Universität St.Gallen

Treatment effects and panel data

Michael Lechner

June 2013 Discussion Paper no. 2013-14

School of Economics and Political Science,
Department of Economics

University of St. Gallen

Editor: Martina Flockerzi
University of St. Gallen
School of Economics and Political Science
Department of Economics
Bodanstrasse 8
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35
Email seps@unisg.ch

Publisher: School of Economics and Political Science
Department of Economics
University of St. Gallen
Bodanstrasse 8
CH-9000 St. Gallen
Phone +41 71 224 23 25

Electronic Publication: Fax +41 71 224 31 35
http://www.seps.unisg.ch

Treatment effects and panel data

Michael Lechner

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 28 14
Fax +41 71 224 23 02
Email michael.lechner@unisg.ch
Website www.sew.unisg.ch

Abstract

It is a major achievement of the econometric treatment effect literature to clarify under which conditions causal effects are non-parametrically identified. The first part of this chapter focuses on the static treatment model. In this part, I show how panel data can be used to improve the credibility of matching and instrumental variable estimators. In practice, these gains come mainly from the availability of outcome variables measured prior to treatment. Such outcome variables also foster the use of alternative identification strategies, in particular so-called difference-in-difference estimation. In addition to improving the credibility of static causal models, panel data may allow credibly estimating dynamic causal models, which is the main theme of the second part of this chapter.

Keywords

Matching, instrumental variables, local average treatment effects, difference-in-difference estimation, dynamic treatment effects.

JEL Classification

C22, C23, C32, C33.

1. Introduction

In the last three decades two rapidly developing fields had an immense effect on how microeconometric empirical studies are conducted in our days. On the one hand, the literature on panel econometrics clarified how the increasing availability of panel data sets could improve the estimation of econometric models by exploiting the fact that repeated observations from a unit of the population are available. This led to more precise and more robust estimation strategies. Many of these methods made it into our standard econometric textbooks and became part of the standard econometric curriculum. This handbook, as well as the recently published 3rd edition of the 'Econometrics of Panel Data' (Mátyás and Sevestre, 2008), give a good account of the latest (as well as less new) developments in this field.

On the other hand, the so-called treatment effects literature exploded over the last two decades as well. It is a major achievement of the econometric treatment effects literature to clarify under which conditions causal effects are non-parametrically identified. This also led to a much better understanding of how to choose appropriate research designs and of 'what we are really estimating'. This is particularly so for the case when effects are heterogeneous, which is the prominent case in that literature. Furthermore, this literature, which is not yet as mature as the panel econometrics one, puts emphasis on identifying causal effects under as weak as possible (and plausible) conditions, which limits the role of tightly specified statistical parametric models. To the contrary, non- and semi-parametric methods are emphasized. Angrist and Pischke (2010) give a good account of these ideas and show how they influence the way microeconometric studies are done, while Heckman, LaLonde, and Smith (1999), and Imbens and Wooldridge (2009) provide rich surveys of the econometric methods. Angrist and Pischke (2009) give a (graduate) text book treatment of this topic,[1]

[1] This book also has a chapter on panel data but for the special case of difference-in-difference estimation, which so far provided the main formal link between treatment effects and panel data.

which also received top journal space in the ‘Forum on Estimation of Treatment Effects’ in the Journal of Economic Literature (e.g. Deaton, 2010, Heckman, 2010, Imbens, 2010) and the Symposium on ‘Con out of Econometrics’ by the Journal of Economic Perspectives (2010).

The first part of this chapter shows how panel data can be used to improve the credibility of methods usually used in the (static) treatment effects literature. These improvements come mainly from the availability of outcome variables measured prior to treatment. However, in addition to improving the credibility of static causal models, panel data are essential for estimating causal effects obtained from dynamic causal models, which is the main theme of the second part of this chapter. This chapter also shows that such dynamic causal effects have only weak links to parameters usually appearing in the dynamic panel data model literature (e.g. Arellano, Bond, 1991).

This survey has many omissions indeed. As the panel econometrics literature as well as the literature on treatment effects is huge, we had to omit several important topics to stick with the space constraints of such a handbook. First of all, all the semiparametric panel data literature is completely ignored. The interested reader is referred to the chapter by Bo Honoré (11) in this handbook. Duration models are another important omission from this chapter, for reasons of lack of space and not because of lack of relevance, although for example in the work of Abbring and van den Berg (2003) there is a clear link between panel data and the identification and estimation of causal effects. Furthermore, we ignore the extensive literature on distributional treatment effects (e.g. Firpo, 2007) as well as a substantial part of the more structural dynamics treatment literature (see, e.g. Abbring and Heckman, 2007, and Heckman and Navarro-Lozano, 2007). Finally, recent developments on testing as well as developing instrumental variable assumptions are ignored as well (for a recent example, see Klein, 2010).

Even for the subject not omitted, this chapter will neither review the whole panel literature nor the whole treatment effects literature. Instead it focuses on parts of the treatment literature where panel data are particularly helpful and important. 'Importance' and 'helpfulness' of course entirely depend on which empirical subject is analysed, in addition to some degree of subjective judgement. Therefore, this chapter takes an applied perspective in the sense of prominently using an empirical example to exemplify ideas and of treating formal assumptions and properties rather informally (and relating the reader to the corresponding papers in the literature instead). In the same thrust, we exemplify the main ideas in a very simple linear regression setting.

The main empirical example is the evaluation of active labour market programmes, which will now be introduced. This literature tries to answer the question whether unemployed benefit from participating in some public financed training or employment programme. Of course, the effects of interest of such programmes have many dimensions, usually including individual reemployment chances and earnings.[2] Many of such studies are based on reasonably large administrative data sets which allow observing individuals before, during, and after participating in a programme. Thus, usually the empirical analysis is based on panel data. The econometric methods developed in the treatment effects literature are used in the respective empirical analyses, because their main advantages mentioned above are deemed to be important. Furthermore, a diverse set of different identification and estimation strategies is employed.[3]

In the next section, the static treatment effects model is introduced. We consider three approaches that figure prominently in the applied literature. Starting with matching and

[2] The meta study by Card, Kluve, and Weber (2010) gives a comprehensive overview of recent studies in that field, and Lechner, Miquel, and Wunsch (2011) provide a recent prototypical example for Germany.

[3] For a different field applying treatment effects models with panel data, see for example Lechner (2009a), who analyses the impact of individual sports activity on labour market outcomes using the German GSOEP panel data.

regression type methods, we continue with differences-in-differences methods and instrumental variable estimation. In Section 3, we discuss matching and regression type methods within a dynamic treatment effects framework. Finally, Section 4 concludes and the Appendix contains some derivations omitted from the main body of the text.

2. The static treatment model

2.1 Notation

This chapter features the most simple static treatment effects model, namely a model in which the treatment is binary. This simplification allows concentrating on the key issues without an overly complex technical apparatus hiding the main insights relevant for empirical work even if more general models are used in some of the applied work. In this model, like in any other treatment effects model, we are interested in the effect of a ceteris paribus change of the treatment (e.g., participating in a programme), D, on the outcomes, (e.g., earnings) Y. To denote the values of the outcomes that would occur if $D=1$ or $D=0$, respectively, we define so-called potential outcomes Y^d (i.e. Y^1 and Y^0) (Rubin, 1974).[4] By construction, both potential outcomes can never be observed simultaneously. The link between observable and potential outcomes is given by the following observation rule, i.e. $Y = D\,Y^1 + (1-D)\,Y^0$. The observation rule directly implies that $E(Y_t^{d_1} \mid D_1 = d_1) = E(Y_t \mid D_1 = d_1)$.

Other variables that play a role as control variables are denoted by X (e.g. past education or the labour market history), while instrumental variables are denoted by Z. Depending on the context, X may be scalars or vectors of random variables. If not mentioned explicitly otherwise, X and Z are assumed not to be influenced by the treatment and are in this sense exogenous.

[4] Capital letters denote random variables, while small letter denote either their realisations or fixed values.

With respect to timing, we assume that the treatment occurs after period 0, i.e. $t = 0$, and before period 1. In other words, in the static model the treatment variable D_t equals 0 prior to period 1. In period 1, it switches from 0 to 1 with positive probability. Otherwise, D_t is time constant. Thus, fixing/knowing D_1 is like fixing/knowing all D_t. There may be further pre- and post-treatment periods, denoted by t. Thus, the random variables consist of several elements over time, which will be stacked on top of each other, e.g. $Y = (\ldots, Y_0, Y_1, \ldots)'$. If not mentioned otherwise, all the random variables, describing a larger population of interest (unemployed individuals in our example), and functions thereof are assumed to have as many moments as required for the particular analysis. Finally, assume that we observe data from D and Y, as well as X and Z if needed. The data is obtained from N independent draws from the population described by these random variables. In other words, (d_i,y_i,x_i,z_i) are i.i.d. in the cross-sectional dimension 'i' but may be arbitrarily correlated over time.

It is usually the goal of a treatment effect analysis to uncover causal effects aggregated over specific subpopulations while allowing individual causal effects to vary across observations in a general way.[5] In this chapter, we consider the average treatment effect (ATE, γ_t), the average treatment effect for the treated (ATET), $\gamma_t(1)$ and non-treated (ATENT), $\gamma_t(0)$, as well as the local average treatment effect (LATE, $\gamma_t(z)$), all in a particular period t.[6] These effects are defined as follows:

$$\begin{aligned}
\gamma_t &= E(Y_t^1 - Y_t^0); && && (ATE)\\
\gamma_t(d_1) &= E(Y_t^1 - Y_t^0 \mid D_1 = d_1), && \forall d_1 \in \{0,1\}; && (ATET, ATENT)\\
\gamma_t(z) &= E(Y_t^1 - Y_t^0 \mid Complier(z)); && \forall t \in \{\ldots,0,1,\ldots\}. && (LATE)
\end{aligned}$$

[5] Again, by considering only averages the focus is on the simplest case, but conceptionally most of the considerations below carry over to quantile treatment effects, or other objects for which the knowledge of the marginal distribution of the potential outcome is sufficient. Furthermore, we also do not discuss a large range of other parameters that relate for example to continuous instruments (see Heckman and Vytlacil, 2005, for an extensive discussion of such parameters).

[6] By some inconsistency of notation, $\gamma_t(d_1)$ refers to a specific population defined by the value of d_1, while $\gamma_t(z)$ refers to some population that is implied by the use of a specific instrument z (see below).

In our empirical example the ATE is the relevant object of estimation when interest is in the expected effect of the programme for a randomly chosen unemployed, while ATET will be the expected effect of a programme for a randomly chosen participant, and ATENT for a non-participant.[7] Note that ATE can be directly derived from ATET and ATENT, because $\gamma_t = \gamma_t(1)P(D_1 = 1) + \gamma_t(0)[1 - P(D_1 = 1)]$. Therefore, we focus most of the discussion only on the ATET, since the arguments for the ATENT are symmetric and ATE can always be obtained by a combination of the two. Note that one of the two terms that appear in the sum that defines ATET and ATENT can be expressed in terms of variables that have sample counterparts, i.e. $E(Y_t^{d_1} \mid D_1 = d_1) = E(Y_t \mid D_1 = d_1)$ (and can therefore consistently be estimated by the respective sample mean). However, this is not so for the second term, $E(Y_t^{d_1} \mid D_1 = 1 - d_1)$, called the counterfactual, which requires further assumptions for identification, i.e. assumptions required to express the counterfactual in terms of the observable variables *Y*, *X*, and *D* (and *Z*).

Finally, the LATE parameter, introduced by Imbens and Angrist (1994), measures the mean programme effect for a randomly drawn unemployed from a so-called complier population. A complier population is characterised by the fact that for every member a change in the value of the instrument leads to a change in the value of the treatment. Thus, in the empirical example, a complier is a person who would have a different programme participation status when faced with different values of the instrument.[8] While the data is

[7] Note that this way of coding *D* implies that we measure the effect of the intervention that occurred in period 1 only. Analysing further changes in *D* is relegated to the section on dynamic treatment models.

[8] Indeed, there may be two such groups with become either more or less likely to receive treatment for an identical change of the instrument. Usually, one of those groups, called defiers, is assumed to be absent (see Imbens, Angrist, 2004).

informative about who is a participant and who is a non-participant, it is silent about who is a complier. Usually, we expect all these effects to be zero in the pre-treatment periods, $t \leq 0$.[9]

In the following, we consider several identification and estimation strategies that are popular in empirical studies trying to uncover causal effects and discuss the value of the availability of panel studies for these strategies. All these strategies provide potential solutions when a direct comparison of the means of y_t for observations with $d_{1i}=1$ (treated) and $d_{1i}=0$ (controls) will be confounded by some other observable or unobservable variable that jointly influences the potential outcomes (Y_t^d) and the treatment (D_1).

To illustrate some of the ideas and to simplify a comparison with standard (linear) panel data methods, we specify simple linear models for the conditional expectations of the potential outcomes. These models will not be the most general possible. In particular, we abstract among other things from effect heterogeneity (implying $\gamma_t = \gamma_t(1) = \gamma_t(0)$) which plays a key role in the treatment effects literature. However, keeping this parametric example simple allows obtaining additional intuiting for most major ideas discussed in this survey.[10]

$$E(Y_t^0 \mid X_0 = x_0, D_0 = 0, D_1 = d_1, ..., D_T = d_T) = \alpha_t + x_0\beta_t + d_1\delta_t,$$
$$E(Y_t^1 \mid X_0 = x_0, D_0 = 0, D_1 = d_1, ..., D_T = d_T) = \underbrace{\alpha_t + x_0\beta_t + d_1\delta_t}_{E(Y_t^0 \mid X_0 = x_0, D_1 = d_1)} + \gamma_t;$$
$$\forall x_0 \in X_0, \forall t = \{..., 0, 1, ..., T\}.$$

T denotes the final period of data used. This (simple) specification captures the idea that the different groups (defined by treatment status) may exhibit the same effect of the treatment (allowed to be variable over time), but may have different levels in the potential outcomes. If subjects self-select or are selected into the treatments, $d_1\delta_t$ may be termed the time-varying

[9] If this is not true, for example due to changing behaviour in anticipation of treatment, then it is sometimes possible to adjust the calendar date of the treatment (i.e. period 0) just prior to the first period when such reaction could be expected.

[10] Note that these simple linear specifications sometimes allow specialised identification and estimation strategies exploiting these parametric features. As this is not the purpose of this example, such cases will be ignored in the discussions below.

conditional-on-*X* selection effect. The value of δ_t is inherently linked to the nature of the 'selection process'. For example, if treatment is assigned in a random experiment, then δ_t equals zero. If differences of average outcomes between treated and control individuals result from differences in x_0 only, then, again, δ_t equals zero.

Using the observation rule, this model leads to the following conditional expectation for Y_t:

$$E(Y_t \mid X_0 = x_0, D_t = d_t) = \alpha_t + x_0\beta_t + d_t(\gamma_t + \delta_t); \quad \forall x_0 \in X_0, \forall t = \{...,0,1,...,T\}.$$

From this equation it is obvious that additional assumptions are necessary in order to obtain consistent estimates of the treatment effect, γ_t, because it is confounded by the selection effect, δ_t.

2.2 Selection on observables: The conditional independence assumption

Non-parametric identification

In this section we analyse the case when information on background variables *X* is rich enough such that the potential outcomes are unconfounded (conditionally independent of D_1) given *X* (conditional independence assumptions, CIA). This is formalized as[11]

$$Y_t^0, Y_t^1 \coprod D_1 \mid X_0 = x_0, \quad x_0 \in \chi_0, \quad \forall t > 0.$$

This assumption states that the potential outcomes are independent (denoted by $\coprod$) of treatment in period 1 conditional on X_0 for the values of x_0 in χ_0. In addition, assume that there is common support, i.e. $0 < P(D_1|X_0=x_0) < 1$, and that X_0 is not influenced by D_1.[12]

[11] $A \coprod B \mid C = c$ means that *each element* of the vector of random variables *B* is independent of each element of the random vector *A* conditional on the random vector *C* taking values of *c* in the sense (see Dawid (1979).

[12] See the excellent survey by Imbens (2004) who extensively discusses this case.

These assumptions imply that $E(Y_t^{d_1} \mid D_1 = 1 - d_1) = E[E(Y_t \mid X_0 = x_0, D_1 = d_1) \mid D_1 = 1 - d_1]$ so that ATE, ATET, and ATENT are identified, i.e. they can be expressed in terms of random variables for which realisations are available for all sampled members of the population:

$$\gamma_t(1) = E(Y_t \mid D_1 = 1) - E[E(Y_t \mid X_0 = x_0, D_1 = 0) \mid D_1 = 1],$$
$$\gamma_t(0) = E[E(Y_t \mid X_0 = x_0, D_1 = 1) \mid D_1 = 0] - E(Y_t \mid D_1 = 0), \quad \forall x_0 \in \chi_0, \quad \forall t > 0.$$

For the linear model outlined above, note that as indicated already in the previous section, CIA implies that $\delta_t = 0$ for $t > 0$. Thus, the treatment effects can easily be obtained by standard regression methods.

Why are panel data helpful in this essentially static setting? Firstly, having further post-treatment time periods available allows estimating the dynamics of the effects. Secondly, whether this selection-on-observable assumption is plausible depends on the particular pre-treatment information available, as the data needs to contain all variables that jointly influence the treatment and the post-treatment potential outcomes. Thirdly, assuming some homogeneity over time, we may argue that the CIA also holds for outcomes prior to the treatment. If this is true, and if the treatment effect is zero for those pre-treatment periods ($\gamma_t = \gamma_t(1) = \gamma_t(0) = 0, \ t \leq 0$), we may conduct placebo tests (pre-programme tests in the language of Heckman and Hotz, 1989). If we find statistically significant non-zero effects, this will be an indication that the CIA does not hold prior to treatment. This in turn may indicate that it does not hold in the post-treatment periods either.[13]

Understanding the outcome dynamics in the empirical example is important because many programmes have initial negative effects, so-called lock-in effects (for example, due to a reduced job search while participating in a programme). Positive effects, if any, appear only later (see e.g. Lechner, Miquel, and Wunsch, 2011). The issue about using pre-treatment

[13] To be precise, such tests can be informative about confounders that are simultaneously related to the current treatment and the past and current outcomes.

variables to control for confounding may be even more important. In the case of labour market evaluations, Lechner and Wunsch (2011), for example, show the importance of controlling for variables capturing an informative individual labour market history to avoid biased estimation. It is probably true for many applications that key elements of *X* are pre-treatment outcome variables ($X_0 = (Y_{-\tau},...,Y_0, \tilde{X}_0)$; the last element in this vector denotes some other exogenous confounders). One reason for this may be that they contain the same or similar unobservables as the post-treatment variables and that such unobservables are likely to be correlated with D_1. Finally, placebo tests may or may not be an appropriate tool in practice. For example, in many countries unemployment is a requirement to become eligible for the programmes of the active labour market policy. In such case, the sample will be selected such that everybody is unemployed in $t = 0$ (otherwise there would be no common support). Then, if we estimate an effect for the pre-treatment period $t = 0$ in a placebo experiment, we will always find a zero effect, at least for the outcome variable unemployment. Thus the test has no power for this variable in this period. As outcomes are likely to be correlated over time, and as different outcome variables, like earnings and various employment indicators, are also correlated in the cross-sectional dimension, the test may generally lack power in such situations.

Estimation

For our 'toy-linear' model, the CIA implies:

$$E(Y_t \mid X_0 = x_0, D_t = d_t) = \alpha_t + x_0\beta_t + d_t\gamma_t; \quad \forall x_0 \in X_0, \forall t = \{1,...,T\}.$$

Thus, the treatment effect, γ_t, is consistently estimated by a cross-sectional regression (in the post-treatment periods) in which the observable outcome, Y_t, is regressed on X_0 and D_1. Remember that the vector of confounding variables here includes functions of past outcomes as well as other exogenous variables for which the realised values are known in period 0.

Indeed, there is no gain by using panel data methods in this model with linear confounding correction, time varying coefficients, and a treatment effect that does not vary with confounders and treatment, be it fixed or random effects. Cross-sectional regressions for post-treatment periods ($t > 0$) are consistent estimators of the treatment effects. Of course, if some of the coefficients are constant over time, then panel data methods may lead to more efficient estimates. Similar arguments will be valid if conditional expectations of the potential outcomes are nonlinear functions of the confounders, or if the effects vary with the confounders.

However, estimating a parametric model is unnecessarily restrictive since the identifying assumptions provide non-parametric identification of the mean causal effects. Therefore, it is not surprising that the literature emphasised methods that do not require the parametric assumptions (and the implied restrictions on effect heterogeneity). To estimate the effects non-parametrically, we need a non-parametric regression of $P(D_1=1|X_0=x_0)$ for weighting-type estimators (Hirano, Imbens, Ridder, 2003). Alternatively, for regression based estimators a non-parametric regression of $E(Y_t|X_0=x_0, D_1=0)$ for the ATET, of $E(Y_t|X_0=x_0, D_1=1)$ for the ATENT, or both for the ATE is required (Imbens, Newey, Ridder, 2007). At least one of those non-parametric regressions is also needed for many other non-parametric methods, like for most versions of matching (Rubin, 1979). This is the case because in the selection-on-observables framework all methods, whether parametric or non-parametric, are explicitly or implicitly based on adjusting the distribution of X_0 in the $D_1=1$ and $D_1=0$ subsamples such that the adjusted distribution of the confounders is very similar for treated and non-treated. If this is successful using the same adjustment for the outcome variables gives the desired mean causal effects. The higher the dimension of X_0, the more difficult it is to create this kind of comparability in all dimensions of X_0, and thus the curse of dimensionality comes into its damaging play.

The results of Rosenbaum and Rubin (1983) reduce this problem in some sense as they show that it is sufficient to make those subpopulations comparable with respect to a one-dimensional random variable, instead of the high-dimensional X_0. This one-dimensional random variable is the so-called propensity score, $p(X_0)$, which is the probability of treatment given the confounders, i.e. $p(x_0):=P(D_1=1|X_0=x_0)$. The methods used most in empirical work are semiparametric in the sense that the propensity score is estimated by (flexible) parametric models. Then this score is used either for weighting, regression-type adjustments, or matching estimation. Since there is nothing specific to panel data when using these methods in this context, we will refer the reader to the excellent surveys by Imbens (2004) and Imbens and Wooldridge (2009). Several Monte Carlo studies compare the performance of the various estimators, like Frölich (2004), Busso, DiNardo, and McCrary (2009a, 2009b) and the very extensive study of Huber, Lechner, and Wunsch (2013). The latter compares more than hundred different estimators using what they call an 'empirical Monte Carlo' study, which is a Monte Carlo design that shares many features with real empirical studies. In the latter study, a particular radius matching estimator with bias adjustment showed some superior large and small sample properties.

2.3 Selection on unobservables I: Difference-in-difference methods

2.3.1 Semi-parametric identification

Whereas matching-type methods discussed in the previous section may not necessarily require data from different periods, such data are essential for difference-in-difference (DiD) methods. The basic idea of the DiD concept is to have (at least) four different subsamples available for the empirical analysis: One group that has already been subject to the treatment (observed in $t > 0$), one group that will be subject to the treatment in the future (observed in $t \leq 0$), and another two groups not subject to the treatment that are observed in the same

periods as the two treatment groups. If the treatment has no effect in period 0 and if the outcomes of the treatment and non-treatment groups develop in the same fashion over time (usually called either 'common-trend' or 'bias stability' assumption), then, conceptionally, we may either (i) use period '0' to estimate the bias of any estimator based on selection-on-observables (since the true effect is 0 in period 0) and use this estimate to purge the similar estimate in $t > 0$ from this bias, or (ii) use the change of the outcome variables of the non-treatment group over time together with the pre-treatment outcomes of the future treated to estimate what would have happened to the treated group in $t > 0$ had they not been treated.

These ideas are indeed old and can at least be traced back to a paper by Snow (1855). He was interested in whether cholera was transmitted by (bad) air or (bad) water. Snow (1855) 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, to isolate the effect of the water quality from other confounders. In our days there are many applications of these methods, mainly in applied microeconomics. They are also well explained in most modern econometric textbooks (see for example the excellent discussions in Angrist and Pischke, 2009). Since these methods are also contained in several excellent surveys on treatment effects (e.g., Blundell and Costa Dias, 2009, and Imbens and Wooldridge, 2009), I keep this section brief and reiterate a few panel data related points that appeared in my recent survey on DiD estimation (Lechner, 2011a).

The common-trend assumption,

$$
\begin{aligned}
&E\left(Y_1^0 \mid X_0 = x_0, D_1 = 1\right) - E\left(Y_0^0 \mid X_0 = x_0, D_1 = 1\right) = \\
&E\left(Y_1^0 \mid X_0 = x_0, D_1 = 0\right) - E\left(Y_0^0 \mid X_0 = x_0, D_1 = 0\right) = \\
&E\left(Y_1^0 \mid X_0 = x_0\right) - E\left(Y_0^0 \mid X_0 = x_0\right), \qquad \forall x_0 \in \chi_0,
\end{aligned}
$$

together with the assumptions that (i) D_1 has no effect prior to treatment, i.e. $\gamma_t(d) = 0, t \leq 0$, (ii) the covariates are not influenced by the treatment, and that (iii) there is

the necessary common support, $\gamma_t(1)$, are identified.[14] Note that identification is not non-parametric (as in the previous section) in the sense that the validity of the common-trend assumption depends on the chosen transformation (unit of measurement) of the outcome variables. In other words, if the common-trend assumption is deemed to be correct, for example, for earnings it will be violated for non-linear transformations of earnings, like log-earnings (at least for non-trivial cases). As it is usually difficult to explain why such an assumption should only be valid for a particular functional form, this is a limitation of this method (see the generalisation of Athey and Imbens, 2006, which is however more difficult to apply).[15]

Going back to our 'toy' model and forming the differences for the non-participation potential outcomes, we obtain the following expressions:

$$E(Y_t^0 - Y_\tau^0 \mid X_0 = x_0, D_1 = 1) = (\alpha_t - \alpha_\tau) + x_0(\beta_t - \beta_\tau) + (\delta_t - \delta_\tau),$$
$$E(Y_t^0 - Y_\tau^0 \mid X_0 = x_0, D_1 = 0) = (\alpha_t - \alpha_\tau) + x_0(\beta_t - \beta_\tau); \qquad \tau \in \{...,0\}, t \in \{1,...,T\}.$$

Thus, the required condition for the common trend assumption to hold is that the impact of the selection effect is time constant $(\delta_t - \delta_\tau) = 0$, at least for the (minimum of) two periods used for estimation. This may seem to be somewhat more general than in the case of CIA which required the absence of post-treatment confounding, i.e. $\delta_t = 0$. However, since obviously $\delta_t = 0$ does not imply $(\delta_t - \delta_\tau) = 0$, the two methods are not nested. On top of this, DiD also requires the absence of pre-treatment effects $(\gamma_\tau = 0, \tau \leq 0)$.[16] Under these

[14] This literature usually attempts only to identify effects for treated. Although identifying effects for non-treated would technically just involve a redefinition of the treatment, this setting is usually unattractive in empirical studies, because it requires three treated groups one of which becomee non-treated from period 0 to period 1.

[15] See Lechner (2011a) for more discussion on how to deal with non-linearities in this approach.

[16] Note that although this is not required by CIA in general, once pre-treatment outcomes are used as covariates they must be exogenous. Of course, this is only plausible in the absence of pre-treatment effects.

assumptions, we obtain the following conditional expectations for the observable outcome variables:

$$E(Y_t \mid X_0 = x_0, D_1 = d_1, ..., D_T = d_T) = \alpha_t + x_0\beta_t + \underline{1}(t > 0)d_t\gamma_t + d_1\delta .$$

$\underline{1}(\cdot)$ denotes the indicator function which is one if its argument is true. Thus the treatment effects can be recovered by regression methods.

2.3.2 Estimation

The name of the estimation strategy is already indicative of the underlying estimation principle in general. If the common-trend assumption holds conditional on X_0, then the estimate of the effect conditional on X can be obtained by forming the differences of the pre- and post-treatment periods' outcomes of the treated and subtracting the differences of the pre- and post-treatment periods' outcomes of the non-treated. In the (virtual) second step the conditional-on-X effects are averaged with weights implied by the distribution of X among the treated.

For our simple linear model this leads to the specification given above. Clearly, the treatment effects cannot be estimated with one period of data alone because of the presence of the selection term, $d_1\delta$, which was absent in the model of the section discussing the selection-on-observables only. Within a post-treatment cross-section this selection effect cannot be distinguished from the causal effect $\underline{1}(t > 0)d_t\gamma_t$. Further note that panel data are not necessary for estimating the parameters of this equation. Repeated cross-sections of at least one pre-treatment and one post-treatment period are sufficient as long as they contain also the information about the (past and future) treatment status and confounders.

If the linear model is not deemed to be appropriate for modelling the conditional expectations, then there are some non-linear and/or less parametric methods available, many

of which are discussed in Lechner (2011a). Therefore, for the sake of brevity the interested reader is referred to that paper.

2.3.3 The value of panel data compared to repeated cross-sections

In the previous sections we saw that panel data allowed us to (i) follow the outcome dynamics, (ii) compute more informative control variables, and (iii) check the credibility of the identifying assumptions with placebo tests. While (ii) always requires panel data, for (i) and (iii) it is only essential to have data from additional periods (so that repeated cross-sections are sufficient). The same is true for DiD.

If panel data are available the linear DiD estimator can be estimated by fixed effects methods:[17] 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. Therefore, it is not surprising that adding fixed individual effects instead of the treatment group dummy d in the regression formulation leads to the same quantity to be estimated (e.g. Angrist and Pischke, 2009). This way it becomes obvious, as it was for our 'toy'-model, that the usual advantages attributed to fixed effects models, like controlling of time constant endogeneity and selectivity within a linear setting, are also advantages of the difference-in-difference approach.

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 (because it is likely that pre-treatment outcomes are correlated with post-treatment outcomes as well as

[17] The remaining part of this section follows closely section 3.2.8 of Lechner (2011a).

selection, either directly, or because the unobservables that influence those three quantities are correlated).[18] This conditioning is not possible with repeated cross-sections since we do not observe pre- and post-treatment outcomes of the same individuals.

The corresponding matching-type assumptions for the case when lagged outcome variables are available (and used) imply the following:

$$E\left(Y_t^0 \mid Y_0 = y_0, X_0 = x_0, D_1 = 1\right) = E\left(Y_t^0 \mid Y_0 = y_0, X_0 = x_0, D_1 = 0\right), \quad \forall t > 0.$$

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, as mentioned before, is that it allows for time constant confounding unobservables ($\delta_t \neq 0$) while requiring common-trends ($\delta_t = \delta_\tau$), whereas matching does not require common-trends ($\delta_t \neq \delta_\tau$) but assumes that conditional on pre-treatment outcomes confounding unobservables are irrelevant ($\delta_t = 0$). As δ_t, δ_τ capture the effects of variables jointly influencing selection as well as outcomes, their interpretation depend on the conditioning sets used. For example, if the selection process is entirely governed by x_0 and y_τ, then controlling for those variables implies $\delta_t = 0$. In this case matching may be used and there is no need for any assumptions concerning the selection process in period τ. More generally, one may argue that conditioning on the past outcome variables already controls for the part of the unobservables that manifested themselves in the lagged outcome variables.

[18] Note that although such an intuition of controlling for more information is plausible in many applications, it is easy to create an example with a larger and a smaller conditioning set for which CIA holds in the *smaller* but not in the larger set.

One may try to combine the positive 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 differencing of DiD and to focus on the post-treatment variables alone. Thus, such a procedure implicitly requires the matching assumptions. 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 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, Chabé-Ferret (2012) gives several examples in which a difference-in-difference strategy leads to a consistent estimator while matching conditional on past outcomes may be biased. However, even for those examples given, the assumptions necessary for the consistency of DiD require substantive knowledge on how the selection bias impacts the potential outcomes, which are similar to our toy-model. He also shows simulations that indicate that for the case when the assumptions for matching on lagged outcomes as well as for DiD are not exactly fulfilled, both estimators are biased, but matching appears to be more robust than DiD. He concludes that for the cases for which one or the other set of assumptions is not clearly preferred on theoretical grounds, results from both estimation strategies should be presented.

2.4 Selection on unobservables II: Instrumental variables

2.4.1 Non-parametric identification

Either when selection-on-observables or differences-in-differences approaches are not credible, or when the instrument-specific LATE parameter is the more interesting parameter compared to the ATE, ATET, or ATENT,[19] then instrumental variable estimation may be the method of choice. The seminal paper by Imbens and Angrist (1994) increased considerably our understanding of which kind of causal effect is estimated by 2SLS when effects are heterogeneous. This literature was further extended by Heckman (1997), Vytlacil (2002), and Heckman and Vytlacil (2005) for continuous instruments as well as Abadie (2003) and Frölich (2007) for ways to deal with covariates. These papers also clarify that with heterogeneous effects the IV assumptions have to be strengthened somewhat. In other words, on top of the assumption that the instrument has no effect on the outcomes other than by changing the treatment (exclusion restriction, no direct effect assumption), the assumption that a change in the instrument affects the treatment only in one direction (i.e. it either increases or decreases treatment probability for all), the so-called monotonicity assumption, is required as well.

As before, the key question for this chapter is about the role of panel data in IV estimation. As before, the first benefit panel data provide is that observing more post treatment outcomes allows uncovering how the effects of the treatment in period 1 evolve over time. Secondly, instruments may not be valid unconditionally and current period control variables may not help as they might already be affected by the treatment. In this case observing more pre-treatment variables may be very helpful. In our example of active labour

[19] For example, Frölich and Lechner (2010) analyse the effects of active labour market programmes and argue that the compliers that relate to their instruments are close to a population that would join the programmes if they were marginally extended. In fact, for the policy question about the effects of extending the programmes estimating such a parameter would be more interesting than estimating the ATE, the ATET, or the ATENT.

market policy evaluation, Frölich and Lechner (2010) used an instrument that measured on which side of a regional boarder within a local labour market an unemployed lived. The rational for this instrument was that this fact mattered for their programme participation probability but not (directly) for their labour market success. The concern in the paper was that they might have chosen one or the other side of the boarder by considerations that could involve other characteristics, like tax rates and past labour market success. These factors may be however related to outcomes via different channels than programme participation thus violating the exclusion restriction. With panel data we are able to condition on such past events and thus improve the credibility of the instrument. The third benefit one might derive from panel data is that past values of some variables that are not time constant may provide instruments. A word of caution is in order in this instant, because there are a couple of empirical papers that use lagged outcomes as instruments without giving the explicit reasoning that would justify doing so. This is somewhat at odds with the arguments made in the previous section about the value of lagged outcomes as a confounding control variable, because by definition a confounder has a direct effect on outcomes thus violating one of the key assumptions required for consistent IV estimation. In other words, as it is likely that those lagged outcome variables depend on the same unobservable than the current period outcome variables do, one needs very explicit arguments why this should not matter with respect to the exclusion restriction in the particular study at hand. The fourth benefit of panel data, namely placebo tests, is that it may allow estimating effects for periods in which the true effect is known to be zero (and the instrument is valid as well), thus providing some empirical evidence on the credibility of the instrument.

2.4.2 Estimation

The easiest way to conceptualize the linear model is to follow exactly the same steps as for selection-on-observables, and to assume that one of the confounders that are contained in

the required conditioning set X_0 is unobservable. Thus the linear model for the observable outcomes derived above cannot be a basis for consistent estimation by regression methods. Note that by the definition of a confounder as a variable jointly correlated with treatments and outcomes, this leads to the endogeneity of D_t in the regression formulated in terms of observable variables. In this case, and if a valid instrument is available, the panel econometric IV methods for linear models, described for example in Baltagi (2008) and Biorn and Krishnakumar (2008), may be applied to obtain estimates that are consistent under the linearity and homogeneity assumptions discussed in the previous section.

For non- or semiparametric estimation similar problems concerning the dimension of the confounders in case of selection-on-observables occur. Frölich (2007) showed that the IV estimate is a ratio of estimators that would be consistent under a no-confounding assumption of the relation of the instrument and the outcome. In fact, IV estimates can be obtained by dividing the effect of the instrument Z on the outcome Y_t by the effect of Z on D_1 each time controlling for variables, X_0, that are jointly related to the instrument and to the outcome or the treatment.[20] Since these are similar estimation strategies as described in the section on selection-on-observables the same tools for reducing the dimension are available. The only difference is of course that the propensity score in this case is the probability of the binary instrument (instead of the treatment) being one given the confounders, i.e. $p^z(x_0) := P(D_1=1|X=x_0)$.

[20] Note that although the same notation X is used here for both variables, usually these 'instrument confounders' may be different from the 'treatment confounders' that are required under the CIA.

3. A dynamic treatment model

3.1 Motivation and basic structure of the model

The static treatment model, which is widely used in micro econometrics even when panel data are used, allows for dynamics in the sense that the effects of *the* treatment, D_1, are allowed to vary over time, and that variables measured in pre-treatment periods were used to tackle the confounding problem in different ways. The treatment itself, however, was not allowed to change over time more than once (from period 0 to period 1). In this section we present a model that allows for more treatment dynamics. For such a model, the availability of panel data is essential.

Robins (1986) suggested an explicitly dynamic causal framework based on potential outcomes. It allows the definition of causal effects of dynamic interventions and clarifies the resulting endogeneity and selectivity problems. Identification is achieved by sequential selection-on-observable assumptions (see Abbring, 2003, for a comprehensive summary).[21] His approach was subsequently applied in epidemiology and biostatistics (e.g. Robins, 1989, 1997, 1999, Robins, Greenland, and Hu, 1999, for discrete treatments; Gill and Robins, 2001, for continuous treatments; and many other applications by various authors) to define and estimate the effect of time-varying treatments in discrete time. It is common in that literature to estimate the effects by parametric models usually based on the so-called G-computation algorithm as suggested by Robins (1986).

Lechner and Miquel (2010, LM10 further on) extend Robins' (1986) framework to different causal parameters. Since the assumptions used in LM10 are similar to the selection-

[21] Until now identification of dynamic treatment models by instrumental variable methods and generalized difference-in-difference methods is a rather unexplored area, although there are some results in Miquel (2002, 2003), that awaits further research. Therefore, this section focuses entirely only on the sequential-selection-on-observable approach proposed by Robins (1986) in his seminal paper. Alternative reduced form approaches have been suggested for example by Fitzenberger, Osikuminu, and Paul (2010).

on-observables or conditional independence assumptions (CIA) of the static model, Lechner (2009b) proposed dynamic extensions of the matching and inverse-probability-weighting estimators discussed above, which are more robust than parametric models. The applications of this approach in economics are limited so far.[22] One reason is that this approach, in particular in its semi-parametric and non-parametric form requires larger and more informative data than required for estimating causal effects in a static treatment effects model.

Below, the definitions of the dynamic causal model as well as the identification results derived by Robins (1986) and Lechner and Miquel (2010) are briefly reviewed.[23] To ease the notational burden, we use a three-periods-binary-treatment model to discuss the most relevant issues that distinguish the dynamic from the static model.[24] Using again our labour market programme evaluation example for illustration, suppose that, as before, there is an initial pre-treatment period, $D_0=0$, plus two subsequent periods in which different treatment states (participation in a programme) are realized. Denote the history of variables up to period *t* by a bar below that variable, i.e. $\underline{d}_2 = (0, d_1, d_2)$.[25] Therefore, in this setting all treatment combinations are fully described by the four sequences (0,0), (1,0), (0,1), and (1,1). The potential outcomes are indexed by these treatment combinations, $Y_t^{\underline{d}_1}$ $(t \geq 1)$ or $Y_t^{\underline{d}_2}$ $(t \geq 2)$. They are measured at the end of each period, whereas treatment status is measured at the beginning of each period. For each sequence length of length of one or two periods (plus the initial period), one of the respective potential outcomes is observable:

[22] Exceptions are Lechner and Wiehler (2013) who analyse the effects of the timing and order of Austrian active labour market programs and LM10 who analyse the effects of the German active labour market policies. A further exception is Ding and Lehrer (2003) who use this framework and related work by Miquel (2002, 2003) to evaluate a sequentially randomized class size study using difference-in-difference-type estimation methods. Lechner (2008) discusses practical issues when using this approach for labour market evaluations.

[23] The dynamic potential outcome framework is also useful to compare concepts of causality used in microeconometrics and time series econometrics (see Lechner, 2011b, for details).

[24] As before, there may be more periods available to measure pre- or post-treatment outcomes though.

[25] Therefore, the first element of this sequence, d_0, is mainly ignored in the notation as it does not vary.

$$Y_t = D_1 Y_t^1 + (1 - D_1) Y_t^0, \quad \forall t \geq 1;$$

$$Y_t = D_1 Y_t^1 + (1 - D_1) Y_t^0 = D_1 D_2 Y_t^{11} + (1 - D_1) D_2 Y_t^{01} + D_1 (1 - D_2) Y_t^{10} + (1 - D_1)(1 - D_2) Y_t^{00}; \quad \forall t \geq 2.$$

Finally, note that the confounders, X_t, will be explicitly considered to be time varying and may contain functions of Y_t. Like the outcomes they are observable at the end of each period.

As for the static model, the causal effect of the sequences is formalized using averages of potential outcomes. The following expression defines the causal effect (for period *t*) of a sequence of treatments up to period 1 or 2, $\underline{d}_\tau^k$, compared to an alternative sequence of the same or a different length, $\underline{d}_{\tau'}^l$, for a population defined by one of those sequences or a third sequence, $\underline{d}_{\tilde{\tau}}^j$:

$$\gamma_t^{\underline{d}_\tau^k, \underline{d}_{\tau'}^l}(\underline{d}_{\tilde{\tau}}^j) = E(Y_t^{\underline{d}_\tau^k} \mid \underline{D}_{\tilde{\tau}} = \underline{d}_{\tilde{\tau}}^j) - E(Y_t^{\underline{d}_{\tau'}^l} \mid \underline{D}_{\tilde{\tau}} = \underline{d}_{\tilde{\tau}}^j),$$

$$0 \leq \tilde{\tau}; 1 \leq \tau, \tau' \leq 2, \tilde{\tau} \leq \tau', \tau; \tilde{\tau}, \tau', \tau \leq t; \; k \neq l, \; k \in (1, ..., 2^\tau), \; l \in (1, ..., 2^{\tau'}), \; j \in (1, ..., 2^{\tilde{\tau}}).$$

The treatment sequences indexed by *k*, *l*, and *j* may correspond to $d_1=0$ or $d_1=1$ if τ (or τ') denotes period 1, or to the longer sequences $(d_1, d_2)= (0,0), (0,1), (1,0)$, or $(1,1)$ if τ (or τ') equals two. LM10 call $\gamma_t^{\underline{d}_\tau^k, \underline{d}_{\tau'}^l}$ the dynamic average treatment effect (DATE). Accordingly, $\gamma_t^{\underline{d}_\tau^k, \underline{d}_{\tau'}^l}(\underline{d}_\tau^k)$ is termed DATE on the treated (DATET). There are also cases in-between, like $\gamma_t^{\underline{d}_2^k, \underline{d}_2^l}(\underline{d}_1^l)$, for which the conditioning set is defined by a sequence shorter than the one defining the causal contrast. Finally, note that the effects are symmetric for the same population ($\gamma_t^{\underline{d}_\tau^k, \underline{d}_{\tau'}^l}(\underline{d}_\tau^k) = -\gamma_t^{\underline{d}_{\tau'}^l, \underline{d}_\tau^k}(\underline{d}_\tau^k)$, but $\gamma_t^{\underline{d}_\tau^k, \underline{d}_{\tau'}^l}(\underline{d}_\tau^k) \neq \gamma_t^{\underline{d}_\tau^k, \underline{d}_{\tau'}^l}(\underline{d}_{\tau'}^l)$). This feature, however, does not restrict effect heterogeneity.

Let us now postulate a simple linear model that serves the same purpose as in the case of the static model:

$$E(Y_t^0 \mid X_0 = x_0, D_1 = d_1) = \alpha_{1,t} + x_0\beta_{1,t} + d_1\delta_{1,t},$$
$$E(Y_t^1 \mid X_0 = x_0, D_1 = d_1) = \alpha_{1,t} + \underbrace{x_0\beta_{1,t} + d_1\delta_{1,t}}_{E(Y_t^0 \mid X_0 = x_0, D_1 = d_1)} + \gamma_t^1, \quad \forall x_0 \in \chi_0, \quad \forall t = \{..., 0, 1, ..., T\}.$$

Therefore, for the 'observable' outcomes the observation rule implies the following:

$$E(Y_t \mid X_0 = x_0, D_1 = d_1) = \alpha_{1,t} + x_0\beta_{1,t} + d_1(\gamma_t^1 + \delta_{1,t}).$$

Note that this part of the specification that relates to the effects of the treatments in period 1 only is specified exactly as for the static model to ease comparison (with the exception of not conditioning on all D_t, which has a different meaning in the dynamic than in the static model). Therefore, it is also clear that identification of $\gamma_{1,t}$ is exactly as for the static model discussed in the previous section. Therefore, from now on we concentrate on sequences that include treatment status in period 2 as well.

The key features that the toy model for dynamic treatments is supposed to capture are related to the impact of confounders already influenced by the treatment in period one as well as the selection effects that come from selecting into D_1 in period 1 and into D_2 in period 2. The following specifications contain these features while keeping all other complications to a minimum:

$$E(Y_t^{00} \mid \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T) = \alpha_{2,t} + x_0\beta_{2,t} + y_1\phi_t + d_1d_2\lambda_t^{11} + (1-d_1)d_2\lambda_t^{01} + d_1(1-d_2)\lambda_t^{10};$$
$$E(Y_t^{d_1', d_2'} \mid \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T) = \underbrace{\alpha_{2,t} + x_0\beta_{2,t} + y_1\phi_t + d_1d_2\lambda_t^{11} + (1-d_1)d_2\lambda_t^{01} + d_1(1-d_2)\lambda_t^{10}}_{E(Y_t^{00} \mid \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T)}$$
$$+ \gamma_t^{d_1'd_2'};$$
$$\forall x_0, y_1, d_2, d_1.$$

Note that the coefficients $\lambda_t^{d_1,d_2}$ denote the subsample specific selection effects. Next, we derive the conditional expectation of the observable outcome for this case using the observation rules given above.

$$E(Y_t \mid \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T) = E(Y_t \mid \underline{X}_1 = \underline{x}_1, \underline{D}_2 = \underline{d}_2) =$$
$$= \alpha_{2,t} + x_0\beta_{2,t} + y_1\phi_t + d_1d_2(\gamma_t^{11} + \lambda_t^{11}) + (1-d_1)d_2(\gamma_t^{01} + \lambda_t^{01}) + d_1(1-d_2)(\gamma_t^{10} + \lambda_t^{10}).$$

In a similar fashion as before, this equation shows that the treatment effects are not identified without further assumptions that concern the selection effects, $\lambda_t^{d_1d_2}$. It is also important to note that this model is not a typical dynamic panel data model, as the conditioning is not on y_{t-1}, but on y_1, i.e. it does not depend on t.

3.2 Identification

The weak dynamic conditional independence assumption (W-DCIA) postulates that the variables that jointly influence selection at each stage of the sequence as well as the outcomes are observable in the time period corresponding to that stage:

$$\begin{aligned} &a) \quad Y_t^{00}, Y_t^{10}, Y_t^{01}, Y_t^{11} \coprod D_1 \mid X_0 = x_0, \\ &b) \quad Y_t^{00}, Y_t^{10}, Y_t^{01}, Y_t^{11} \coprod D_2 = d_2 \mid D_1 = d_1, \underline{X}_1 = \underline{x}_1, \end{aligned} \quad \forall \underline{x}_1 \in \underline{\chi}_1,\ \forall t \geq 1.$$

$\underline{\chi}_1 = (\chi_0, \chi_1)$ denotes the support of X_0 and X_1. Part a) of W-DCIA states that the potential outcomes are independent of treatment choice in period 1 (D_1) conditional on X_0. This is the standard version of the static CIA. Part b) states that conditional on the treatment in period 1 and on the confounding variables of periods 0 and 1, $\underline{X}_1$, potential outcomes are independent of participation in period 2 (D_2).

In Appendix A it is shown that using this assumption and the observation rule gives us the relation between shorter and longer sequences of potential outcomes (which also provides the link between the static and the dynamic models):

$$E(Y_t^{d_1} \mid D_1 = d_1) = E(Y_t^{d_1 1} \mid D_1 = d_1, D_2 = 1)\, p^{D_2}(d_1) + E(Y_t^{d_1 0} \mid D_1 = d_1, D_2 = 0)\left[1 - p^{D_2}(d_1)\right];$$

$$E(Y_t^{1-d_1} \mid D_1 = d_1) = \underset{X_0 \mid D_1 = d_1}{E}\left[E(Y_t^{(1-d_1)1} \mid X_0 = x_0, D_1 = 1-d_1, D_2 = 1)\, p^{D_2|X}(x_0, 1-d_1)\right]$$
$$+ \underset{X_0 \mid D_1 = d_1}{E}\left[E(Y_t^{(1-d_1)0} \mid X_0 = x_0, D_1 = 1-d_1, D_2 = 0)\left[1 - p^{D_2|X}(x_0, 1-d_1)\right]\right];$$

$$p^{D_2}(d_1) := P(D_2 = 1 \mid D_1 = d_1); \quad p^{D_2|X}(x_0, d_1) := P(D_2 = 1 \mid X_0 = x_0, D_1 = d_1).$$

Thus the expectation of the outcomes of the shorter sequences is a weighted average of the expectation of the two longer sequences that have the same first period treatment as the shorter sequence.

To see whether the W-DCIA is plausible in our example, the question is which variables influence programme participation in each period as well as subsequent labour market outcomes and whether such variables are observable. If the answer to the latter question is yes (and if there is common support, i.e. there are individuals with the same observable characteristics that are observed in both treatment sequences of interest), then there is identification, even if some or all conditioning variables in period 2 are influenced by the labour market and programme participation outcomes of period 1. LM10 show that, for example, quantities that are for subpopulations defined by treatment status in period 1 or 0 only, like $E(Y_2^{11})$, $E(Y_2^{11} \mid D_1 = 0)$, and $E(Y_2^{11} \mid D_1 = 1)$, are identified. Mean potential outcomes for subpopulations defined by treatment status in period 1 and 2 are only identified if the sequences coincide in the first period (e.g., $E[Y_2^{11} \mid \underline{D}_2 = (1,0)]$). However, $E[Y_2^{11} \mid \underline{D}_2 = (0,0)]$ or $E[Y_2^{11} \mid \underline{D}_2 = (0,1)]$ are not identified. Thus, $\gamma_t^{\underline{d}_\tau^k, \underline{d}_\tau^l}$ and $\gamma_t^{\underline{d}_\tau^k, \underline{d}_\tau^l}(d_1^j)$ are identified $\forall d_1^k, d_2^k, d_1^l, d_2^l, d_1^j, d_2^j \in \{0,1\}$, but $\gamma_2^{\underline{d}_2^k, \underline{d}_2^l}(\underline{d}_2^j)$ is not identified if $d_1^l \neq d_1^k$, $d_1^l \neq d_1^j$, or $d_1^k \neq d_1^j$. The relevant distinction between the populations defined by participation states in period 1 and subsequent periods is that in period 1, treatment choice is random conditional on exogenous variables, which is the result of the initial condition stating that $D_0 = 0$ holds for everybody. However, in the second period, randomization into these treatments is conditional

on variables already influenced by the first part of the treatment. W-DCIA has an appeal for applied work as a natural extension of the static framework. However, W-DCIA is not strong enough to identify the classical treatment effects on the treated which would define the population of interest using one of the complete sequences (for all three periods), if the sequences of interest differ in period 1.

Let us now consider identification in our linear example. Note that part b) of the W-DCIA implies that $E(Y_t^{d_1',d_2'} \mid \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T) = E(Y_t^{d_1',d_2'} \mid \underline{X}_1 = \underline{x}_1, D_1 = d_1)$, thus $\lambda_t^{11} = \lambda_t^{10} (= \lambda_t^1)$ and $\lambda_t^{01} = \lambda_t^{00} (= 0)$, thus we have:

$$E(Y_t^{00} \mid \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T) = E(Y_t^{00} \mid \underline{X}_1 = \underline{x}_1, D_1 = d_1) = \alpha_{2,t} + x_0\beta_{2,t} + y_1\phi_t + d_1\lambda_t^1;$$
$$E(Y_t^{d_1',d_2'} \mid \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T) = E(Y_t^{d_1',d_2'} \mid \underline{X}_1 = \underline{x}_1, D_1 = d_1) = \alpha_{2,t} + x_0\beta_{2,t} + y_1\phi_t + d_1\lambda_t^1 + \gamma_t^{d_1'd_2'}.$$

$$E(Y_t \mid \underline{X}_1 = \underline{x}_1, \underline{D}_T = \underline{d}_T) = \alpha_{2,t} + x_0\beta_{2,t} + y_1\phi_t + d_1d_2(\gamma_t^{11} + \lambda_t^1) + \\ + (1-d_1)d_2\gamma_t^{01} + d_1(1-d_2)(\gamma_t^{10} + \lambda_t^1).$$

Thus, γ_t^{01} as well as $\alpha_{2,t}, \beta_{2,t}, \phi_t$ are identified. Furthermore since part a) of the W-DCIA implies that $\delta_{1,t} = 0$, the effects of the treatment of the first period, $\alpha_{1,t}, \beta_{1,t}, \gamma_t^1$, are identified as well. However, there is still a selection effect, λ_t^1, that hinders full identification of the causal effects, i.e. only $(\gamma_t^{11} + \lambda_t^1)$ and $(\gamma_t^{10} + \lambda_t^1)$ are identified from this regression.

However, this is already enough because the treatment effects are linked in the sense that γ_t^1 must, by definition, be a weighted average of γ_t^{11} and γ_t^{10}. Indeed the appendix shows that $\gamma_t^1 = (\gamma_t^{11} - \gamma_t^{10}) \underset{X_0|D_1=0}{E}\left[P(D_2 = 1 \mid X_0 = x_0, D_1 = 1)\right] - \gamma_t^{01} P(D_2 = 1 \mid D_1 = 0) + \gamma_t^{10}$ holds in this model, which provides identification (because (i) $\underset{X_0|D_1=0}{E}\left[P(D_2 = 1 \mid X_0 = x_0, D_1 = 1)\right]$ and $P(D_2 = 1 \mid D_1 = 0)$ are identified; (ii) γ_t^{11} can be expressed in terms of identified terms and γ_t^{10}

; (iii) thus $(\gamma_t^{11} + \lambda_t^1)$ and $(\gamma_t^{10} + \lambda_t^1)$ depend only on two further unknowns and have, for non-trivial cases, a solution).

For the general model, LM10 show that to identify all treatment parameters, W-DCIA must be strengthened by essentially imposing that the confounding variables used to control selection into the treatment of the second period are not influenced by the selection into the first-period treatment. This can be summerized by an independence condition like $Y_t^{d_2} \coprod \underline{D}_2 \mid \underline{X}_1 = \underline{x}_1$ (LM10 call this the strong dynamic conditional independence assumption, *S-DCIA*). Note that the conditioning set includes the outcome variables from the first period. This is the usual conditional independence assumption used in the multiple static treatment framework (with four treatments; see Imbens, 2000, and Lechner, 2001). In other words, when the control variables (including the outcome variables in period 1) are not influenced by the previous treatments, the dynamic problem collapses to a static problem of four treatments with selection on observables. An example of such a situation would be an assignement to a two subsequent training programmes which was made already before the first programme began and for which there is no chance to drop out once assigned to both programmes.

Any attempt of non-parametrically estimating these effects faces the same problem that distributional adjustments based on a potentially high-dimensional vector of characteristics and intermediate outcomes are required. However, as before for the static case, propensity scores are available to allow the construction of semi-parametric estimators (see LM10 for details).

3.3 Estimation

Lechner (2008, L08 further on) shows that for the model using W-DCIA these propensity scores are convenient tools for constructing sequential propensity score matching

and reweighting estimators.[26] Using such propensity scores, the following identification result based on W-DCIA is the key ingredient for building appropriate estimators:

$$E(Y_2^{\underline{d}_2^k} \mid D_1 = d_1^j) = \underset{p^{d_1^k}(x_0)}{E}\left\{ \underset{p^{\underline{d}_2^k|d_1^k}(\underline{x}_1)}{E}[E(Y_2 \mid \underline{D}_2 = \underline{d}_2^k, \underline{p}^{\underline{d}_2^k|d_1^k, d_1^k}(\underline{x}_1)) \mid D_1 = d_1^k, p^{d_1^k}(x_0)] \mid D_1 = d_1^j \right\},$$

$$\underline{p}^{\underline{d}_2^k|d_1^k, d_1}(\underline{X}_1) := [p^{\underline{d}_2^k|d_1^k}(\underline{X}_1), p^{d_1}(X_0)], \quad \forall d_1^k, \underline{d}_2^k, d_1^j, d_1 \in \{0,1\},$$

where $p^{\underline{d}_2^k|d_1^k}(\underline{x}_1) := p^{\underline{d}_2^k|d_1^k} P(D_2 = \underline{d}_2^k \mid D_1 = d_1^k, \underline{X}_1 = \underline{x}_1)$ and $p^{d_1}(x_0) = P(D_1 = d_1 \mid X_0 = x_0)$ are the respective participation probabilities. To learn the counterfactual outcome for the population participating in d_1^j (the target population) had they participated in the sequence $\underline{d}_2^k$, characteristics (and thus outcomes) of observations with $\underline{d}_2^k$ must be reweighted to make them comparable to the characteristics of the observations in the target population (d_1^j). The dynamic, sequential structure of the causal model restricts the possible ways to do so. Intuitively, for the members of the target population, observations that share the first element of the sequence of interest (d_1^k) should be reweighted such that they have the same distribution of $p^{d_1^k}(X_0)$ as the target population. Call this artificially created group comparison group one. Yet, to estimate the effect of the full sequence, the outcomes of observations that share $\underline{d}_2^k$ instead of d_1^k are required. Thus, an artificial subpopulation of observations in $\underline{d}_2^k$ that has the same distribution of characteristics of $p^{d_1^k}(X_0)$ and $p^{\underline{d}_2^k|d_1^k}(\underline{X}_1)$ as the artificially created comparison group 1 is required. The same principle applies for dynamic average treatment effects in the population (DATE).

[26] Of course, other static matching-type estimators (e.g. Huber, Lechner, Wunsch, 2013) can be adapted to the dynamic context in a similar way.

All proposed estimators in L08 have the same structure: They are computed as weighted means of the outcome variables observed in the subsample $D_2 = d_2^k$. The weights, $w(\cdot)$, depend on the specific effects of interest and are functions of the balancing scores.

$$\widehat{E(Y_2^{\underline{d}_2^k} \mid D_1 = d_1^j)} = \sum_{i \in \underline{d}_2^k} w_i^{\underline{d}_2^k, d_1^j} (\underline{p}^{\underline{d}_2^k \mid d_1^k, d_1^k}(\underline{x}_{1,i}), d_1^k)\, y_i; \quad w_i^{\underline{d}_2^k, d_1^j} \geq 0; \quad \sum_{i \in \underline{d}_2^k} w_i^{\underline{d}_2^k, d_1^j} = 1; \tag{1}$$

$$\widehat{E(Y_2^{\underline{d}_2^k})} = \sum_{i \in \underline{d}_2^k} w_i^{\underline{d}_2^k} (\underline{p}^{\underline{d}_2^k \mid d_1^k, d_1^k}(\underline{x}_{1,i}), d_1^k)\, y_i; \quad w_i^{\underline{d}_2^k} \geq 0; \quad \sum_{i \in \underline{d}_2^k} w_i^{\underline{d}_2^k} = 1. \tag{2}$$

It remains to add a note on estimation of our linear toy model. Following the considerations in the identification part, the model consists of two linear regressions based on the following two equations:

$$E(Y_t \mid X_0 = x_0, D_1 = d_1) = \alpha_{1,t} + x_0 \beta_{1,t} + d_1 \gamma_t^1;$$
$$E(Y_t \mid \underline{X}_1 = \underline{x}_1, \underline{D}_2 = \underline{d}_2) = \alpha_{2,t} + x_0 \beta_{2,t} + y_1 \phi_t + d_1 d_2 (\gamma_t^{11} + \lambda_t^1) + (1 - d_1) d_2 \gamma_t^{01} + d_1 (1 - d_2)(\gamma_t^{10} + \lambda_t^1).$$

In a second step the estimated coefficients together with the link between the one-period and two-period treatment effects are used to uncover the causal effects. Since these effects are assumed to be homogenous, the W-DCIA is sufficient to identify all relevant quantities of this model. It is important to note, though, that the outlined procedure is very different from estimating a classical dynamic or static linear or non-linear panel data model.

4. Concluding remarks

For many empirical applications panel data are essential for the credible identification and precise estimation of causal effects. The first part of this chapter, which discussed matching and instrumental variable estimation in the static treatment model, showed how the additional information provided by panel data can be used to measure pre-treatment variables that improve the credibility of those strategies. Furthermore, if several post-treatment periods

were available, more interesting effects capturing the outcome dynamics can be estimated. The latter was also true for the so-called difference-in-difference approach, although the use of the pre-treatment outcomes differ for this approach: lagged outcomes do not appear in the conditioning set but were used instead to form pre-treatment-post-treatment outcome differences. Thus, the latter approach is not robust to non-linear transformations of the outcome variables while the former two approaches are robust to such transformations. Another difference between IV, matching, and difference-in-difference approaches is that for the latter panel data are not strictly necessary as repeated cross-sections will do. Finally, for all approaches based on a static treatment framework panel data may allow for so-called placebo tests, i.e. estimating effects for periods for which it is known that they should be zero. Such tests are another tool of improving the credibility of the chosen identifying assumptions.

The second part of this chapter showed how panel data can be used to identify and estimate causal parameters derived from dynamic treatment effect models, an area which did not yet receive much attention in econometrics. Therefore, the results on non-parametric identification and non- or semi-parametric estimation are mainly limited to the case of imposing sequential selection-on-observable assumptions, a case which is popular in other fields as well, like epidemiology. It is a perhaps a surprising insight from this analysis that the parameters usually estimated by linear parametric panel data models and the causal parameters derived from the dynamic treatment models are only loosely related.

There are still many open ends in this literature. For example, in the dynamic treatment models instrumental variable estimation seems to be rather unexplored, while for the static models we just start to understand when it makes more sense to use lagged outcome variables as covariates instead of forming differences and apply a difference-in-difference approach instead. In conclusion, we can expect the intersection of the literatures on panel data and treatment effects to produce many interesting research papers in the near future.

5. References

Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models", *Journal of Econometrics*, 113, 231-263.

Abbring, J. H. (2003): "Dynamic Econometric Program Evaluation", IZA, Discussion Paper, 804.

Abbring, J. H., and G. J. van den Berg (2003), "The nonparametric identification of treatment effects in duration models", *Econometrica*, 71, 1491–1517.

Abbring, J. H., and J. J. Heckman (2007): "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation", *Handbook of Econometrics*, vol. 6B, Elsevier 5145–5303.

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.

Angrist, J. D., and J.-S. Pischke (2010): "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics", *Journal of Economic Perspectives*, 24, 3–30.

Arellano, M., and S. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations", *Review of Economic Studies*, 58, 277-297.

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

Baltagi, B. (2008): *Econometric Analysis of Panel Data*, Wiley.

Biorn, E., and J. Krishnakumar (2008): "Measurement Errors and Simultaneity", in Mátyás L., and P. Sevestre (eds.), *The Econometrics of Panel Data*, Chapter 10, 323-367.

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

Busso, M., J. DiNardo, and J. McCrary (2009a): "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects", forthcoming in the *Journal of Business and Economic Statistics*.

Busso, M., J. DiNardo, and J. McCrary (2009b): "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators", IZA Discussion Paper 3998.

Card, D. J. Kluve, and A. Weber (2010): "Active Labour Market Policy Evaluations: A Meta-Analysis," *Economic Journal*, 120, F452–F477.

Chabé-Ferret, S. (2012): "Matching vs. Differencing when Estimating Treatment Effects with Panel Data: the Example of the Effect of Job Training Programs on Earnings", *Toulouse School of Economics Working Paper*, 12-356.

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

Deaton, A. (2010): "Instruments, Randomization and Learning about Development", *Journal of Economic Literature*, 48, 424–455.

Ding, W., and S. F. Lehrer (2003): "Estimating Dynamic Treatment Effects from Project STAR," mimeo.

Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects", *Econometrica*, 75, 259-276.

Fitzenberger, B., Osikuminu, A., and M. Paul (2010): “The Heterogeneous Effects of Training Incidence and Duration on Labor Market Transitions”, IZA Discussion Paper No. 5269.

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

Frölich, M. (2007): "Nonparametric IV estimation of local average treatment effects with covariates", *Journal of Econometrics*, 139, 35-75.

Frölich, M., and M. Lechner (2010): “Exploiting Regional Treatment Intensity for the Evaluation of Labour Market Policies,” *Journal of the American Statistical Association*, 105, 1014-1029.

Gill, R. D., and J. M. Robins (2001): "Causal Inference for Complex Longitudinal Data: The Continuous Case." *The Annals of Statistics*, 1-27.

Heckman, J. J. (1997): "Instrumental Variables", *Journal of Human Resources*, 32, 441-462.

Heckman, J. J. (2010): “Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy,” *Journal of Economic Literature*, 48, 356–398.

Heckman, J. J., and E. Vytlacil (2005): "Causal Parameters, Structural Equations, Treatment Effects and Randomized Evaluation of Social Programs", *Econometrica*, 73, 669–738.

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., and Navarro-Lozano, S. (2007): “Dynamic Discrete Choice and Dynamic Treatment Effects”, *Journal of Econometrics*, 136, 341–396.

Heckman, J. J., R. J. LaLonde, J. A. Smith (1999): “The economics and econometrics of active labor market programs”. In: Ashenfelter, O., Card, D. (Eds.), *Handbook of Labor Economics*, vol. 3A. North-Holland, New York, 1865–2097 (Chapter 31).

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

Huber, M., M. Lechner, and C. Wunsch (2013): “The Performance of Estimators Based on the Propensity Score,” *Journal of Econometrics*.

Imbens G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62, 467–475.

Imbens, G. W. (2000): "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87, 706-710.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review", *Review of Economics and Statistics*, 86, 4-29.

Imbens, G. W. (2010): “Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)”, *Journal of Economic Literature* 48, 399–423.

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

Imbens, G. W., W. Newey, and G. Ridder (2007): "Mean-squared-error Calculations for Average Treatment Effects", IRP discussion paper.

Klein, T. J. (2010): "Heterogeneous treatment effects: Instrumental variables without monotonicity?," *Journal of Econometrics*, 155(2), 99-116,

Lechner, M. (2001): "Identification and estimation of causal effects of multiple treatments under the conditional independence assumption," in Lechner, M., Pfeiffer, F. (eds.), *Econometric Evaluation of Active Labour Market Policies*, Physica, Heidelberg, 43-58.

Lechner, M. (2008): "Matching estimation of dynamic treatment models: Some practical issues", in D. Millimet, J. Smith, and E. Vytlacil (eds.), *Advances in Econometrics, Volume 21, Modelling and Evaluating Treatment Effects in Econometrics*.

Lechner, M. (2009a): "Long-run labour market and health effects of individual sports activities", *The Journal of Health Economics*, 28, 839-854.

Lechner, M. (2009b): "Sequential Causal Models for the Evaluation of Labor Market Programs," *Journal of Business & Economic Statistics*, 27, 71-83.

Lechner, M. (2011a): "The Estimation of Causal Effects by Difference-in-Difference Methods", *Foundations and Trends in Econometrics*, 4/3, 165–224.

Lechner, M. (2011b): "The Relation of Different Concepts of Causality used in Time Series and Microeconometrics", *Econometric Reviews*, 30, 109-127.

Lechner, M., and C. Wunsch (2011): "Sensitivity of Matching-Based Program Evaluations to the Availability of Control Variables", *Labour Economics*.

Lechner, M., and R. Miquel (2010): "Identification of the Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions ", *Empirical Economics*, 39, 111-137.

Lechner, M., and S. Wiehler (2013): "Does the order and timing of active labor market programs matter?" forthcoming in the *Oxford Bulletin of Economics and Statistics.*

Lechner, M., R. Miquel, and C. Wunsch (2011): "Long-Run Effects of Public Sector Sponsored Training in West Germany", *Journal of the European Economic Association*, 9, 742-784.

Mátyás, L., and P. Sevestre (2008): "*The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practic* ", 3rd ed., Springer.

Miquel, R. (2002): "Identification of Dynamic Treatments Effects by Instrumental Variables," University of St. Gallen, Department of Economics, Discussion Paper, 2002-11.

Miquel, R. (2003): "Identification of Effects of Dynamic Treatments with a Difference-in-Differences Approach." University of St. Gallen, Department of Economics, Discussion paper, 2003-06.

Robins, J. M. (1986): "A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect," *Mathematical Modeling*, 7:1393-1512, with 1987 "Errata to: A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect," *Computers and Mathematics with Applications*,

14:917-921; 1987 "Addendum to: A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect," *Computers and Mathematics with Applications*, 14:923-945; and 1987 Errata to: "Addendum to 'A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect'," *Computers and Mathematics with Applications*, 18: 477.

Robins, J. M. (1989): "The Analysis of Randomized and Nonrandomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in Sechrest, L., Freeman, H., Mulley, A. (eds.), *Health Service Research Methodology: A Focus on Aids*, 113-159, Washington, D.C., Public Health Service, National Centre for Health Services Research.

Robins, J. M. (1997): "Causal Inference from Complex Longitudinal Data. Latent Variable Modelling and Applications to Causality," in Berkane, M. (Ed.), *Lecture Notes in Statistics* (120), Springer, New York, 69-117.

Robins, J. M. (1999): "Association, Causation, and Marginal Structural Models," *Synthese*, 121, 151-179.

Robins, J. M., S. Greenland, and F. Hu, (1999): "Estimation of the Causal Effect of a Time-varying Exposure on the Marginal Mean of a Repeated Binary Outcome," *Journal of the American Statistical Association*, 94, 687-700.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 66, 688-701.

Rubin, D. B. (1979): "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies", *Journal of the American Statistical Association*, 74, 318-328.

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

Vytlacil, E. (2002): „Independence, monotonicity and latent variable models: an equivalence result", *Econometrica*, 70, 331-341.

Appendix: Relation of different potential outcomes in the dynamic treatment model

In this appendix we provide the derivations that lead to the link of the potential outcomes of sequences of length one and two. First, define the following short-cut notation for the conditional selection probabilities.

$$p^{D_2}(d_1) := P(D_2 = 1 | D_1 = d_1); \quad p^{D_2|X}(x_0, d_1) := P(D_2 = 1 | X_0 = x_0, D_1 = d_1).$$

Next, we use the observation rule to establish the desired relation:

$$E\left[D_1 Y_t^1 + (1 - D_1) Y_t^0 | D_1 = 1\right] = E\left[Y_t^1 | D_1 = 1\right] = E\left[D_2 Y_t^{11} + (1 - D_2) Y_t^{10} | D_1 = 1\right].$$

Using iterated expectations, for the general case we obtain the following expression:

$$\begin{aligned} E(Y_t^{d_1} | D_1 = d_1) &= E(Y_t^{d_1} | D_1 = d_1, D_2 = 1) p^{D_2}(d_1) + E(Y_t^{d_1} | D_1 = d_1, D_2 = 0)\left[1 - p^{D_2}(d_1)\right] \\ &= E(Y_t^{d_1 1} | D_1 = d_1, D_2 = 1) p^{D_2}(d_1) + E(Y_t^{d_1 0} | D_1 = d_1, D_2 = 0)\left[1 - p^{D_2}(d_1)\right]. \end{aligned}$$

Next, we want to establish a similar link for the mean counterfactual $E(Y_t^{d_1} | D_1 = 1 - d_1)$, which requires the use of part a) of the W-DCIA.

$$\begin{aligned} &E(Y_t^{1-d_1} | D_1 = d_1, X_0 = x_0) \overset{W-DCIA\,a)}{=} E(Y_t^{1-d_1} | X_0 = x_0, D_1 = 1 - d_1) \Rightarrow \\ &E(Y_t^{1-d_1} | D_1 = d_1) = \underset{X_0 | D_1 = d_1}{E} E(Y_t^{1-d_1} | X_0 = x_0, D_1 = 1 - d_1) = \\ &\quad = \underset{X_0 | D_1 = d_1}{E}\left[E(Y_t^{(1-d_1)1} | X_0 = x_0, D_1 = 1 - d_1, D_2 = 1) p^{D_2|X}(x_0, 1 - d_1)\right] \\ &\quad + \underset{X_0 | D_1 = d_1}{E}\left[E(Y_t^{(1-d_1)0} | X_0 = x_0, D_1 = 1 - d_1, D_2 = 0)\left[1 - p^{D_2|X}(x_0, 1 - d_1)\right]\right] \end{aligned}$$

This formula can now be used to connect the treatment effects as well:

$$\begin{aligned} \gamma_t^1(1) &= E(Y_t^1 | D_1 = 1) - E(Y_t^0 | D_1 = 1) = \\ &= E(Y_t^{11} | D_1 = 1, D_2 = 1) p^{D_2}(1) + E(Y_t^{10} | D_1 = 1, D_2 = 0)\left[1 - p^{D_2}(1)\right] - \\ &\quad - \underset{X_0 | D_1 = 1}{E}\left[E(Y_t^{01} | X_0 = x_0, D_1 = 0, D_2 = 1) p^{D_2|X}(x_0, 0)\right] \\ &\quad - \underset{X_0 | D_1 = 1}{E}\left[E(Y_t^{00} | X_0 = x_0, D_1 = 0, D_2 = 0)\left[1 - p^{D_2|X}(x_0, 0)\right]\right]; \end{aligned}$$

$$\gamma_t^1(0) = \underset{X_0|D_1=0}{E}\Big[E(Y_t^{11} \mid X_0 = x_0, D_1 = 1, D_2 = 1)\, p^{D_2|X}(x_0,1)\Big] +$$
$$+ \underset{X_0|D_1=0}{E}\Big[E(Y_t^{10} \mid X_0 = x_0, D_1 = 1, D_2 = 0)\Big[1 - p^{D_2|X}(x_0,1)\Big]\Big] -$$
$$-\Big[E(Y_t^{01} \mid D_1 = 0, D_2 = 1)\, p^{D_2}(0) + E(Y_t^{00} \mid D_1 = 0, D_2 = 0)\Big[1 - p^{D_2}(0)\Big]\Big].$$

Finally, we consider the special case of the dynamic linear toy model postulated for $\gamma_t^1(0)$:

$$\gamma_t^1 = \gamma_t^1(1) = \gamma_t^1(0) = \underset{X_0|D_1=0}{E}\Big[E(Y_t^{11} \mid X_0 = x_0, D_1 = 1, D_2 = 1)\, p^{D_2|X}(x_0,1)\Big] +$$
$$+ \underset{X_0|D_1=0}{E}\Big[E(Y_t^{10} \mid X_0 = x_0, D_1 = 1, D_2 = 0)\Big[1 - p^{D_2|X}(x_0,1)\Big]\Big] -$$
$$-\Big[E(Y_t^{01} \mid D_1 = 0, D_2 = 1)\, p^{D_2}(0) + E(Y_t^{00} \mid D_1 = 0, D_2 = 0)\Big[1 - p^{D_2}(0)\Big]\Big] =$$
$$= \underset{X_0|D_1=0}{E}\Big[E(Y_t^{11} - Y_t^{10} \mid X_0 = x_0, D_1 = 1, D_2 = 1)\, p^{D_2|X}(x_0,1) + E(Y_t^{10} \mid X_0 = x_0, D_1 = 1, D_2 = 0)\Big] -$$
$$-\Big[E(Y_t^{01} - Y_t^{00} \mid D_1 = 0, D_2 = 1)\, p^{D_2}(0) + E(Y_t^{00} \mid D_1 = 0, D_2 = 0)\Big] =$$
$$= \underset{X_0|D_1=0}{E}\Big[E(Y_t^{11} - Y_t^{10} \mid X_0 = x_0, D_1 = 1, D_2 = 1)\, p^{D_2|X}(x_0,1)\Big] -$$
$$-\Big[E(Y_t^{01} - Y_t^{00} \mid D_1 = 0, D_2 = 1)\, p^{D_2}(0)\Big] +$$
$$+ \underset{X_0|D_1=0}{E}\Big[E(Y_t^{10} \mid X_0 = x_0, D_1 = 1, D_2 = 0)\Big] - E(Y_t^{00} \mid D_1 = 0, D_2 = 0) =$$
$$= \underset{X_0|D_1=0}{E}\Big[E(Y_t^{11} - Y_t^{10} \mid X_0 = x_0, D_1 = 1, D_2 = 1)\, p^{D_2|X}(x_0,1)\Big] -$$
$$-\Big[E(Y_t^{01} - Y_t^{00} \mid D_1 = 0, D_2 = 1)\, p^{D_2}(0)\Big] +$$
$$+ \underset{X_0|D_1=0}{E}\Big[E(Y_t^{10} - Y_t^{00} \mid X_0 = x_0, D_1 = 1, D_2 = 0)\Big] =$$
$$= (\gamma_t^{11} - \gamma_t^{10})\, \underset{X_0|D_1=0}{E}\Big[p^{D_2|X}(x_0,1)\Big] - \gamma_t^{01} p^{D_2}(0) + \gamma_t^{10}.$$