On the testability of common trends in panel data without placebo periods*

Martin Huber†

May 3, 2024

Abstract

We demonstrate and discuss the testability of the common trend assumption imposed in Difference-in-Differences (DiD) estimation in panel data when not relying on multiple pre-treatment periods for running placebo tests. Our testing approach involves two steps: (i) constructing a control group of non-treated units whose pre-treatment outcome distribution matches that of treated units, and (ii) verifying if this matched control group and the original non-treated group share the same time trend in average outcomes. Testing is motivated by the fact that in several (but not all) panel data models, a common trend violation across treatment groups implies and is implied by a common trend violation across pre-treatment outcomes. For this reason, the test verifies a sufficient, but (depending on the model) possibly stronger than necessary condition for DiD-based identification. We investigate the finite sample performance of a testing procedure that is based on double machine learning, which permits controlling for covariates in a data-driven manner, in a simulation study and also apply it to labor market data from the National Supported Work Demonstration.

JEL Classification: C12, C21

Keywords: treatment effects, causality, difference-in-differences, hypothesis test, panel data.

*I am grateful to Kaspar Wüthrich for helpful comments and suggestions.
†University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland; martin.huber@unifr.ch

1 Introduction

The Difference-in-Differences (DiD) approach is a popular method for identifying treatment effects by comparing differences in average outcomes before and after treatment introduction across treated and non-treated groups, as exemplified by seminal applications such as Snow (1855) and Ashenfelter (1978). A crucial assumption for identification in DiD models is the common trend assumption, which imposes that the mean potential outcome without treatment follows the same time trend in both treatment groups, possibly conditional on observed covariates, as discussed in Abadie (2005), among many others. While the common trend assumption is not directly testable for the before-and-after treatment periods used for effect estimation, placebo tests can be conducted if multiple pre-treatment periods are available.

In this paper, we propose an alternative test for the common trend assumption that does not require multiple pre-treatment periods, but is applicable if the DiD analysis utilizes panel data with the same units observed in pre- and post-treatment periods. The testing approach consist of two steps: (i) constructing a control group of non-treated observations that resembles or matches the treatment group in terms of the distribution of pre-treatment outcomes, and (ii) examining whether this matched control group and the original (i.e. non-matched) non-treated group share the same time trend in terms of their average outcomes under non-treatment. The intuition behind this approach is that in lieu of the treatment group, whose non-treated outcome trend cannot be observed, we utilize the matched control group, which is comparable to the treatment group in terms of pre-treatment outcomes but does not receive the treatment, to check whether its non-treated outcomes follow the same trend as those of the original non-treated group. This testing approach is motivated by the fact that, in several (albeit not all) panel models, a common trend across treatment groups implies and is implied by a common trend across pre-treatment outcomes. This follows in several models from a correlation between the treatment and pre-treatment outcomes under treatment selection bias, which is the reason for applying the DiD approach.

Similar to placebo tests in pre-treatment periods, we note that the suggested method is not a direct test of the common trend assumption, as the potential outcomes of treated units are never observed. However, for a range of models that appear plausible in empirical applications, the average pre- and post-treatment outcomes of the matched control group serve as proxies for the mean potential outcomes that the treatment group would have realized under non-treatment, allowing us to verify the common trend assumption. If the average outcomes across the matched control and total non-treated groups exhibit the same trend, it suggests that differences in the

pre-treatment outcome distributions of treated and non-treated groups do not imply differences in the time trends of the mean potential outcome under non-treatment.

As a caveat, we acknowledge the existence of models, as considered in Ghanem et al. (2022), in which the condition we test may be violated despite the satisfaction of common trends across treatment groups required for DiD-based identification. Such a scenario arises, for instance, when there are time-varying unobservables that affect the outcome in the same period but are not correlated over time and are unrelated to the treatment, which is known as strict exogeneity, while time constant unobservables affect the outcomes in either period and the treatment, but satisfy the common trend assumption. In such cases, the pre-treatment outcome becomes correlated with the outcome trend through differences in time-varying unobservables over time, leading to the failure of our testable condition. However, the common trend assumption required for DiD-based identification remains valid, as it does not depend on utilizing pre-treatment outcomes. For this reason, we emphasize that the test verifies a sufficient condition for identification, which is stronger than necessary. However, in scenarios where strict exogeneity, which precludes effects of past outcomes or time-varying unobservables associated with the outcomes on the treatment, seems doubtful, our testing approach provides valuable guidance w.r.t. DiD-based identification.

Our testing approach is easily implemented based on existing treatment effect estimators like matching as e.g. discussed in Rosenbaum and Rubin (1983) and Rosenbaum and Rubin (1985), regression, inverse probability weighting (IPW), see Horvitz and Thompson (1952), or doubly robust (DR) estimators, see Robins et al. (1994), Robins et al. (1995), and Robins and Rotnitzky (1995). First, such estimators are utilized to generate a matched control group among non-treated observations whose pre-treatment outcome distribution corresponds to that of the treated group. Second, DiD estimation is applied to the matched control group and the total non-treated group to verify if the time trends in average outcomes are the same for both groups. In this approach, we can easily control for pre-treatment covariates if the common trend assumption is assumed to hold conditionally given those covariates. Moreover, if the set of potential covariates is large, we may control for them in a data-driven manner based on approaches that combine treatment evaluation based on DR estimators with machine learning, such as the double machine learning (DML) framework proposed by Chernozhukov et al. (2018).

We investigate the finite sample performance of our DML-based testing procedure in a simulation study, which points to decent empirical size and power in the simulation designs considered. We also provide an empirical application to labor market data from the National Supported Work Demonstration (NSW) previously analyzed in LaLonde (1986), Dehejia and Wahba

(1999), and Dehejia and Wahba (2002). The NSW, our treatment of interest, was a labor market program in the 1970s aimed at enhancing the employment prospects of disadvantaged individuals through subsidized employment and job training. In the non-experimental NSW data, in which treated subjects from the experimental NSW study were combined with non-treated individuals from the U.S. Panel Study of Income Dynamics, our test rejects the null hypothesis that the matched control and original non-treated groups follow the same average outcome trend when flexibly controlling for several pre-treatment covariates like age, education, ethnicity, and employment status.

Our paper is related to placebo tests for DiD based on pre-treatment periods, as for instance discussed in the surveys by Lechner (2011), and Roth et al. (2022). However, instead of pre-treatment periods, it exploits the panel structure of the data to construct such tests. Furthermore, our testing approach is closely related to that of Huber and Kueck (2022) on testing identification of treatment effects in observational data based on an suspected instrument that is associated with the treatment and is supposedly unrelated with the outcome other than via the treatment. In fact, utilizing the pre-treatment outcome as 'instrument' and the outcome difference in post- and pre-treatment periods as outcome variable in the framework of Huber and Kueck (2022) is the base for the DiD testing approach suggested in this paper.

The remainder of this study is organized as follows. Section 2 discusses assumptions required for identifying the average treatment effect on the treated (ATET) by DiD and testing the common trend assumption by our method. Section 3 discusses causal scenarios in which our testable implications as well as DiD-based identification fails, but effects are identified based on a selection-on-observables assumption w.r.t. the pre-treatment outcome. Section 4 discusses the testable implications and DiD-based identification in the presence of time varying outcome shocks. Section 5 considers conditional versions of the common trend assumptions and the testable implications when controlling for observed covariates. Section 6 proposes machine learning-based tests of common trends that control for covariates in a data-driven manner. Section 7 provides a simulation study analyzing the finite sample performance of the test. Section 8 presents an application to NSW data. Section 9 concludes.

2 Identifying assumptions and testable implications

We consider a panel data DiD setting with n units, $i = 1, \ldots, n$, that are observed over two periods, $t \in \{0, 1\}$. Let us denote by D_i a binary treatment variable whose causal effect is of interest. In the pre-treatment period ($t = 0$), no unit receives the treatment; in the post-

treatment period ($t = 1$), a subset of units, referred to as the treatment group, receives the treatment, while the remaining units, referred to as the control group, remain untreated. The observed outcome for unit i in period t is denoted by Y_{it}, such that Y_{i0} and Y_{i1} are the observed pre- and post-treatment outcomes, respectively. To define the causal effects of interest, we make use of the potential outcome framework (Neyman, 1923; Rubin, 1974). We denote by $Y_{it}(d)$ the potential outcome of unit i that would be hypothetically realized under treatment $d \in \{0, 1\}$ in period t.[1] In general, we will use upper case letters to refer to random variables and lower case letters for specific values of these random variables. Moreover, we will suppress the index i to alleviate the notation.

Subsequently, we discuss two assumptions necessary for identifying causal effects in DiD settings. The firs one is the common trend assumption, which imposes that the time trend in the mean potential outcomes under non-treatment is constant across treatment groups.

Assumption 1 (Common trends)**.**

$$E[Y_1(0) - Y_0(0)|D = 1] = E[Y_1(0) - Y_0(0)|D = 0] = E[Y_1(0) - Y_0(0)].$$

It is this common trend assumption our approach aims to test, even though we cannot verify it directly as the potential outcomes under non-treatment of treated units are never observed. The second assumption (often implicitly) imposed when applying DiD is a no anticipation assumption.

Assumption 2 (No anticipation)**.** $Y_0(1) = Y_0(0)$.

Assumption 2 implies that subjects not yet treated in the pre-treatment period do not anticipate their treatment in a way that already influences their pre-treatment outcomes. As our testing approach does not aim at testing this assumption, we will maintain Assumption 2 throughout the paper.

Under Assumptions 1 and 2, the average treatment effect on the treated (ATET) in period $t = 1$, $\Delta_{D=1}$, is identified by the difference in the before-after (treatment introduction) change

[1] By denoting the potential outcome $Y(d)$ as a function of a subject's own treatment status $D = d$ only, we implicitly impose the assumption that someone's potential outcomes are not affected by the treatment status of others. This is known as the 'stable unit treatment value assumption' (SUTVA), see for instance the discussion in Rubin (1980) and Cox (1958), and is invoked throughout.

of average observed outcomes across treatment groups:

$$\Delta_{D=1} = \underbrace{E[Y_1|D=1] - E[Y_0|D=1]}_{\text{before-after change among treated}} - \underbrace{\{E[Y_1|D=0] - E[Y_0|D=0]\}}_{\text{before-after change among non-treated}}. \tag{2.1}$$

Next, we will consider a couple of further assumptions, which are required for a fruitful application of out test. The first assumption is that the mean potential outcomes under non-treatment in the pre-treatment period differ across treated and non-treated groups.

Assumption 3 (Treatment selection bias)**.**

$$E[Y_0(0)|D=1] \neq E[Y_0(0)|D=0].$$

Assumption 3 implies that selection into treatment is non-random w.r.t. pre-treatment outcomes. In fact, allowing for such a treatment selection bias is the main motivation for applying DiD approaches, which permit consistent estimation of the ATET if the difference in mean potential outcomes under non-treatment across treatment groups remains constant or stable over time. The latter selection bias stability condition is equivalent to the common trend assumption, which can be easily seen by rearranging the first equality in Assumption 1 to $E[Y_1(0)|D=1] - E[Y_1(0)|D=0] = E[Y_0(0)|D=1] - E[Y_0(0)|D=0]$. Bias stability is for instance satisfied in a fixed effects model with panel data, in which a time constant unobserved term (like innate ability) is correlated with the treatment (e.g. a training program) and has a time constant effect on the outcomes (e.g. wage) across pre- and post-treatment periods. Assumption 3 is testable in the data by verifying whether the difference in mean outcomes across treated and non-treated groups in the pre-treatment period is different from zero. To see this, note that

$$\begin{aligned} E[Y_0|D=1] - E[Y_0|D=1] &= E[Y_0(1)|D=1] - E[Y_0(0)|D=1] \\ &= E[Y_0(0)|D=1] - E[Y_0(0)|D=1], \end{aligned} \tag{2.2}$$

where the first equality follows from the observational rule, implying that $Y_t = Y_t(1)$ for observations with $D=1$, and the second from the no anticipation assumption.

We note that under a violation of Assumption 3 and the absence of treatment selection bias in the pre-treatment period, testing the common trend assumption becomes obsolete, as the DiD approach collapses to taking mean outcome differences in the post-treatment period

only. To see this, note that $E[Y_1|D=1] - E[Y_0|D=1] - \{E[Y_1|D=0] - E[Y_0|D=0]\}$ is equal to $E[Y_1|D=1] - E[Y_1|D=1]$ if $E[Y_0|D=1] = E[Y_0|D=0]$. Accordingly, the common trend assumption $E[Y_1(0) - Y_0(0)|D=1] = E[Y_1(0) - Y_0(0)|D=0]$ simplifies to $E[Y_1(0)|D=1] = E[Y_1(0)|D=0]$ if $E[Y_0(0)|D=1] = E[Y_0(0)|D=0]$, implying that the treatment is assumed to be mean independent of the potential outcomes under non-treatment. For this reason, we will maintain Assumption 3 throughout, otherwise the application of a DiD strategy does not appear meaningful.

Our next assumption imposes that the common trend assumption also holds for the treated group and a matched control group that resembles the treated group in terms of the distribution of $Y_0(0)$, i.e. the potential outcome under non-treatment in the pre-treatment period.

Assumption 4 (Common trends across matched groups)**.**

$$E[E[Y_1(0) - Y_0(0)|Y_0, D=0]|D=1] = E[Y_1(0) - Y_0(0)|D=1].$$

Assumption 4 states that the trend of the average outcome in the matched control group is representative for that of the mean potential outcome under non-treatment in the treated group. Consequently, the joint satisfaction of Assumptions 1 and 4 implies that

$$E[E[Y_1(0) - Y_0(0)|Y_0, D=0]|D=1] = E[Y_1(0) - Y_0(0)|D=0] = E[Y_1(0) - Y_0(0)]. \quad (2.3)$$

By the observational rule, this yields the testable implication

$$E[E[Y_1 - Y_0|Y_0, D=0]|D=1] = E[Y_1 - Y_0|D=0]. \quad (2.4)$$

Verifying implication (2.4) requires that for every value of the pre-treatment outcome among the treated, non-treated observations with the same value of the pre-treatment outcome exist. This is a common support restriction, which means that the conditional treatment probability given any value of the pre-treatment outcome, the so-called propensity score, must be smaller than one, as imposed in our next assumption:

Assumption 5 (Common support)**.**

$$\Pr(D=1|Y_0) < 1.$$

While imposing Assumption 4 in addition to Assumption 1 appears to be the weakest possible statistical assumption for implementing our testing approach, it is a high level condition whose intuition might not seem straightforward. However, it turns out that in several models, both Assumptions 1 and 4 are implied by a more easily interpretable assumption imposing common trends across pre-treatment outcomes:

Assumption 6 (Common trends across pre-treatment outcomes)**.**

$$E[Y_1(0) - Y_0(0)|Y_0] = E[Y_1(0) - Y_0(0)].$$

The reason for Assumption 6 generally implying Assumption 1 ais that by Assumption 3, Y_0 is correlated with D due to treatment selection bias. This in turn means that if the outcome trend is not correlated with the treatment, as imposed in Assumption 1, it is not correlated with pre-treatment outcomes either, as imposed in Assumption 6. Only in very specific models it can be the case that differential trends conditional on Y_0, i.e. $E[Y_1(0) - Y_0(0)|Y_0 = y, D = 0] \neq E[Y_1(0) - Y_0(0)|Y_0 = y, D = 1]$ for some y, offset each other in particular ways such that Assumptions 4 and 1 as well as implication (2.4) hold.

Under Assumption 1, the additional satisfaction of Assumption 6 generally implies that the treatment and the average outcome trend remain independent even conditional on pre-treatment outcome Y_0, apart from such special cases with offsetting associations of the outcome trend and Y_0 across values of D:

$$E[Y_1(0) - Y_0(0)|D = 0] = E[Y_1(0) - Y_0(0)|Y_0] = E[Y_1(0) - Y_0(0)|Y_0, D = 0]. \tag{2.5}$$

Therefore, when invoking Assumptions 1 and 6 rather than Assumptions 1 and 4, the testable implication (2.4) may be simplified by dropping the outer expectation right of the equals sign:

$$E[Y_1 - Y_0|Y_0, D = 0] = E[Y_1 - Y_0|D = 0]. \tag{2.6}$$

This result, which we formalize in Theorem 1 further below, is closely related to Theorem 2 in Huber and Kueck (2022). The latter states that if an instrumental variable and an outcome are independent conditional on the treatment, then both the instrument and the treatment are exogenous apart from special cases that violate causal faithfulness, an assumption ruling out offsetting causal effects, as discussed further below. Our pre-treatment outcome Y_0 and outcome change $Y_1 - Y_0$ correspond to the instrument and the outcome, respectively, in the framework

of Huber and Kueck (2022). It is worth noting that we only condition on $D = 0$ (not $D = 1$) when formulating the testable implication, as the common trend assumptions only concern the potential outcomes under non-treatment. In models respecting causal faithfulness, the satisfaction of equation (2.6) thus implies that both the pre-treatment outcome and treatment are mean independent of the outcome trend under non-treatment, which is sufficient (albeit stronger than necessary) for the satisfaction of Assumption 1 required for identification of the ATET.

To gain further intuition, we subsequently discuss Assumption 6 in the context of in parametric and nonparametric models where confounding of the treatment D and the outcome change $Y_1 - Y_0$ is due to unobserved fixed effects. We start by considering the following parametric outcome and treatment models:

$$Y_T = \alpha_T + \beta_T D + U, \quad D = I\{\alpha_D + U > 0\} \tag{2.7}$$

where U is an unobserved fixed (i.e. time-constant) effect that affects the pre- and post-treatment outcomes Y_0, Y_1 as well as the treatment, thus causing treatment endogeneity or selection bias. β_T is the treatment effect, which is period-dependent and by our linear model homogeneous across individuals. When imposing Assumption 2 to rule out anticipation effects as we do in the subsequent discussion, $\beta_0 = 0$, while β_1 is non-zero if the treatment has an effect on the outcome. α_T is a period-specific constant term such that a non-zero difference between α_0 and α_1 implies a non-zero time trend in the outcome. $I\{\cdot\}$ in the treatment model denotes the indicator function, which is equal to 1 if its argument is satisfied and zero otherwise. Furthermore, α_D is a constant term. Thus, the treatment is characterized by a simple binary choice model where treatment assignment depends on the size of the fixed effect.

In this model, the potential outcomes under treatment and non-treatment in the post-treatment period correspond to $Y_1(1) = \alpha_1 + \beta_1 + U$ and $Y_1(0) = \alpha_1 + U$, respectively, and $Y_1(1) - Y_1(0) = \beta_1$ gives the homogeneous treatment effect. Concerning the potential outcomes in the pre-treatment period, we have that $Y_0(1) = Y_0(0) = \alpha_0 + U$. Furthermore, the time trend of the potential outcome under non-treatment is homogeneous, i.e. for any individual it holds that $Y_1(0) - Y_0(0) = \alpha_1 - \alpha_0$. Therefore, the time trend is independent of the pre-treatment outcome Y_0 and the treatment state D, such that Assumption 6 and thus, both Assumptions 1 and 4 are satisfied, as well as the testable implications (2.4) and (2.6). As $E[Y_1|D = 1] - E[Y_0|D = 1] = \alpha_1 - \alpha_0 + \beta_1$, while $E[Y_1|D = 0] - E[Y_0|D = 0]$ yields the time trend $\alpha_1 - \alpha_0$, it holds that the DiD approach, $E[Y_1|D = 1] - E[Y_0|D = 1] - \{E[Y_1|D =$

$0] - E[Y_0|D = 0]\}$, yields the treatment effect β_1 as suggested by the testable implications.

It is worth noting that Assumption 6 also holds when generalizing the outcome and treatment models in (2.7) to some extent. First, we can allow for effect heterogeneity, e.g. including an interaction of the treatment with the fixed effect in the outcome model:

$$Y_T = \alpha_T + \beta_T D + \gamma_T DU + U, \tag{2.8}$$

where γ_T is the interaction effect in period T. As for β_0, γ_0 is necessarily zero under the no anticipation assumption. In this case, the DiD approach correctly identifies the ATET as $\beta_1 + \gamma_1 E[U|D = 1]$ and $E[Y_1(0) - Y_0(0)|Y_0, D = d]$, while the common trend is again $E[Y_1(0) - Y_0(0)|Y_0, D = d] = \alpha_1 - \alpha_0$ such that Assumption 6 holds.

Second, we may allow for general functions of D and U in the outcome equation, such as interactions between D and higher order terms of U, or for multiple time constant unobservables, such that U is a vector rather than a scalar. This can be formalized by assuming U to be an unknown function ν of a vector of unobservables $U_1, ...U_K$ where K denotes the number the number of unobservables such that $U = \nu(U_1, ...U_K)$ and assuming the following model:

$$Y_T = \alpha_T + \beta_T(U)D + U, \tag{2.9}$$

where the treatment effect $\beta_T(U)$ is now permitted to be heterogeneous in U if $T = 1$, while $\beta_0(U) = 0$ in the absence of anticipation effects. The ATET corresponds to $E[\beta_T(U)|D = 1]$. Alternatively, we may write (2.9) as

$$Y_T = \alpha_T + \phi_T(D, U), \tag{2.10}$$

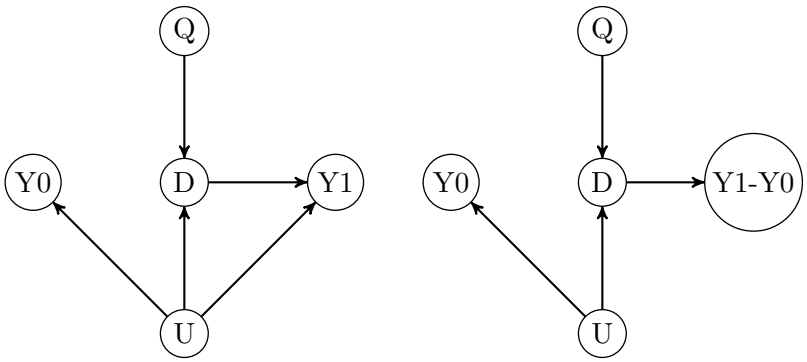
with $\phi_T(D, U) = \beta_T(U)D + U$. This makes explicit that we consider a nonparametric outcome model with an additively separable time trend. Second, the treatment model in (2.7) might be replaced by a nonparametric function of both the time constant unobeservable(s) U and further unobservables, denoted by Q, if the latter are independent of outcome Y_T conditional on U. This implies that conditional on U, there are no time-varying unobserved variables jointly affecting D and Y_T. Formally,

$$D = \psi(U, Q), \tag{2.11}$$

where ψ is an unknown function and Q is independent of Y_T conditional on U.

As an alternative to using equations, we can represent our causal framework by causal graphs as e.g. considered in Pearl (2000), in which we use directed arrows to present average causal effects between variables, which are represented by edges. Figure 1 provides a causal graph which is in line with outcome model (2.9) and treatment model (2.11). In the left graph, the fixed effect U jointly affects the outcomes in either period, while both U and the unobservable Q affect the treatment. Furthermore, as implied by Assumption 3, D is associated with Y_0, because the unobserved term U which influences the outcomes also affects D. For this reason, U confounds the causal association of the treatment and the post-treatment outcome, which motivates the DiD approach. The right graph provides the causal framework when replacing Y_1 by the difference in post- and pre-treatment outcomes $Y_1 - Y_0$. Importantly, the time constant unobservable term U does not affect $Y_1 - Y_0$, as it cancels out due to the additive separability α_T and U in model (2.9), implying that U does not interact with the time trend. Therefore, the average causal association of D and $Y_1 - Y_0$ is unconfounded and DiD identifies the ATET.

Figure 1: Causal framework satisfying Assumption 6



In the previously considered models, the period-specific constant α_T was additively separable and thus, independent of U and D. If this is not the case, Assumption 6 generally fails, as well as Assumptions 1 and 4. Consider for instance a specification where the period-specific constant depends on the fixed effect, e.g. though an interaction of α_T and U:

$$Y_T = \alpha_T U + \beta_T(U)D + U. \tag{2.12}$$

The time trend is now given by $E[Y_1(0) - Y_0(0)|D = d] = (\alpha_1 - \alpha_0) \cdot E[U|D = d]$. In our setup where U assumably affects D, this implies differential trends across treatment groups such that Assumption 1 is violated and DiD does not identify the ATET. Furthermore, $E[Y_1(0) - Y_0(0)|Y_0 = y, D = d] = (\alpha_1 - \alpha_0) \cdot E[U|Y_0 = y, D = d]$ for $d \in \{1, 0\}$ generally differs across pre-

treatment outcome values y, because $Y_0 = (1 + \alpha_0)U$ is a function U. Therefore, Assumptions 1 and 4 are generally violated, as well as the testable implication (2.6).

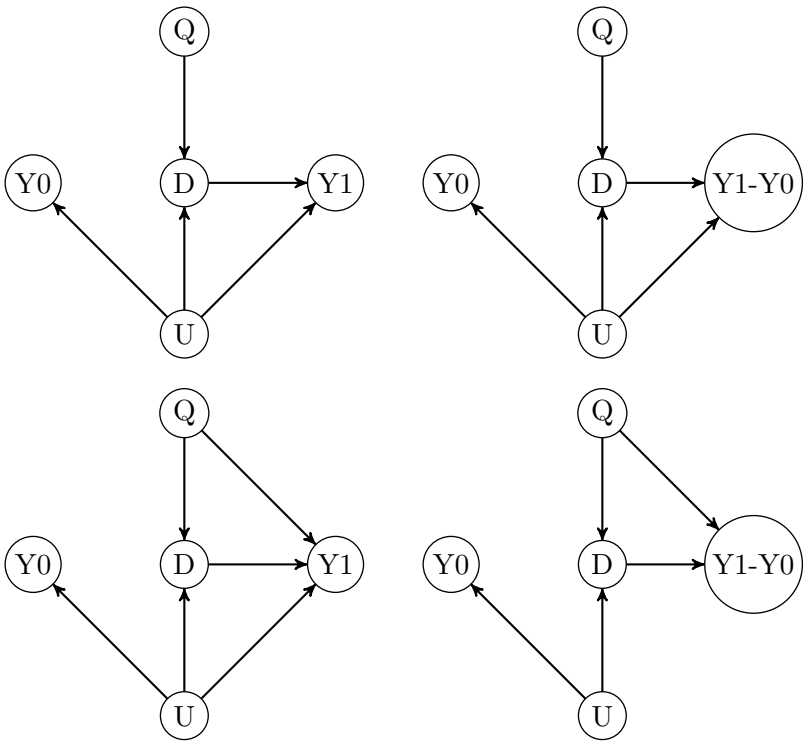
Figure 2 presents causal graphs for scenarios which generally violate the common trend assumptions. In the upper graphs, U affects Y_0 and Y_1 as well as $Y_1 - Y_0$, in line with model 2.12 where the period-specific constant α_T depends on the fixed effect. For this reason, U confounds the causal relation between D and $Y_1 - Y_0$ by jointly affecting D and $Y_1 - Y_0$. In the lower graphs, a time-varying unobservable Q affects both D and Y_1. For instance, sector-specific recessions (Q) might induce retraining activities (D) among certain employees while also affecting wages in future periods (Y_1). In this case, the outcome difference $Y_1 - Y_0$ does not difference out Q, which was not present in the pre-treatment period 0 yet (or took a different value). For this reason Q confounds the causal relation between D and $Y_1 - Y_0$, implying a differential time trend for treated and non-treated groups and a violation of Assumption 6.

In both the scenarios in the upper and lower graphs, it is easy to see that the testable implication (2.6) is violated as Y_0 is associated with $Y_1 - Y_0$ conditional on $D = 0$. In the upper graphs, the correlation is due to U directly affecting both Y_0 and $Y_1 - Y_0$. In the lower graphs, conditioning on $D = 0$ introduces so-called collider bias, i.e. a spurious association between U and Q via D, see e.g. Pearl (2000). Collider bias comes from the fact that both U and Q affect D such that conditioning on $D = 0$ causes U and Q to be associated, which is also known as sample selection bias in economics, see e.g. Heckman (1976). Through this conditional association of U and Q, Y_0 and $Y_1 - Y_0$ are conditionally associated, too, as U affects Y_0 and Q affects $Y_1 - Y_0$.

In our model, a violation of Assumption 6 and implication (2.6) generally also implies the violation of Assumptions 1 and 4 and the testable implication (2.4), apart from special cases. Only under quite specific effect heterogeneities captured by $\beta(U)$ in model (2.12), it may happen that differential trends conditional on Y_0, i.e. $E[Y_1(0) - Y_0(0)|Y_0 = y, D = 0] \neq E[Y_1(0) - Y_0(0)|Y_0 = y, D = 1]$ for some y, average out in particular ways such that Assumptions 4 and 1 as well as implication (2.4) hold. Yet, one might prefer testing this latter condition (rather than the maybe more intuitive condition (2.6) related to the stronger Assumption 6), because it is based on the weakest possible model assumptions for testing identification based on DiD. For binary outcomes, for instance, Assumption 6 is only satisfied if the common trend is equal to zero such that any non-treated outcome is constant over time, which may seem implausible. In contrast, Assumption 4 can hold in admittedly quite specific models with non-zero trends in which Assumption 6 is violated.

Theorem 1 demonstrates that implication (2.6) is satisfied if and only if both Assumptions 1 and 6 among non-treated observations hold, conditional on Assumption 3 and the causal

Figure 2: Causal frameworks not satisfying Assumption 6



faithfulness (or stability) assumption, which rules out special cases of models in which causal mechanisms between variables cancel each other out exactly, see for instance Spirtes et al. (2000) and Pearl (2000). As an example, consider the upper graphs of Figure 2 and assume that the confounding effects due to U directly affecting Y_0 and the collider bias introduced by conditioning on $D = 0$ exactly offset each other. In this case, the implication (2.6) would hold despite a violation of Assumption 1. For this reason, we invoke causal stability, which excludes such special cases of offsetting mechanisms. The proof of Theorem 1 is provided in Appendix A.1.

Theorem 1.

$$
\begin{aligned}
& E[Y_1(0) - Y_0(0)|Y_0] = E[Y_1(0) - Y_0(0)|D = 0] = E[Y_1(0) - Y_0(0)] \\
\iff \quad & E[Y_1 - Y_0|Y_0, D = 0] = E[Y_1 - Y_0|D = 0].
\end{aligned}
$$

Conditional on causal faithfulness (no offsetting mechanisms) and Assumption 3, the testable implication (2.6) is necessary and sufficient for the joint satisfaction of Assumptions 1 and 6.

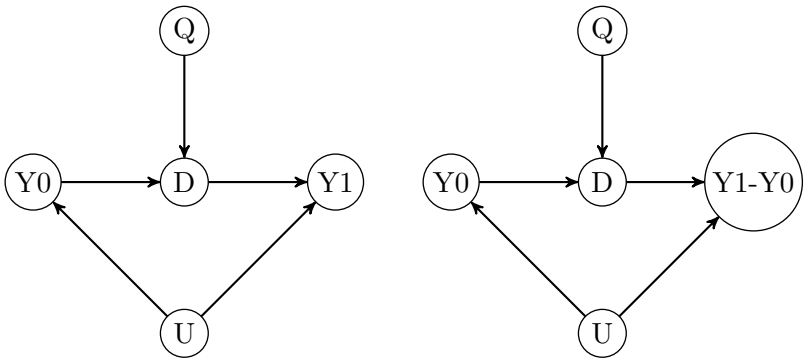
By Theorem 1, the satisfaction of the testable implication (2.6) is sufficient for DiD-based identification, as both Assumptions 1 and 6 hold. However, it is important to note that there also exist further panel models in which DiD-based identification is feasible even under a viola-

tion of implication (2.6) such that the latter is not a necessary condition, as discussed further below. Furthermore, we point out that our test generally cannot detect violations of the common trend assumption if they exclusively concern the potential outcomes under non-treatment of treated subjects, but not of non-treated subjects. Such a situation would arise if there existed distinct potential outcome models for treated and non-treated observations and period-fixed effect-interactions like $\alpha_T U$ in equation (2.12) exclusively occurred among treated, but not among non-treated observations. However, this may not appear very realistic in many empirical applications, because we would suspect a certain overlap in the (unobserved) background characteristics of treated and non-treated groups. This in turn should make it likely that if such period-fixed effect-interactions exist, they should occur in both treatment groups, such that our test should be able to detect such violations of the common trend assumption among the non-treated.

3 Identification based on selection-on-observables

In this section, we explore a case where identification remains feasible despite violations of our testable implications, utilizing a different method than DiD, which entails distinct models and assumptions from those discussed previously. Let us consider a modification of the treatment model provided in equation (2.11), assuming that (i) the pre-treatment outcome Y_0 affects treatment D, and (ii) U influences D solely through Y_0, but not directly. Additional, we assume that the common trend assumption fails such that the time trend interacts with U, as in equation (2.12). This scenario implies the satisfaction of a so-called selection-on-observables assumption, while the conventional common trend assumption stipulated in Assumption 1 is violated.

Figure 3: Causal framework satisfying selection on observables



The causal graph in Figure 3 illustrates such a scenario. Interestingly, even though Assump-

tion 1 does not hold, Assumptions 6 and 4 are satisfied, because $E[U|Y_0 = y, D = 1] = E[U|Y_0 = y, D = 0]$. The latter implies that common trends across treatment states hold conditional on pre-treatment outcomes:

$$E[Y_1(0) - Y_0(0)|Y_0, D = 1] = E[Y_1(0) - Y_0(0)|Y_0, D = 0]. \tag{3.1}$$

By the law of iterated expectations, equation (3.1) is a sufficient condition for Assumption 4. However, whether or not Assumption 4 holds, Assumption 1 is violated such that equation (2.3) and the testable implications (2.4) are violated, correctly suggesting that the DiD approach is inconsistent. This is due to U jointly affecting D and $Y_1 - Y_0$, because U is not additively separable from the time trend.

Nevertheless, it is worth pointing out that the satisfaction of equation (3.1) implies that the treatment is mean independent of potential outcomes in the post-treatment period conditional on the pre-treatment outcome, such that treatment selection is based on an observable, namely Y_0. To see this, note that for some value $Y(0) = y$, equation (3.1) can be rewritten as

$$\begin{aligned}
& E[Y_1(0) - Y_0(0)|Y_0 = y, D = 1] = E[Y_1(0) - Y_0(0)|Y_0 = y, D = 0], \\
= \; & E[Y_1(0) - Y_0(0)|Y_0(0) = y, D = 1] = E[Y_1(0) - Y_0(0)|Y_0(0) = y, D = 0], \\
= \; & E[Y_1(0)|Y_0(0) = y, D = 1] - y = E[Y_1(0)|Y_0(0) = y, D = 0] - y, \\
\Leftrightarrow \; & E[Y_1(0)|Y_0(0) = y, D = 1] = E[Y_1(0)|Y_0(0) = y, D = 0],
\end{aligned} \tag{3.2}$$

where the second line follows from $Y_0 = Y_0(0)$ under Assumption 2. For this reason, controlling for pre-treatment outcomes permits identifying average treatment effects, e.g. based on matching treated and non-treated observations with comparable pre-treatment outcomes, see Imbens (2004) for a review on estimation approaches in this context. However, our testing framework is tailored to testing common trends, and for this reason, it cannot be used to verify the satisfaction of the selection-on-observables assumption, implying that the treatment is as good as random after controlling for Y_0, which underlies this matching approach. The violation of (2.4) points to a violation of Assumption 1, no matter whether (3.2) holds. On the other hand, the satisfaction of (2.4) implies the satisfaction of common trend assumption, but the selection-on-observables assumption may either hold or fail, namely if U affects D directly (contrarily to Figure 3), rather than exclusively through Y_0.

4 Models with time-varying outcome shocks

The models examined thus far have not incorporated time-varying outcome shocks, i.e., time-varying unobservable factors or error terms affecting the outcomes in the same period. It is important to note that introducing such error terms, which is plausible in many causal scenarios, may lead to models where Assumption 1, the common trend assumption required for DiD-based identification, holds, while Assumptions 6 and 4, as well as the testable implications (2.4) and (2.6), are violated. In such instances, our testing approach may incorrectly reject identification based on DiD asymptotically. To illustrate this, let us, for instance, reconsider the parametric models in expression (2.7) (albeit nonparametric models could be extended analogously) and add a time-varying error term to the outcome equation, denoted by V_T, which is (conditionally) mean zero, i.e., $E[V_T|D,U] = E[V_T] = 0$. The linear outcome model now becomes

$$Y_T = \alpha_T + \beta_T D + U + V_T, \tag{4.1}$$

while we maintain the treatment model provided in (2.7).

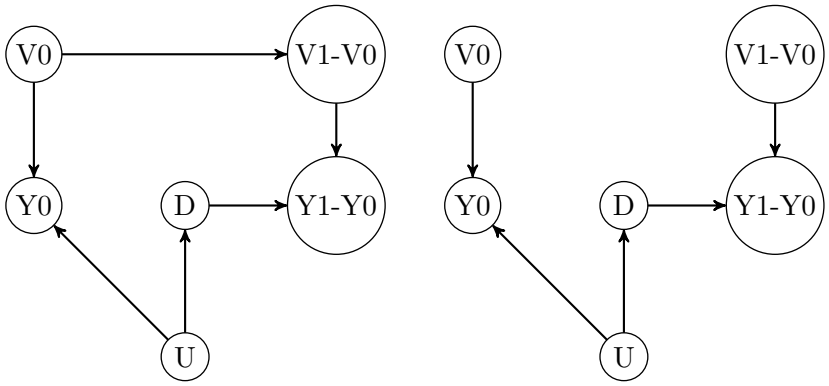
Equation (4.1) corresponds to conventional linear panel regression models as considered in classical fixed effects regression, see for instance Wooldridge (2002). The condition that time-varying unobservables V_T are uncorrelated with D given U, as implied by $E[V_T|D,U] = 0$, is known as strict exogeneity in such models. V_T can be interpreted as conditionally random outcome shock that does not entail confounding of the treatment D and the outcome change $Y_1 - Y_0$. In particular, $E[Y_1(0) - Y_0(0)|D = d] = \alpha_1 - \alpha_0 + E[V_1 - V_0|D = d] = \alpha_1 - \alpha_0$, because $E[V_T|D] = 0$ as V_T neither influences nor is influenced by D according to our model, such that Assumption 1 holds. Furthermore, we have that $E[Y_1(0) - Y_0(0)|Y_0] = \alpha_1 - \alpha_0 + E[V_1 - V_0|Y_0]$. However, $E[V_1 - V_0|Y_0]$ is in general not zero, because the conditioning variable Y_0 is a function of the pre-treatment error V_0, which is generally correlated with the difference $V_1 - V_0$.[2] Therefore, Assumptions 6 and 4 as well the testable implications (2.4) and (2.6) are generally violated. Summing up, Assumption 1, the common trend assumption which is necessary for DiD-based identification, holds, while the sufficient (but not necessary) testable implications are violated.

The left graph in Figure 4 illustrates this scenario. Via the correlation of V_0 and $V_1 - V_0$, Y_0 is correlated with $Y_1 - Y_0$, which is an example for a so-called back-door path between Y_0 and $Y_1 - Y_0$ in the denomination of Pearl (2000). For this reason, Assumption 6 fails, while Assumption 1 holds, because U affects D, but not $Y_1 - Y_0$. See also the discussion in **??**,

[2] For instance, assume the simplistic case that V_1 is a constant. In this case, $V_1 - V_0$ is perfectly negatively correlated with V_0.

which highlights that strict exogeneity is sufficient for the common trend assumption required for ATET identification.

Figure 4: Causal framework satisfying Assumption 1 and violating (left) or satisfying (right) Assumption 6



It is worth mentioning that there exists a special case in which V_0 and $V1 - V0$ are not correlated, namely if V_1 follows a so-called random walk, such that it is strongly serially correlated with V_0:

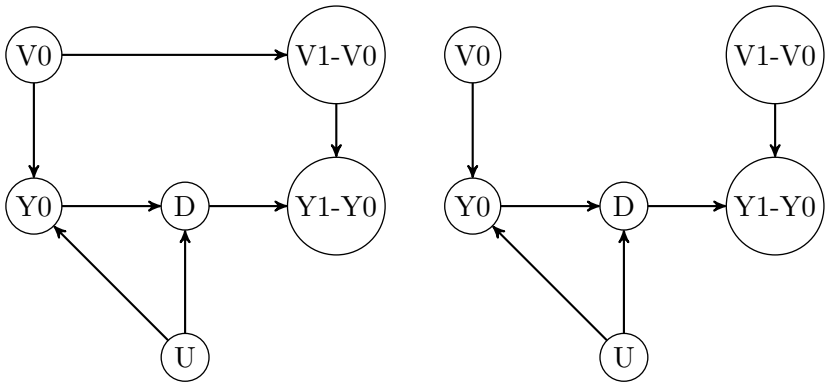
$$V_1 = V_0 + W_1 \Leftrightarrow V_1 - V_0 = W_1, \tag{4.2}$$

where W_1 is an independent unobservable (or random shock) in period 1, $E[W_1|V_0] = E[W_1] = 0$, implying that $E[V_1|V_0] = V_0$. Under a random walk, $V_1 - V_0 = W_1$ is not correlated with error V_0, such that there is no correlation of Y_0 and $Y_1 - Y_0$. This is illustrated in the right graph of Figure 4. However, we emphasize that apart from a random walk, the presence of time-varying unobservables in the outcome equation implies a violation of our testable implications, even if Assumption 1 holds.

In this context with time-varying unobserved variables V_T, it is interesting to look at the implications of our test under a modification of our model that permits for a non-zero impact of the pre-treatment outcome Y_0 on treatment D, as previously considered in Figure 3. Consequently, we modify the causal models in Figure 4 by also adding a direct effect of Y_0 on D, as provided in Figure 5 and also considered in Chabé-Ferret (2017). This constitutes a violation of strict exogeneity, with its implications more thoroughly discussed in Imai and Kim (2019), as V_0 affects D via Y_0. In this causal model, testing Assumption 6 is well aligned with verifying the satisfaction of Assumption 1. In the left graph, Assumption 6 is violated for the very same reasons as discussed for the left graph of Figure 4. However, also Assumption 1 is violated, as

V_0 jointly affects the outcome change $Y_1 - Y_0$ via $V_1 - V_0$ and the treatment D via Y_0. This back-door path between the treatment and the outcome change implies a failure of the common trend assumption across treatment groups such that DiD does not identify the ATET, as also discussed in Ghanem et al. (2022). Such a back-door path does, however, not exist in the case that $V_1 - V_0$ is not correlated V_1, i.e. under a random walk, as provided in the right graph of Figure 5, where the pre-treatment error V_0 is not associated with the outcome change $Y_1 - Y_0$. For this reason, both Assumptions 1 and 6 are satisfied. In summary, in scenarios where strict exogeneity, which precludes effects of past outcomes or past time-varying unobservables associated with the outcomes on the treatment, appears doubtful or implausible, our test provides valuable guidance for DiD-based identification.

Figure 5: Causal framework violating (left) or satisfying (right) Assumptions 1 and 6



5 Common trends conditional on covariates

In many applications, common trends might only appear plausible when comparing treated and non-treated observations that are similar in terms of observed covariates, henceforth denoted by X, as for instance discussed in Abadie (2005). When e.g. assessing the effect of a minimum wage on earnings, the common trend assumption may only seem realistic for the earnings of treated and non-treated individuals that are employed in the same industry, as earnings trends may differ across industries. This implies that the common trend assumption does not hold unconditionally across regions introducing and not introducing a minimum wage, if the region differ in terms of industry structure. However, by controlling for covariates through adjusting their distribution in the non-treated group to match that of the treated group, the common trend assumption 1 is assumed to hold conditional on X, as formalized in Assumption 7:

Assumption 7 (Common trends across treatment groups conditional on covariates)**.**

$$E[Y_1(0) - Y_0(0)|X, D = 1] = E[Y_1(0) - Y_0(0)|X, D = 0] = E[Y_1(0) - Y_0(0)|X].$$

Furthermore, let us assume that also Assumptions 2, 3, 5 hold conditional on covariates, as postulated in Assumptions 8 to 10.

Assumption 8 (No anticipation conditional on covariates)**.**

$$\Pr(Y_0(1) = Y_0(0)|X) = 0.$$

Assumption 9 (Treatment selection bias conditional on covariates)**.**

$$E[Y_0(0)|D = 1] \neq E[Y_0(0)|D = 0].$$

Assumption 10 (Common support conditional on covariates)**.**

$$\Pr(D = 1|X, Y_0) < 1.$$

Under Assumptions 7, 8, and 10 (or the slightly weaker common support restriction $\Pr(D = 1|X) < 1$), the ATET is identified in analogy to (5.1) by

$$\begin{aligned}
\Delta_{D=1} &= E[Y_1(1)|D = 1] - E[Y_1(0)|D = 1] \\
&= E[Y_1(1)|D = 1] - E[Y_0(0)|D = 1] - E[E[Y_1(0)|X, D = 1] + E[Y_0(0)|X, D = 1]|D = 1] \\
&= E[Y_1|D = 1] - E[Y_0|D = 1] - \{E[E[Y_1|X, D = 0] - E[Y_0|X, D = 0]|D = 1]\}, \qquad (5.1)
\end{aligned}$$

where the second quality follows from the law of iterated expectations.

Furthermore, when controlling for X, Assumption 4 becomes

Assumption 11 (Common trends across matched groups conditional on covariates)**.**

$$E[E[Y_1(0) - Y_0(0)|Y_0, X, D = 0]|X, D = 1] = E[Y_1(0) - Y_0(0)|X, D = 1].$$

In analogy to equation (2.4), the satisfaction of both Assumptions 7 and 11 implies the testable implication

$$E[E[Y_1 - Y_0|Y_0, X, D = 0]|X, D = 1] = E[Y_1 - Y_0|X, D = 0], \tag{5.2}$$

given that the common support assumption 10 holds. By averaging the conditional means in equation (5.2) over the distribution of X in the treated group, which is required for the identification of the ATET in (5.1), we obtain a further testable implication:

$$E[E[Y_1 - Y_0|Y_0, X, D = 0]|D = 1] = E[E[Y_1 - Y_0|X, D = 0]|D = 1], \tag{5.3}$$

which follows from $E[E[E[Y_1 - Y_0|Y_0, X, D = 0]|X, D = 1]|D = 1] = E[E[Y_1 - Y_0|Y_0, X, D = 0]|D = 1]$ by the law of iterated expectations.

Conditional on X, Assumption 6 changes to

Assumption 12 (Common trends across pre-treatment outcomes conditional on covariates)**.**

$$E[Y_1(0) - Y_0(0)|Y_0, X] = E[Y_1(0) - Y_0(0)|X].$$

Under Assumptions 7 and 12, it holds in analogy to implication (2.6) that

$$E[Y_1 - Y_0|Y_0, X, D = 0] = E[Y_1 - Y_0|X, D = 0]. \tag{5.4}$$

Finally, averaging the conditional means in equation (5.4) over the distribution of X conditional on $D = 1$ yields once again the previously mentioned testable implication (5.3).

6 Testing

This section suggests a machine learning-based method for testing implication (5.3) to jointly verify the conditional common trend assumptions 7 and 11, but also briefly sketches possible approaches for testing implication (5.4) when replacing Assumption 11 by the stronger Assumption 12. Furthermore, it briefly discusses methods for testing implication (2.4) to jointly verify common trend assumptions 1 and 4, which do not involve covariates X, as well as implication (2.6) when replacing Assumption 4 by the stronger Assumption 6. Starting with the latter scenario, let us denote the conditional mean difference in outcomes given Y_0 and $D = 0$

as $\mu(y_0) = E[Y_1 - Y_0|Y_0 = y_0, D = 0]$. Implication (2.6) is equivalent to the following null hypothesis H_0:

$$H_0 : \mu(y_0) - \mu(y_0') = 0 \quad \forall y_0 \neq y_0' \text{ in the support of } Y_0, \tag{6.1}$$

If the outcome Y_t is binary, equation (6.1) can be easily tested by an OLS regression of $Y_1 - Y_0$ on the binary Y_0 and a constant in the sample of non-treated observations and an inspection of the p-value of the coefficient on Y_0. If the outcome Y_t is non-binary but discrete, one may regress $Y_1 - Y_0$ on a constant and indicator variables for the discrete values of Y_0, with only one value not being modeled by an indicator to avoid perfect multicollinearity. Running an F-test for the joint coefficients of the indicator variables yields a test of the null hypothesis (6.1). If the outcome is continuously distributed, the null hypothesis may e.g. be tested by running a univariate nonparametric regression of $Y_1 - Y_0$ on Y_0 in the sample of non-treated observations and test whether the first derivatives w.r.t. Y_0 are zero for all observations, see for instance the kernel regression-based statistical significance test suggested by Racine (1997).

When assuming that common trends hold conditional on covariates X as postulated in Assumptions 7 and 12, the null hypothesis to be tested becomes

$$H_0 : \mu(y_0, x) - \mu(y_0', x) = 0 \quad \forall y_0 \neq y_0' \text{ in the support of } Y_0 \text{ and } x \text{ in the support of } X, \tag{6.2}$$

with $\mu(y_0, x) = E[Y_1 - Y_0|Y_0 = y_0, X = x, D = 0]$ denoting the conditional mean difference in outcomes given Y_0, X, and $D = 0$. This null hypothesis may e.g. be verified by tests that investigate whether the mean squared error increases significantly when dropping Y_0 from the outcome regression model $\mu(Y_0, X)$, see e.g. the approaches suggested in Huber and Kueck (2022) and Kook and Lundborg (2024).

Next, let us consider testing implication (2.4) for jointly verifying Assumptions 1 and 4. In this case, the null hypothesis is given by

$$H_0 : E[\mu(Y_0)|D = 1] - E[Y_1 - Y_0|D = 0] = 0, \tag{6.3}$$

We may estimate the two conditional means in equation (6.3) by their sample analogs conditional of $D = 1$ and $D = 0$, respectively, and $\mu(Y_0)$ by regression (as discussed before) in order to construct a test for a violation of the null, typically based on the t-statistic.

Finally, when assuming that such common trend assumptions only hold conditional on X,

we consider the testable implication (5.3) for jointly verifying Assumptions 7 and 11 based on the following null hypothesis:

$$H_0 : E[\mu(Y_0, X)|D = 1] - E[\mu(X)|D = 1] = 0, \tag{6.4}$$

with $\mu(y_0, x) = E[Y_1 - Y_0|Y_0 = y_0, X = x, D = 0]$ and $\mu(x) = E[Y_1 - Y_0|X = x, D = 0]$. Using insights from doubly robust (DR) statistics as for instance considered in Robins et al. (1994) and Robins and Rotnitzky (1995), it follows that (6.4) is equivalent to

$$\begin{aligned} H_0 : \theta &= 0, \text{ where} \\ \theta &= E\left[\frac{\mu(Y_0, X) \cdot D}{\Pr(D = 1)} + \frac{(Y_1 - Y_0 - \mu(Y_0, X)) \cdot (1 - D) \cdot p(Y_0, X)}{(1 - p(Y_0, X) \cdot \Pr(D = 1))}\right] \\ &- E\left[\frac{\mu(X) \cdot D}{\Pr(D = 1)} + \frac{(Y_1 - Y_0 - \mu(X)) \cdot (1 - D) \cdot p(X)}{(1 - p(X) \cdot \Pr(D = 1))}\right], \end{aligned} \tag{6.5}$$

with $p(X) = \Pr(D = 1|X)$ and $p(Y_0, X) = \Pr(D = 1|Y_0, X)$ denoting the conditional treatment probabilities, also known as propensity scores.[3] We may estimate the expectations in expression (6.5) by their respective sample analogs and the plug-in (or nuisance) parameters $\mu(Y_0, X)$, $p(Y_0, X)$, $\mu(X)$, and $p(X)$ by machine learning (ML) algorithms, in order to control for important confounders in a data-driven way, which appears particularly attractive if X is high-dimensional. This approach is known as double machine learning (DML); see Chernozhukov et al. (2018) for a seminal discussion, as well as Chang (2020) and Zimmert (2020) for DiD-based DML estimators.

DML is conventionally combined with so-called cross-fitting, which consists of estimating the model parameters of the plug-in parameters (like regression coefficients) and the mean difference θ in expression (6.5) in non-overlapping subsets of the data, with the roles of the subsets for the estimation steps being sequentially swapped. As no observation enters both estimation steps at the same time, cross-fitting avoids correlations between the estimation of the model parameters of $\mu(Y_0, X)$, $p(Y_0, X)$, $\mu(X)$, and $p(X)$ on the one hand and of the mean difference θ on the other hand and thus, overfitting bias. Averaging over the mean differences obtained from the various subsets of the data yields the final estimate for testing the null hypothesis in expression (6.5). Under specific regularity conditions discussed in Chernozhukov et al. (2018), in particular if the ML-based estimators of the plug-in parameters converge to the respective true models

[3] To see the equivalence of the expressions in (6.4) and (6.5), consider the first expectation in either expression. Note that by basic probability theory $E\left[\frac{\mu(Y_0,X)\cdot D}{\Pr(D=1)}\right] = E[\mu(Y_0, X)|D = 1]$, while by the law of iterated expectations $E\left[\frac{(Y-\mu(Y_0,X))\cdot(1-D)\cdot p(Y_0,X)}{(1-p(Y_0,X)\cdot\Pr(D=1))}\right] = 0$ because $E[Y - \mu(Y_0, X)|Y_0, X, D = 0] = 0$. An analogous result can be shown for the equivalence of the second expectation in expressions (6.3) and (6.4), respectively.

with a rate of at least $n^{1/4}$, DML estimation of the mean differences in expression (6.5) is root-n-consistent and asymptotically normal. Furthermore, the results in Chernozhukov et al. (2018) imply that asymptotic variance of the estimator of θ corresponds to σ^2/n where n is the sample size and σ^2 is the variance of the difference of the terms in the expectations of expression (6.5):

$$\sigma^2 = Var\left(\frac{\mu(Y_0,X)\cdot D}{\Pr(D=1)} + \frac{(Y_1-Y_0-\mu(Y_0,X))\cdot(1-D)\cdot p(Y_0,X)}{(1-p(Y_0,X)\cdot\Pr(D=1))} - \frac{\mu(X)\cdot D}{\Pr(D=1)} - \frac{(Y_1-Y_0-\mu(X))\cdot(1-D)\cdot p(X)}{(1-p(X)\cdot\Pr(D=1))}\right). \tag{6.6}$$

Also the variance σ^2 can be consistently estimated under specific regularity conditions.

7 Simulation study

This section provides a simulation study to investigate the finite sample behavior of our testing approach based on the following data generating process:

$$\begin{aligned} Y_0 &= U + \beta_{V_0} V_0, \quad D = I\{X'\beta_X + U + Q > 0\}, \\ Y_1 &= 1 + D + X'\beta_X + (1+\beta_U)U - \beta_Q Q + V_1, \\ X &\sim N(0,\sigma_X^2), V_0 \sim N(0,1), V_1 \sim N(0,1), U \sim N(0,1), Q \sim N(0,1), \\ &\quad \text{with } X, V_0, V_1, U, Q \text{ being independent of each other.} \end{aligned}$$

The pre-treatment outcome Y_0 is a function of the fixed effect U and the period-specific unobservable V_0 if the coefficient $\beta_{V_0} \neq 0$, while the time-varying constant α_0 is equal to zero. The binary treatment D is a function of the fixed effect U, a further unobservable Q, and covariates X for coefficient vector $\beta_X \neq 0$. The post-treatment outcome Y_1 is a linear function of the time-varying constant $\alpha_1 = 1$, treatment D, whose effect equals 1, covariates X for $\beta_X \neq 0$, and the time-varying unobservable V_1. Y_1 is also a function of unobservable Q if the coefficient $\beta_Q \neq 0$, as well as of the fixed effect U. We note that for $\beta_U \neq 0$, the outcome trends are heterogeneous in U such that the common trend assumption is violated, because U is not netted out when subtracting Y_0 from Y_1 (even conditional on D and X).

The covariate vector X has a multivariate normal distribution with a zero mean and a covariance matrix σ_X^2 that is obtained by setting the covariance of the ith and jth covariate in X to $0.5^{|i-j|}$. Furthermore, U and Q are standard normally distributed and independent of each other, as well as of X. β_X gauges the effects of the covariates on Y_1 and $Y_1 - Y_0$ alike (because

X does not affect Y_0) as well as on D, respectively, and thus, the magnitude of confounding due to observables. The ith element in the coefficient vector β_X is set to $0.7/i$ for $i = 1, ..., p$, with p denoting the number of covariates. This implies a linear decay of covariate importance in terms of confounding of D and the average trend of the potential outcomes under non-treatment, $Y_1(0) - Y_0(0)$.

We consider testing whether the estimate of the DR-based mean difference θ in expression (6.5) is statistically different from zero, which may point to a violation of the common trends assumption. We estimate this mean difference based on DML as discussed in Section 6, using two subsets (or folds) of the data for cross-fitting. We estimate the model specifications of the plug-in parameters by lasso regression (linear regression for estimating $\mu(Y_0, X)$ and $\mu(X)$ and logit regression for estimating $p(Y_0, X)$ and $p(X)$) with cross-validated penalty terms, see Tibshirani (1996). Observations with propensity score estimates of $p(Y_0, X)$ or $p(X)$ that are very close to one, namely larger than 0.99 (or 99%), are dropped from the estimation in order to avoid an explosion of the propensity score-based weights. Furthermore, the propensity score estimates of the remaining observations are normalized such that they add up to one in the estimation of either mean in expression (6.5), see for instance Knaus (2021) for a discussion of normalized DML. We analyse the performance of our testing approach in 1000 simulations and with two sample sizes of $n = 1000$ and 4000 observations of $\{Y_0, Y_1, D, X\}$ when setting the number of covariates to $p = 200$. In addition to the test, we also estimate the ATET provided in expression (5.1) by normalized DML with 2-fold cross-validation and lasso regression for estimating the plug-in parameters, in order to investigate the bias of DiD-based estimation in the various simulations. This estimator corresponds to a standard DR procedure for ATET evaluation conditional on covariates X, see e.g. Chernozhukov et al. (2018), with the difference that we consider $Y_1 - Y_0$ rather than Y_1 as outcome variable, due to our DiD framework with panel data.

In our first simulation study, we set $\beta_U = 0$ (no trend heterogeneity in the fixed effect), $\beta_Q = 0$ (no confounding due to Q), and $\beta_{V_0} = 0$ (implying that $V_1 - V_0 = V_1$, which is uncorrelated with Y_0). In this setting, all conditional common trend assumptions given X are satisfied, namely Assumptions 7, 11, and 12. The upper panel of Table 1 provides the results. In line with the satisfaction of common trends, the average estimate of the DR-based difference θ in expression (6.5) ('mean est') is close to zero for either samples size across the 1000 simulations. Furthermore, the estimator seems to converge at $\sqrt{n}$-rate, as the standard deviation ('std') is roughly reduced by half when quadrupling the sample size from 1000 to 4000. We also see that at least for the larger sample size, the average standard error of the estimator ('mean se') is

similar to the actual standard deviation. The average p-value ('mean pval') of rejecting the null hypothesis of of $\theta = 0$ based on a t-test amounts to 42 and 47% for the smaller and larger sample size, respectively. Furthermore, the empirical rejection rate, i.e., the share of p-values in the simulations that are at most 0.05 (or 5%), corresponds to 10 and 6% under the smaller and larger sample size, respectively. This impliex a decent empirical size for $n = 4000$. Finally, both the bias ('bias') and the root mean squared error ('rmse') of the DML-based DiD estimator of the ATET are moderate, as common trends hold conditional on covariates, and decrease in the sample size.

In a next step, we set $\beta_U = 0.5$ while maintaining $\beta_Q = \beta_{V_0} = 0$, such that the trend in the mean potential outcomes is heterogeneous in the fixed effect U and Assumptions 7, 11, and 12 are violated. The second upper panel of Table 1 provides the results for this scenario. The average estimate of θ now amounts to roughly 0.49 under the smaller and 0.55 under the larger sample size, respectively. Furthermore, the average p-values are close to zero and the rejection rates are 99% and 100% for $n = 1000$ and $n = 4000$, respectively. We also notice that the DML-based DiD estimator of the ATET is considerably biased under either sample size. Next, we set $\beta_Q = 0.5$ and $\beta_U = \beta_{V_0} = 0$ such that Q confounds the treatment and the trend of potential outcomes, again entailing a violation of the common trend assumptions. The results are presented in the third panel of Table 1. Under the lower sample size, the rejection rate of the test reaches 39% with an average p-value of 18%, while it amounts to 99% under the larger sample size, with an average p-value close to zero. also in this scenario, the DiD estimator of the ATET is prone to non-negligible bias.

Finally, we examine a data generating process where Assumption 7, sufficient for DiD-based identification conditional on X, is met, but Assumptions 11 and 12 are violated. To this end, we set $\beta_{V_0} = 0.5$ and $\beta_U = \beta_Q = 0$, resulting in V_0 affecting Y_0, and the difference $V_1 - V_0$ being correlated with V_0 and thus with Y_0. This mirrors the scenario depicted in the left graph of Figure 4 when keeping X fixed. As shown in the bottom panel of Table 1, the rejection rates of the test are 92% and 100% for $n = 1000$ and $n = 4000$, respectively. However, the bias of the DiD estimator of the ATET remains moderate and decreases with sample size, because Assumption 7 holds.

8 Empirical application

This section presents an empirical application using labor market data from LaLonde (1986) concerning the National Supported Work Demonstration (NSW), a U.S. federally funded pro-

Table 1: Simulations satisfying or violating common trends

sample size	mean est	std	mean se	mean pval	rejection	bias	rmse
	$\beta_{V_0}=0$, $\beta_U=0$, $\beta_Q=0$: Assumptions 7, 11, and 12 hold						
1000	-0.03	0.10	0.08	0.42	0.10	0.14	0.18
4000	-0.01	0.06	0.05	0.47	0.06	0.04	0.06
	$\beta_{V_0}=0$, $\beta_U=0.5$, $\beta_Q=0$: Assumptions 7, 11, and 12 violated						
1000	0.49	0.12	0.09	0.00	0.99	0.73	0.73
4000	0.55	0.06	0.05	0.00	1.00	0.61	0.62
	$\beta_{V_0}=0$, $\beta_U=0$, $\beta_Q=0.5$: Assumptions 7, 11, and 12 violated						
1000	0.17	0.11	0.10	0.18	0.39	-0.45	0.46
4000	0.26	0.06	0.06	0.00	0.99	-0.57	0.57
	$\beta_{V_0}=0.5$, $\beta_U=0$, $\beta_Q=0$: Assumption 7 holds, Assumptions 11 and 12 violated						
1000	-0.30	0.10	0.08	0.03	0.92	0.16	0.20
4000	-0.32	0.05	0.05	0.00	1.00	0.06	0.08

Notes: columns 'mean est', 'std', and 'mean se' provide the average estimate of Δ, its standard deviation, and the average of the estimated standard error across simulations, respectively. 'rejection pval' and 'rejection rate' give the average p-value of the test and the empirical rejection rate when setting the level of statistical significance to 0.05 (or 5%), respectively. 'bias' yields the bias of DML-based DiD estimation of the ATET.

gram in the 1970s aimed at enhancing the employment prospects of disadvantaged individuals through subsidized employment and job training. The dataset comprises both an experimental sample, where the program was randomly assigned, and observational data, where treated subjects from the experiment are combined with non-treated subjects from the Panel Study of Income Dynamics (PSID). Researchers such as LaLonde (1986), Dehejia and Wahba (1999), and Dehejia and Wahba (2002) analyzed these data to explore how closely non-experimental estimators, based on the selection-on-observables assumption, approximate the effect estimate in the experimental data. Additionally, Smith and Todd (2005) consider DiD-based approaches and generally found them to be closer to the experimental estimate compared to competing methods (such as matching) based on the selection-on-observables assumption.[4]

Here, we test whether the estimate of the DR-based mean difference θ in expression (6.5) is statistically different from zero in the PSID sample analyzed in Dehejia and Wahba (1999), which represent a subsample of the NSW data of LaLonde (1986) and is available in the *causalsense* package by Blackwell (2018) for R. The pre-and post-treatment outcomes Y_0, Y_1 are real earnings in US dollars in 1975 and 1978, respectively, while the treatment D is a binary indicator for program assignment. The covariates X include age, years of education as well as binary indicators for not possessing a high school diploma, ethnicity, marital status, and being unemployed in the pre-treatment year 1975. To increase model flexibility, we also include

[4]However, we point out that one should interpret these findings cautiously, as the experimental estimate itself is subject to non-negligible uncertainty or variance due to the limited experimental sample size.

interaction terms between each pair of covariates and squared terms for the non-binary covariates age and education, entailing all in all 32 variables in X. As for the simulations discussed in Section 7, estimation is based on DML with 2-fold cross-fitting and lasso regression for the estimation of the plug-in parameters, where we drop (or trim) observations with propensity score estimates of $p(Y_0, X)$ or $p(X)$ that are larger than 0.99 (or 99%).

Table 2 reports the results for the PSID sample, namely the estimate of the test statistic θ ('test stat'), its standard error ('se'), and p-value ('pval'). Additionally, we report the DiD-based ATET estimate from DML with 2-fold cross-fitting, lasso regression, and trimming ('ATET') along with its standard error and p-value. The test statistic amounts to 990.56 US dollars with a standard error of 310.58. Consequently, the p-value is close to zero, implying the rejection of the null hypothesis of common trends as postulated in equation (6.5) at any conventional level of statistical significance. The DiD-based ATET estimate is 2430.84 dollars, which is somewhat larger than the experimental estimate of 1794.30 (with a heteroscedasticity-robust standard error of 672.68) that is based on the treated and non-treated observations in the experimental sample.

Table 2: Empirical application

test stat	se	pval	ATET	se	pval
990.56	310.58	0.00	2430.84	1448.34	0.01

Notes: columns 'test stat', 'se', and 'pval' are the estimate of θ in expression (6.5), its standard error, and its p-value, respectively. Columns 'ATET', 'se', and 'pval' are the DiD-based estimate of the ATET, its standard error, and its p-value, respectively.

9 Conclusion

We suggested a placebo test for the DiD-related common trend assumption in panel data with the same units in both the pre- and post-treatment periods that does not rely on multiple pre-treatment periods. The testing approach is based on constructing a control group of non-treated observations that resembles the treatment group in terms of the distribution of pre-treatment outcomes and verifying whether this matched control group and the total (i.e. non-matched) non-treated group share the same time trend in terms of their average outcomes (under non-treatment). If the matched control and total non-treated groups share the same average outcome trend, this suggests that differences in the pre-treatment outcome distributions of treated and non-treated groups do not imply differences in the time trends of the mean potential outcome under non-treatment. However, we point out that there exist specific models in which the

implication that we test is violated, even though the common trend assumption required for DiD-based identification is satisfied. With this regard, our test verifies a sufficient condition for identification, which is, however, stronger than necessary. We also investigated the finite sample performance of our DML-based testing procedure in a simulation study and found it to perform very satisfactorily in the simulation designs considered. Finally, we provided an empirical illustration utilizing labor market data from the National Supported Work Demonstration.

A Appendix

A.1 Proof of Theorem 1

We subsequently provide a proof for Theorem 1. The latter states that conditional on causal faithfulness and Assumption 3, $E[Y_1 - Y_0|Y_0, D = 0] = E[Y_1 - Y_0|D = 0]$ is a necessary and sufficient condition for $E[Y_1(0) - Y_0(0)|Y_0] = E[Y_1(0) - Y_0(0)|D = 0] = E[Y_1(0) - Y_0(0)]$ such that both Assumptions 1 and 6 hold.

We show this result based on a proof by contradiction. To this end, let us assume that in addition to causal faithfulness and Assumption 3, also Assumptions 1 and 6 hold, while $E[Y_1(0) - Y_0(0)|Y_0, D = 0] \neq E[Y_1(0) - Y_0(0)|D = 0]$. The latter inequality necessarily implies that D is a collider between Y_0 and the time trend $Y_1(0) - Y_0(0)$ in the sense of Pearl (2000), meaning that controlling for D introduces a non-zero correlation between Y_0 and $Y_1(0) - Y_0(0)$. To see this, first note that by Assumptions 1 and 6, Y_0 is mean independent of $Y_1(0) - Y_0(0)$, while by Assumption 3, Y_0 is correlated with D. Under causal faithfulness, the correlation of Y_0 and D may only come from a non-zero average causal effect of Y_0 on D and/or from unobserved variables affecting both Y_0 and D. If conditioning on D introduces a correlation between Y_0 and $Y_1(0) - Y_0(0)$, then there must necessarily exist a confounder affecting both D and $Y_1(0) - Y_0(0)$. This follows from a combination of the so-called d-separation criterion of Pearl (1988),[5] which implies that D is a collider if either a confounder affects both D and $Y_1(0) - Y_0(0)$ or $Y_1(0) - Y_0(0)$ has a non-zero average effect on D. However, the latter possibility is infeasible, because the outcome Y_1 is measured in the post-treatment period and can thus not affect the (earlier) treatment D, such that (reverse) causality from Y_1 to D is ruled out. The remaining possibility of confounders affecting both D and $Y_1(0) - Y_0(0)$ violates (and thus, contradicts) Assumption 1. This in turn implies that if Assumption 1 holds such that D is not a collider and controlling on D does not introduce a correlation between $Y_1(0) - Y_0(0)$ and Y_0, then $E[Y_1 - Y_0|Y_0, D = 0] \neq E[Y_1 - Y_0|D = 0]$ must be due to a correlation of Y_0 and $Y_1(0) - Y_0(0)$. The latter, however, violates (and thus, contradicts) Assumption 6.

We have demonstrated that the joint satisfaction of causal faithfulness as well as Assumptions 1, 3, and 6 necessarily implies $E[Y_1 - Y_0|Y_0, D = 0] = E[Y_1 - Y_0|D = 0]$. Next, we also show the converse, namely that conditional on causal faithfulness and Assumption 3, $E[Y_1 - Y_0|Y_0, D = 0] = E[Y_1 - Y_0|D = 0]$ necessarily implies the joint satisfaction of Assumptions 1 and 6. In our proof by contradiction, we assume that $E[Y_1 - Y_0|Y_0, D = 0] = E[Y_1 - Y_0|D = 0]$ holds while Assumptions 1 and/or 6 are violated. The latter violations imply confounders of D and $Y_1(0) - Y_0(0)$ and/or a correlation of Y_0 and $Y_1(0) - Y_0(0)$. For the reasons discussed before, any of these violations imply $E[Y_1 - Y_0|Y_0, D = 0] \neq E[Y_1 - Y_0|D = 0]$

[5] Formally, the d-separation criterion implies that two (sets of) variables A and B are d-separated when conditioning on a (set of) control variable(s) C if and only if (i) the path between A and B contains a triple $A \rightarrow M \rightarrow B$ (causal chain) or $A \leftarrow M \rightarrow B$ (confounding) such that variable (set) M is in C (i.e. controlled for), (ii) the path between A and B contains a collider $A \rightarrow S \leftarrow B$ such that variable (set) S or any variable (set) causally affected by S is not in C (i.e. not controlled for).

and thus create a contradiction. Therefore, it follows that

$$
\begin{aligned}
& E[Y_1(0) - Y_0(0)|Y_0] = E[Y_1(0) - Y_0(0)|D = 0] = E[Y_1(0) - Y_0(0)] \qquad \text{(A.1)} \\
\iff \quad & E[Y_1 - Y_0|Y_0, D = 0] = E[Y_1 - Y_0|D = 0],
\end{aligned}
$$

such that $E[Y_1 - Y_0|Y_0, D = 0] = E[Y_1 - Y_0|D = 0]$ is necessary and sufficient for Assumption 1 and 6, conditional on causal faithfulness and Assumption 3.

References

Abadie, A., 2005. Semiparametric difference-in-differences estimators. Review of Economic Studies 72, 1–19.

Ashenfelter, O., 1978. Estimating the effect of training programms on earnings. The Review of Economics and Statistics 6, 47–57.

Blackwell, M., 2018. causalsens: Selection Bias Approach to Sensitivity Analysis for Causal Effects. URL: https://CRAN.R-project.org/package=causalsens. r package version 0.1.2.

Chabé-Ferret, S., 2017. Should we combine difference in differences with conditioning on pre-treatment outcomes. working paper, Toulouse School of Economics .

Chang, N.C., 2020. Double/debiased machine learning for difference-in-differences models. The Econometrics Journal 23, 177–191.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal 21, C1–C68.

Cox, D., 1958. Planning of Experiments. Wiley, New York.

Dehejia, R.H., Wahba, S., 1999. Causal effects in non-experimental studies: Reevaluating the evaluation of training programmes. Journal of American Statistical Association 94, 1053–1062.

Dehejia, R.H., Wahba, S., 2002. Propensity-score-matching methods for nonexperimental causal studies. The Review of Economics and Statistics 84, 151–161.

Ghanem, D., Sant'Anna, P.H., Wüthrich, K., 2022. Selection and parallel trends. arXiv preprint 2203.09001 .

Heckman, J.J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5, 475–492.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663–685.

Huber, M., Kueck, J., 2022. Testing the identification of causal effects in data. arXiv preprint 2203.15890 .

Imai, K., Kim, I.S., 2019. When should we use unit fixed effects regression models for causal inference with longitudinal data? American Journal of Political Science 63, 467–490.

Imbens, G.W., 2004. Nonparametric estimation of average treatment effects under exogeneity: a review. The Review of Economics and Statistics 86, 4–29.

Knaus, M.C., 2021. Double machine learning based program evaluation under unconfoundedness arXiv:arXiv preprint 2003.03191.

Kook, L., Lundborg, A.R., 2024. Algorithm-agnostic significance testing in supervised learning with multimodal data. arXiv preprint 2402.14416 .

LaLonde, R., 1986. Evaluating the econometric evaluations of training programs with experimental data. American Economic Review 76, 604–620.

Lechner, M., 2011. The estimation of causal effects by difference-in-difference methods. Foundations and Trends in Econometrics 4, 165–224.

Neyman, J., 1923. On the application of probability theory to agricultural experiments. essay on principles. Statistical Science Reprint, 5, 463–480.

Pearl, J., 1988. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo.

Pearl, J., 2000. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge.

Racine, J., 1997. Consistent significance testing for nonparametric regression. Journal of Business & Economic Statistics 15, 369–378.

Robins, J.M., Rotnitzky, A., 1995. Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association 90, 122–129.

Robins, J.M., Rotnitzky, A., Zhao, L., 1994. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 90, 846–866.

Robins, J.M., Rotnitzky, A., Zhao, L., 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 90, 106–121.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rosenbaum, P.R., Rubin, D.B., 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician 39, 33–38.

Roth, J., Sant'Anna, P.H.C., Bilinski, A., Poe, J., 2022. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. arXiv preprint 2201.01194 .

Rubin, D., 1980. Comment on 'randomization analysis of experimental data: The fisher randomization test' by d. basu. Journal of American Statistical Association 75, 591–593.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701.

Smith, J., Todd, P., 2005. Does matching overcome lalonde's critique of nonexperimental estimators? Journal of Econometrics 125, 305–353.

Snow, J., 1855. On the Mode of Communication of Cholera.

Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D., 2000. Causation, prediction, and search. Springer.

Tibshirani, R., 1996. Regresson shrinkage and selection via the lasso. Journal of the Royal Statistical Society 58, 267–288.

Wooldridge, J., 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge.

Zimmert, M., 2020. Efficient difference-in-differences estimation with high-dimensional common trend confounding. arXiv preprint 1809.01643 .