and subject to substantial problems (like excessive computation time, convergence problems, or a nonunique objective function, etc.). However, there is a large literature in microeconometrics regarding these issues and their possible solutions. The fact that there is a type of DiD assumption involved does not create additional problems. Even in the parametric case, there may be DiD-specific problems with inference from the standard estimators that will, however, be briefly addressed in the section about inference below.

4.2 Semiparametric and nonparametric models

When no covariates are present, estimation can still be based on the regression formulation without being restrictive in any sense. Thus, standard linear estimators, like OLS, can be used and will have desirable properties. OLS based on a constant, group and time dummies, as well as their interaction only is identical to the typical DiD comparison of the four sample means. When covariates are included and the chosen specification is not rich enough to lead to a saturated model (usually because the number of observations is too small within each cell to allow for a complete set of interactions), the consistency of OLS is not guaranteed. It may depend on the validity of the functional form assumptions. Generally, parametric regression is restrictive because it depends on a linear index function to capture the influence of covariates on the outcomes and it restricts effect heterogeneity (see the in-depth discussion of the properties of linear estimation in such settings by Angrist and Pischke, 2009). Whether the resulting bias is a serious one or only a small one, is usually impossible to judge in an application without comparing the parametric results to results obtained by more flexible semi- or nonparametric models.

The issue of allowing covariates to enter the estimator in a flexible way can, however, be tackled similarly to the approaches taken by the semiparametric matching literature. To see this point, start by rewriting the average treatment effect on the treated in the following way:

$$\begin{aligned} ATET_1 &= E_{X|D=1}\theta_1(x) \\ &= E\left(Y_1 \mid D=1\right) - E_{X|D=1}E\left(Y_1 \mid X=x, D=0\right) \\ &\quad -\left[E_{X|D=1}E\left(Y_0 \mid X=x, D=1\right) - E_{X|D=1}E\left(Y_0 \mid X=x, D=0\right)\right]. \end{aligned} \tag{15}$$

The first observation is that in order to estimate $ATET_1$ consistently, a consistent estimate of the conditional on *X* effect, $\theta_1(x)$, is not necessary. Any estimate that will consistently estimate $ATET_1$ will have to reweight the outcomes observed in the three 'counterfactual' subpopulations (defined by *T* and *D*) such that these weights applied to the covariates will make the covariate distribution in the particular subpopulation identical to the one observed in the target population (*T = 1, D = 1).*

The second observation is that the structure of the estimand is very similar to the one for matching estimation for which a large literature exists (see for example the survey by Imbens, 2004). The difference is that instead of adjusting a covariate distribution in just the one subsample of nontreated, within a two-period DiD framework three covariate distributions have to be adjusted because there are three nontreated groups. In other words, the combination of three consistent matching estimators in the respective subsamples will lead to a consistent estimator of the population effects (if there are additional time periods available, the number of matching estimations increases by two for every additional period). It is also worth noting that the so-called propensity score properties first shown by Rosenbaum and Rubin (1983) for the case of matching can be applied here as well:

$$\begin{aligned}
&E_{X|D=1}E\left(Y_t \mid X = x, D = d\right) = \\
&= E_{p(X,t,d)|D=1}E\left[Y_t \mid p(X,t,d) = p(x,t,d), D = d\right]; \\
&\quad p(X,t,d) := P(TD = 1 \mid X = x, (T,D) \in \{(t,d),(1,1)\}].
\end{aligned} \qquad (16)$$

The common support assumption (COSU) ensures that these probabilities, called propensity scores in the matching literature, are positive. This general property is in principle already contained in the results derived by Rosenbaum and Rubin (1983). The proof for the DiD case is sketched in Appendix A.1.

So far studies using semiparametric or nonparametric covariate adjustments in DiD estimation are rare. The first paper to observe that DiD estimators can be based on matching seems to be Heckman, Ichimura, and Todd (1997). They base their estimators on local linear kernel estimators, as do Bergemann, Fitzenberger, and Speckesser (2009).

Acemoglu and Angrist (2001), whose empirical analysis is based on a very large sample and a fairly small number of discrete covariates, compute the respective means within the cells defined by the covariates and then weight the resulting differences with the probability of the particular cell in the population.

Eichler and Lechner (2002) use simple propensity score matching methods while Blundell, Meghir, Costa Dias, and van Reenen (2004) use a time specific and cross-section specific propensity score in their matching methods (see also Blundell and Costa Dias, 2009). Ravallion, Galasso, Lazo, and Philipp (2005) use propensity score matching as well to purge their estimates of differences coming from different distributions of trend confounders in the various subsamples.

Abadie (2005) proposes estimators based on weighting the outcomes by the propensity scores combined with linear projections. He also provides distribution theory for the case of using a nonparametric estimator for the propensity score. Finally, in an empirical application Bühler, Helm, and Lechner (2011) also use weighting on the propensity score. They propose to estimate different scores for each of the three comparisons involved. It is worth pointing out that all matching estimators considered in the literature can be used in this DiD setting as well. The paper by Huber, Lechner, and Wunsch (2010) give a comprehensive account of such estimators and their relative performance.

5. Some DID-specific issues about inference

Recently, there has been a renewed discussion about how to conduct inference in a (usually parametric) DiD setting. These papers are concerned with potential correlations of uncertainty over time as well as within groups (in particular with panel data), for example exhibited by group-time specific error terms. Note that if we have period/group-specific randomness (e.g. group-time specific individual random effects),[25] then with a finite number of periods and groups, no consistent estimator can exist, because within group averaging

[25] Of course, (additive separable) treatment group specific uncertainty that is constant over time, as well as time specific uncertainty that is constant over groups, is differenced out by taking the differences of the four group means.

cannot eliminate such variability. The only way to reduce this type of uncertainty is to have more periods and more (nontreatment) groups. The issue of correlations over time of the units in the different groups (like states before and after a policy change) as well as their cross-sectional correlation within groups has been analysed within a regression framework by Bertrand, Duflo, and Mullainathan (2004), Donald and Lang (2007), Hansen (2007a, 2007b) allowing for multiple periods and groups. These papers conclude that inference may be substantially affected by such cluster effects (see also the informative discussion about these issues within a regression-type framework in Imbens and Wooldridge, 2009, p. 69). Conley and Taber (2011) approach this problem somewhat differently. They consider the case of a small number of treated groups and an expanding number of control groups and develop asymptotic distribution theory for test statistics under this scenario. Although in this case DiD estimators for the effect of *D* may be inconsistent (but unbiased), they develop asymptotic standard tests that can be used for inference in this setting.

Despite, or because of, the recent findings on problems with standard inference in DiD estimation, how to best conduct inference very much depends on how uncertainty is specified. Generally, how to do inference remains an open issue in many DiD settings.

6. Further issues

6.1 Very small number of treated and control units / interventions at the aggregate level

There may be a problem for identification, estimation and inference (see previous section) if, for example, one region is subject to a treatment in period *t* (i.e. all units are subject to the new regime), whereas a couple of other regions are not subject to it. First, note that this may create some problem for identification if the other regions have different time trends of the potential outcomes compared to the treatment region. Abadie, Diamond, and Hainmüller (2007) suggest treating this problem essentially as a matching problem. They propose using a weighted average of all the control regions with weights chosen such that the weighted mean of the covariates in the control regions is equal, or at least as similar as possible, to the observed values in the treatment region after the treatment. Furthermore, the authors also argue that even in such aggregate studies, one should account for uncertainty (one may argue that one observes the population and, thus, there is no need to take sampling uncertainty into account) and provide some inference procedures that (i) do not rely on asymptotic theory and are valid in small samples, and (ii) take account of other uncertainty than sampling uncertainty.[26]

6.2 Additional periods

For most of the paper the simplest case of just two periods has been considered. This is sufficient to achieve identification of the usual quantities of interest. Now, we consider the value added by having more (post- and / or pre-treatment) periods available. It is assumed that in all pre-treatment periods the (future) treatment group is identified and that for all pre- and post-treatment periods the same covariates (that are not influenced by the future or past treatments) are observed.

Before expanding on the advantages of having more post- and pre-treatment periods, note that they are only useful if the common trend assumptions holds for all pre- and post-treatment comparisons. This assumption may become particularly implausible, or particular plausible, if the periods are further away from each other, depending on the particular scenario. Although derived in a fairly restrictive framework, the paper by Chabé-Ferret (2010)

[26] The already mentioned paper by Donald and Lang (2007) is also particularly interested in the case of a small number of treated and controls, while the study by Conley and Taber (2011) mentioned above provides an alternative analysis that is geared to the intermediate case of a larger number of nontreated and a small number of treated.

points to cases when, for example, an estimator using pre-post-treatment pairs with equal time distance to the treatment is consistent while using just the most recent pre-treatment period leads to an inconsistent estimator. Unfortunately, he also shows examples in which the opposite is true. Apparently, more research is required on this important topic.

6.2.1 Dynamics of the effects

Suppose we have additional post-treatment knowledge. If the common trend assumptions are valid for all comparisons with the pre-treatment period, then the additional information allows discovering effect dynamics and testing for effects being stable over time. This type of dynamics may be important information from a policy perspective.

If, however, it is known that the true effects are constant over time, performing DiD estimations based on using different single post-treatment periods allows to test the plausibility of the identifying assumption, because if the true effects are the same over time and if the identifying assumptions hold for all post-treatment periods, then the estimates from these pair-wise comparisons should be the same as well.

6.2.2 Checking the credibility of the identifying assumptions

The last remarks in the previous section also apply to the case of just one post-treatment period and many pre-treatment periods. Again, if (and only if) the identification condition is valid for all those periods, then the choice of the pre-treatment period used should not systematically change the estimates.

So-called 'placebo experiments' are also possible to test overidentifying assumptions and make the common trend assumption more plausible. Suppose for example that we have several pre-treatment periods. In this case, we could pretend that actually the treatment happened earlier and then measure the outcome after the pretended treatment but before the treatment actually happened. If we find an effect of this artificial treatment it could have either of the following two reasons: (i) The treatment is anticipated and therefore has an effect even before it starts. This (at least) raises some concerns about when to measure covariates which are not supposed to be influenced by the treatment. (ii) If we can rule out anticipation, this effect of the placebo treatment becomes a specification test for the common trend assumption, because any estimated nonzero effect would have to be interpreted as selection bias and thus would cast serious doubts on the validity of the identifying assumptions.[27]

6.2.3 Efficiency

Apparently, since more periods may generate overidentifying assumptions, such assumptions can be used to obtain more efficient estimators. If the identifying assumptions hold, it must be more efficient to use this additional information. Suppose we have Υ additional pre-treatment periods indexed by τ with valid common trend assumptions. In this case, we can rewrite the estimand in the following way:

$$
\begin{aligned}
&ATET_1 = Diff(1) - \sum_{\tau=-\Upsilon}^{0} w_\tau Diff(\tau); \qquad w_\tau > 0, \sum_{\tau=-\Upsilon}^{0} w_\tau = 1; \\
&Diff(t) = E_{X|D=1} E\left(Y_t \mid X = x, D = 1\right) - E_{X|D=1} E\left(Y_t \mid X = x, D = 0\right).
\end{aligned}
$$

The question now is how to choose the $(\Upsilon - 1)$ free weights w_τ. An obvious way would be to require that they minimize the variance of the overall estimator. The variance of the estimators for $Diff(t)$ can be derived using the usual methods. Intuitively, the smaller the variance of Y_t and the larger the sample size, the more precise this period estimator will be. Furthermore, the more similar the distribution of X in the subsamples in period τ is compared to the target distribution of X in $T=1$, $D=d$, the more precise we expect the estimator to be. Furthermore, this requires an estimator of the covariances of $Diff(t)$ and $Diff(t')$, which

[27] See also Angrist and Pischke (2009) as well as Autor (2003) of how to exploit the timing information in a regression DiD framework.

may be difficult to derive for those semiparametric estimators for which the bootstrap is not a valid variance estimator.[28]

7. Conclusions

The difference-in-difference design for empirical analysis of causal effects has a long history in and outside econometrics. Nowadays, it is certainly one of the most heavily used empirical research designs to estimate the effects of policy changes or interventions in empirical microeconomics. It has the advantage that the basic idea is very intuitive and thus easy to understand for an audience with limited econometric education. Compared for example to matching methods it has the further advantage that there is no need to control for all confounding variables. This means that it can accommodate a certain degree of selectivity based on unobservables correlated with treatment and outcome variables. Its key identifying assumption is the common trend assumption that must hold either unconditionally or conditionally on some observables (that are not influenced by the treatment). If the latter is the case, standard DiD estimation can be combined fruitfully with matching estimation techniques to accommodate such covariates in a flexible way. For that latter case, this paper propose some new propensity score based matching techniques that extend the proposals made by Abadie (2005) and Heckman, Ichimura, Smith, and Todd (1998) to a setting more similar to 'standard' matching estimation.

Unfortunately, like any empirical research design it has also severe disadvantages. First of all, the common trend assumption is functional form, or scale-of-measurement, dependent. For example, if there are common trends in the logs of an outcome variable, then

[28] Note that even in the case of repeated cross-sections with independent observations the estimators of *Diff(t)* may not be independent for different periods, because all periods use the empirical distribution of *X*, or of some function of *X* like the propensity score, to estimate *Diff(t)*. Thus, the two estimates could be correlated.

in all other than some exceptional and uninteresting cases the common trend assumption will not be valid for the untransformed levels of the outcome variable. This makes it much harder to justify the common trend assumption from substantive knowledge about selection and outcomes than for empirical research designs that lead to nonparametric identification, like matching or instrumental variables. A related problem is that the common trend assumption usually cannot be true for variables with bounded support. A potential remedy in this case may be to impose the common trend assumption in some latent model. However, justifying a common trend assumption for some latent objects is most likely to be even more difficult to justify in a credible way. It remains to be explored whether the more complex changes-in-changes model by Athey and Imbens (2006) will mediate these problems to some extent. Finally, depending on the structure of uncertainty, inference may be tricky to impossible in the two-periods-two-groups case. Having many (similar) periods and many (similar) groups (of nontreated) seems to be important, as it allows (i) more precise estimation; (ii) testing for the common trend assumption; (iii) and more reliable inference.

8. References

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

Abadie, A., A. Diamond, and J. Hainmüller (2007): "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program", NBER technical working paper 335.

Acemoglu, D., and J. D. Angrist (2001): "Consequences of Employment Protection? The Case of the Americans with Disabilities Act", *Journal of Political Economy*, 109, 915-957.

Ai, D., and E.C. Norton (2003): "Interaction Terms in Logit and Probit Models", *Economics Letters*, 80, 123-129.

Angrist, J. D., and A. B. Krueger (1999): "Empirical Strategies in Labor Economics", in O. Ashenfelter und D. Card (eds.), *Handbook of Labor Economics*, Vol. III A, Ch. 23, 1277-1366.

Angrist, J. D., and J.-S. Pischke (2009), *Mostly Harmless Econometrics*, New York: Princeton University Press.

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

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

Athey, S., and G. W. Imbens (2006): "Identification and Inference in Nonlinear Difference-in-Difference Models", *Econometrica*, 74, 431-497.

Autor, D. H. (2003): "Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing," *Journal of Labor Economics*, 21, 1-42.

Bergemann, A., B. Fitzenberger, and S. Speckesser (2009): "Evaluating the Dynamic Employment Effects of Training Programs in East Germany Using Conditional Difference-In-Differences", *Journal of Applied Econometrics*, 24, 797–823.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How much should we trust differences-in-differences estimates", *Quarterly Journal of Economics*, 249-275.

Besley, T., and R. Burgess (2004): "Can Labor Regulation Hinder Economic Performance? Evidence From India," *Quarterly Journal of Economics*, 91-134.

Blundell, R., A. Duncan, and C. Meghir (1998): "Estimating Labor Supply Responses Using Tax Reforms", *Econometrica*, 66, 827-861.

Blundell, R., and M. Costa Dias (2009): "Alternative Approaches to Evaluation in Empirical Microeconomics", *Journal of Human Resources*, 44, 565-640.

Blundell, R., C. Meghir, M. Costa Dias, and J. van Reenen (2004): "Evaluating the Employment Impact of a Mandatory Job Search Program", *Journal of the European Economic Association*, 2, 569-606.

Bonhomme, S., and U. Sauder (2011): “Recovering Distributions in Difference-in-Differences Models: A Comparison of Selective and Comprehensive Schooling”, *The Review of Economics and Statistics*, 93, 479-494.

Bühler, S., M. Helm, and M. Lechner (2011): "Trade Liberalization and Growth: Plant-Level Evidence from Switzerland", Discussion paper 2011-33, Department of Economics, University of St. Gallen.

Card, D. (1990): "The Impact of the Mariel Boatlift on the Miami Labor Market", *Industrial and Labor Relations Review*, 43/2, 245-257.

Card, D., and A. B. Krueger (1994): "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *The American Economic Review*, 84, 772-793.

Chabé-Ferret, S. (2010): "To Control or Not to Control? Bias of Simple Matching vs Difference-In-Difference Matching in a Dynamic Framework," *mimeo*.

Conley, T., and C. Taber (2011): "Inference with “Difference In Differences” with a Small Number of Policy Changes," *Review of Economics and Statistics*, 93, 113-125.

Cook, P. J., and G. Tauchen (1982): "The Effect of Liquor Taxes on Heavy Drinking", *Bell Journal of Economics*, 13, 379-390

Cook, T. D., and D. T. Campbell (1979), *Quasi-Experimentation*, Boston: Houghton Mifflin.

Donald, S. G., and K. Lang (2007): “Inference with Difference-in-Differences and Other Panel Data,” *Review of Economics and Statistics*, 89, 221–233.

Duflo, E. (2001): “Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment”, *American Economic Review*, 91, 795-813.

Eichler, M., and M. Lechner (2002): "An evaluation of public employment programmes in the East German State of Sachsen-Anhalt", *Labour Economics: An International Journal*, 9, 143–186.

Eissa, N. (1996): "Labor Supply and the Economic Recovery Act of 1981", in M. Feldstein and J. Poterba (eds), *Empirical Foundations of Household Taxation*, 5-38.

Gruber, J., and J. Poterba (1994): "Tax Incentives and the Decision to Purchase Health Insurance: Evidence from the Self- Employed," *The Quarterly Journal of Economics*, 109, 701-733.

Hansen, C. B. (2007a): “Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When T Is Large,” *Journal of Econometrics*, 141, 597–620.

Hansen, C. B. (2007b): “Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects,” *Journal of Econometrics*, 140, 670–94.

Havnes, T., and M. Mogstad (2010): "Is Universal Child Care Leveling the Playing Field? Evidence from a Nonlinear Difference-in-Difference Approach", IZA discussion paper 4478.

Heckman, J. J. (1996): "Comment on Eissa: Labor Supply and the Economic Recovery Act of 1981", in M. Feldstein and J. Poterba (eds.), *Empirical Foundations of Household Taxation*, 5-38.

Heckman, J. J., and R. Robb (1986): "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes", in H. Wainer (ed.), *Drawing Inferences from Self-Selected Samples*, 63-113.

Heckman, J. J., and V. J. Hotz (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training", *Journal of the American Statistical Association*, 84, 862-880.

Heckman, J. J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data", *Econometrica*, 66, 1017-1098.

Heckman, J. J., R. LaLonde, and J. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs", in: O. Ashenfelter and D. Card (eds.), *Handbook of Labour Economics*, Vol. 3, 1865-2097, Amsterdam: North-Holland.

Huber, M., M. Lechner, and C. Wunsch (2010): “How to control for many covariates? Reliable estimators based on the propensity score”, Discussion paper 2010-30, Department of Economics, University of St. Gallen.

Hunt, J. (1995): "The Effect of Unemployment Compensation on Unemployment Duration in Germany," *Journal of Labor Economics*, 13, 88-120.

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

Lai, A. (2011): “London cholera and the blind-spot of an epidemiology theory”, *Significance*, June 2011, 82-85.

Lechner, M. (2008): "A Note on Endogenous Control Variables in Evaluation Studies," *Statistics and Probability Letters*, 78, 190-195.

Lechner, M. (2008b): "A note on the common support problem in applied evaluation studies,", *Annales d'Économie et de Statistique*, 91-92, 217-234.

Lester, R. A. (1946): "Shortcomings of marginal analysis for the wage-employment problems", *American Economic Review*, 36, 63-82.

Manski, C. F. (2011): „Identification of Treatment Response with Social Interactions“, Department of Economics and Institute for Policy Research, Northwestern University, mimeo.

Meyer, B. D. (1995): "Natural and Quasi-Experiments in Economics", *Journal of Business & Economic Statistics*, 13, 151-161.

Meyer, B. D., W. K. Viscusi, and D. L. Durbin (19995): "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment", *American Economic Review*, 85/3, 322-340.

Miguel, E., and M. Kremer (2004): „Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities“, *Econometrica*, 72, 159–217.

Obenauer, M., and B. von der Nienburg (1915): "Effect of Minimum-Wage Determinations in Oregon". *Bulletin of the U.S. Bureau of Labor Statistics*, 176, Washington, D.C.: U.S. Government Printing Office.

Puhani, P. (2008): "The Treatment Effect, the Cross Difference, and the Interaction Term in Nonlinear 'Difference-in-Difference" Models", IZA Discussion paper 3478, revised Nov. 2008.

Ravallion, M., E. Galasso, T. Lazo, and E. Philipp (2005): “What Can Ex-Participants Reveal about a Program’s Impact”, *Journal of Human Ressources*, 40, 208-230.

Rose, A. M. (1952): "Needed Research on the Mediation of Labour Disputes", *Personal Psychology*, 5, 187-200.

Rosenbaum, P. (2001): "Stability in the Absence of Treatment", *Journal of the American Statistical Association*, 96, 210-219.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-50.

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 Group on the Basis of a Covariate", *Journal of Educational Statistics*, 2, 1-26.

Shadish, W. R., T. D. Cook, and D. T. Campbell, (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston: Houghton-Mifflin.

Simon, J. L. (1966): "The Price Elasticity of Liquor in the U.S. and a Simple Method of Determination", *Econometrica*, 34, 193-205.

Snow J. (1854): “The cholera near Golden Square, and at Deptford,” *Medical Times and Gazette*, 9, 321-322.

Snow, J. (1855), *On the Mode of Communication of Cholera*, 2nd ed., London: John Churchill.

Waldfogel, J. (1998): "The Family Gap for Young Women in the United States and Britain: Can Maternity Leave Make a Difference", *Journal of Labor Economics*, 16, 505-545.

Yelowitz, A. S. (1995): "The Medicaid Notch, Labor Supply, and Welfare Participation: Evidence from Eligibility Expansions", *Quarterly Journal of Economics*, 110, 909-939.

Technical Appendix

A.1 Propensity score property

The propensity score properties discussed in the main part of the text states that it does not matter whether reweighting the outcome variables in a specific subsample is with respect to the distribution of *X* or with respect to the distribution of the propensity score. Of course, this is not novel as it is already implied by the results in Rosenbaum and Rubin (1983). Therefore, we will only sketch the proof in a general way.

Denote the two subpopulations of interest as those characterized by the value of the binary random variable *W*. We want to reweight the outcomes in subpopulation *W=0* according to the distribution of *X* in subpopulation characterized by *W=1,* i.e. the target is to get $E_{X|Z=1}E(Y \mid X = x, Z = 0)$. The claim is that this can also be done by the propensity score, defined as $p(x) = P(W = 1 \mid X)$. Thus, the following property has to be shown to be valid:

$$E_{X|Z=1}E(Y \mid X = x, Z = 0) \overset{!}{=} E_{p(X)|Z=1}E(Y \mid p(X) = p(x), Z = 0).$$

A sketch of the proof is as follows:

$$\begin{aligned} E_{p(X)|Z=1}E(Y \mid p(X) = p(x), Z = 0) &= E_{p(X)|Z=1}E_{X|p(X),Z=0}E(Y \mid X = x, p(X) = p(x), Z = 0) \\ &= E_{p(X)|Z=1}E_{X|p(X),Z=0}E(Y \mid X = x, Z = 0) \\ &= E_{p(X)|Z=1}E_{X|p(X),Z=1}E(Y \mid X = x, Z = 0) \\ &= E_{X|Z=1}\, E(Y \mid X = x, Z = 0). \end{aligned}$$

The first line of the above expression uses iterated expectations, the second line employs the fact that *X* is finer than *p(x),* so that $E(Y \mid X = x, p(X) = p(x), Z = 0)$ $= ?$. The third line is based on the balancing score property derived in Rosenbaum and Rubin (1983) stating that *X* and *Z* are independent conditional on *P(Z=1|X),* and the last line exploits Bayes'

Law and again the fact that *X* is finer than *p(X),* $\Big(f_{X|p(X),Z=1}(x)\, f_{p(X),Z=1}(x) =$ 、 $f_{X,p(X)|Z=1}(x, p(x)) = f_{X|Z=1}(x) \Big)$.

A.2 Independence of differences of potential outcomes over time and treatment

An alternative to the common trend assumptions in means is to assume that the differences of the potential outcomes conditional on covariates are independent of *D*.

$$F\left(Y_1^0 - Y_0^0 \mid X = x, D = 0\right) = F\left(Y_1^0 - Y_0^0 \mid X = x, D = 1\right);$$

In the matching literature it would be more common to use the following notation:

$$\left(Y_1^0 - Y_0^0\right) \coprod D \mid X = x.$$

$A \coprod B \mid C = c$ means that *A* is independent of *B* conditional on *C* being equal to *c*. Next, we prove identification in a similar way as before:

$$\begin{aligned}
&F\left(Y_1^0 - Y_0^0 \mid X = x, D = 1\right) = F\left(Y_1^0 - Y_0^0 \mid X = x, D = 0\right) = F\left(Y_1 - Y_0 \mid X = x, D = 0\right) \\
&if \quad Y_0^0 = Y_0^1 \cdots \Rightarrow \\
&F\left(Y_1^0 - Y_0 \mid X = x, D = 1\right) = F\left(Y_1 - Y_0 \mid X = x, D = 0\right).
\end{aligned}$$

It appears that no further simplification of this expression is possible. Thus contrary to the matching assumptions, DiD does not identify the counterfactual distribution. It is probably easiest to see this result for the variance. The above assumption implies that changes in the variances of the differences over time do not depend on the treatment status:

$$\begin{aligned}
Var\left(Y_1^0 - Y_0^0 \mid X = x, D = 1\right) &= Var\left(Y_1^0 - Y_0^0 \mid X = x, D = 0\right) \\
&= Var\left(Y_1 - Y_0 \mid X = x, D = 0\right).
\end{aligned}$$

From the standard definition of the variance of a difference, we get the following expression:

$$\begin{aligned}Var\left(Y_1^0-Y_0^0 \mid X=x,D=1\right)=&Var\left(Y_1^0 \mid X=x,D=1\right)\\&+Var\left(Y_0^0 \mid X=x,D=1\right)\\&-2CoVar\left(Y_1^0,Y_0^0 \mid X=x,D=1\right).\end{aligned}$$

Putting those two equations together, we obtain the final expression for the variance of the counterfactual outcome:

$$\begin{aligned}&Var\left(Y_1^0 \mid X=x,D=1\right)=\\&=Var\left(Y_1-Y_0 \mid X=x,D=0\right)-Var\left(Y_0^0 \mid X=x,D=1\right)+2\mathrm{Co}var\left(Y_1^0,Y_0^0 \mid X=x,D=1\right)\\&=\underbrace{Var\left(Y_1-Y_0 \mid X=x,D=0\right)}_{\text{identified with panel data if } Y_0^0=Y_0^1}-\underbrace{Var\left(Y_0 \mid X=x,D=1\right)}_{\text{identified if } Y_0^0=Y_0^1}+2\underbrace{Covar\left(Y_1^0,Y_0 \mid X=x,D=1\right)}_{\text{not identified}}.\end{aligned}$$

Clearly, the covariance term is not identified and thus effects on the variance are not identified. This holds whether panel data are available or not.

However, due to the linearity of the expectations operator, mean effects are identified because this common trend assumption implies common trends in means as well (if the means exist), which is enough to identify mean average effects:

$$\begin{aligned}&F\left(Y_1^0-Y_0^0 \mid X=x,D=1\right)=F\left(Y_1^0-Y_0^0 \mid X=x,D=0\right)\\&\Rightarrow E\left(Y_1^0-Y_0^0 \mid X=x,D=1\right)=E\left(Y_1^0-Y_0^0 \mid X=x,D=0\right).\end{aligned}$$