Applied Causal Inference Powered by ML and AI

Victor Chernozhukov* Christian Hansen† Nathan Kallus‡
Martin Spindler§ Vasilis Syrgkanis¶

March 5, 2025

Publisher: Online
Version 0.1.1

* MIT
† Chicago Booth
‡ Cornell University
§ Hamburg University
¶ Stanford University

16 Difference-in-Differences

"Corpus omne perseverare in statu suo quiescendi vel movendi uniformiter in directum, nisi quatenus a viribus impressis cogitur statum illum mutare." ("An object remains in its state of rest or of moving uniformly in a straight direction, unless forced to change that state by impressed forces.")

– Isaac Newton [1].

Here we discuss debiased machine learning (DML) methods for performing inference on average causal effects in panel (or longitudinal) or repeated cross-section data in the difference-in-differences (DiD) framework. We present and discuss the key identifying assumption for the average treatment effect on the treated based on DiD – the so-called "parallel trends" assumption – allowing for high-dimensional observed confounding variables. This assumption suggests a natural estimation strategy that directly applies DML to estimate average treatment effects on the treated using differenced outcomes.

16.1 Introduction

We now consider estimation of causal effects in panel (longitudinal) data where we observe individual units in multiple time periods or repeated cross-section data. While there are many potential approaches for analyzing data with both a cross-sectional and temporal component, we specifically look at difference-in-differences (DiD) and closely related approaches.

DiD and related methods are widely used in empirical work in the social sciences and in policy analysis. The basic DiD structure relies on having two groups of observations – a treatment group and a control group – for two time periods – a pre-treatment and a post-treatment period. Canonical DiD analysis then proceeds by comparing changes in the average pre- and post-treatment outcomes in the treatment group to changes in the average pre- and post-treatment outcomes in the control group. Attaching a causal interpretation to this comparison relies on an assumption that imposes that changes in the treatment group *in the absence of treatment* would have been the same as changes in the control group. This assumption captures the intuition that the treatment group would have evolved along the same path as the control group in the absence of treatment – i.e., the two groups share "parallel trends." Under the parallel trends assumption, the difference between the treatment and control differences between the pre- and post-treatment averages identifies the average treatment effect on the treated (ATET).

Figure 16.1: DiD is perhaps the oldest quasi-experimental research design. John Snow was a London doctor and is often considered the father of modern epidemiology. [2] is essentially an effort to provide convincing evidence that water is the causal agent for cholera transmission. It presents and discusses multiple pieces of evidence – including a DiD. *Source:* https://www.micropia.nl/en/discover/microbiology/john-snow/, accessed 6/7/23.

In this chapter, we review the basic DiD framework. We then focus on DiD in a setting where a researcher wishes to impose *conditional parallel trends*. That is, we consider settings where there are observed variables that are thought to be related to the evolution of the outcome of interest such that parallel trends holds only after conditioning on these variables. After suitably defining the conditional parallel trends assumption, we illustrate that the DML approach to estimating ATET from Chapter 9 can be readily applied within the DiD context.

16.2 The Basic Difference-in-Differences Framework: Parallel Worlds

The basic DiD structure has many appealing features. It is intuitive. It allows for essentially unrestricted differences in baseline outcomes for the treatment and control groups and allows for treatment to depend on those baseline differences. Estimation and inference are also relatively straightforward. Here we review the DiD structure using potential outcomes notation and highlight the key identifying assumptions.

The canonical DiD structure relies on existence of two time periods, denoted $t = 1$ and $t = 2$, and maintains that all observations are in the control state at $t = 1$. As such, we introduce potential outcomes

$$Y_t(d)$$

where $d \in \{0, 1\}$ denotes the treatment state in period $t = 2$. For example, $Y_1(1)$ denotes the period one outcome under treatment – that is, the outcome in the period before treatment is received – and $Y_2(1)$ denotes the period two outcome under treatment. Let $D \in \{0, 1\}$ be the treatment group indicator with $D = 1$ indicating that treatment is received at $t = 2$ and $D = 0$ indicating no treatment in either time period. Observed outcomes in period t may then be represented as $Y_t = DY_t(1) + (1 - D)Y_t(0)$. As in other causal inference contexts, we are left with missing data as we are unable to observe observations simultaneously in the treatment and control state.

One can also define four potential outcomes $(Y_t(0,0), Y_t(0,1), Y_t(1,0), Y_t(1,1))$ for each time period. The DiD structure imposes that $(Y_t(1,0), Y_t(1,1))$ can never be observed so it is impossible to learn about the effects of treatment paths that have treatment occur at t = 1. We choose the simpler representation with a single argument in the potential outcomes for notational clarity. Keeping explicit track of potential outcomes for different treatment paths is important in more complicated settings with more potential treatment paths as may arise with many time periods or more complex treatment variables.

DiD proceeds under the following key assumption:

Assumption 16.2.1 (Parallel Trends and No Anticipation) *Potential outcomes satisfy*

$$\mathrm{E}[Y_2(0) - Y_1(0) \mid D = 1] = \mathrm{E}[Y_2(0) - Y_1(0) \mid D = 0] \quad (16.2.1)$$

and

$$\mathrm{E}[Y_1(0) \mid D = 1] = \mathrm{E}[Y_1(1) \mid D = 1]. \quad (16.2.2)$$

Condition (16.2.1) is the *parallel trends* assumption. It requires that, in expectation, the change in control potential outcomes among the treatment group is the same as the change in the control potential outcomes among the control group. Condition (16.2.2) imposes that receipt of treatment at $t = 2$ does not impact average period 1 potential outcomes. Here, we are effectively ruling out anticipation effects. Importantly, (16.2.2) allows for systematic differences between average potential outcomes among treated and control observations in the pre-treatment period. That is, it does not impose that $\mathrm{E}[Y_1(0) \mid D = 1] = \mathrm{E}[Y_1(0) \mid D = 0]$. Thus, we can accommodate, for example, scenarios where we believe that period two treatment assignment is related to period one outcomes.

Often, the no anticipation assumption is left implicit or ignored. We state it for clarity and because it allows clean definition of the causal effect of interest.

It is worth explicitly noting that the parallel trends assumption is typically functional form dependent. That is, if $\mathrm{E}[Y_2(0) - Y_1(0) \mid D = 1] = \mathrm{E}[Y_2(0) - Y_1(0) \mid D = 0]$, it will generally not be the case that $\mathrm{E}[g(Y_2(0)) - g(Y_1(0)) \mid D = 1] = \mathrm{E}[g(Y_2(0)) - g(Y_1(0)) \mid D = 0]$. For example, suppose the outcome of interest is wages. Parallel trends holding for wage does not imply that parallel trends holds for log(wage), and the DiD estimator based on log(wage) need not recover a causal effect. Intuitively, this functional form dependence arises because parallel trends relies on latent sources of confounding being additively separable so that they are eliminated by the differencing operation.

One situation where parallel trends hold regardless of the outcome transformation is when D is randomly assigned (i.e., when $(Y_1(0), Y_1(1), Y_2(0), Y_2(1) \perp D)$. See [3] for further discussion.

A related framework to DiD that is independent to monotone transformations (at the cost of other restrictions, of course) is the changes-in-changes model of [4].

It is straightforward to verify that ATET is identified under Assumption 16.2.1. Note that the right-hand-side of Eq. (16.2.1) is an observable quantity while the left-hand-side corresponds to the unobservable change in the control potential outcomes of treated units. Parallel trends allows us to impute this latent change from the observed change in the control units. Effectively, we are assuming that the treated observations would have changed in the same way as the control observations in the absence of treatment. Similarly, the right-hand-side of Eq. (16.2.2) is an observable quantity while the left-hand-side is the unobserved average of control potential outcomes in period one for the treated group. Eq. (16.2.2) allows us to impute this baseline average from the observed baseline average in the treatment group. We can then reconstruct the counterfactual average of the control potential outcome in the post-treatment period by adjusting this baseline average by the observed change in average outcomes between the two periods in the control group. Figure 16.2 presents a graphical illustration of the identification argument.

More formally, we can put this together to write the ATET as

$$\begin{aligned}\alpha &= \mathrm{E}[Y_2(1) - Y_2(0) \mid D = 1] \\ &= \mathrm{E}[Y_2(1) \mid D = 1] - \mathrm{E}[Y_2(0) \mid D = 1] \\ &= \mathrm{E}[Y_2(1) \mid D = 1] \\ &\qquad - (\mathrm{E}[Y_1(1) \mid D = 1] + \mathrm{E}[Y_2(0) - Y_1(0) \mid D = 0]) \\ &= \mathrm{E}[Y_2(1) - Y_1(1) \mid D = 1] - \mathrm{E}[Y_2(0) - Y_1(0) \mid D = 0]\end{aligned} \tag{16.2.3}$$

where the third equality follows from a direct application of Assumption 16.2.1. The expression in the last line is exactly the difference between the difference between post- and pre-treatment period average outcomes in the treatment group and the difference between post- and pre-treatment period average outcomes in the control group – hence, difference-in-differences.

One could identify the ATE by augmenting Assumption 16.2.1 with $\mathrm{E}[Y_2(1) - Y_1(1) \mid D = 0] = \mathrm{E}[Y_2(1) - Y_1(1) \mid D = 1]$ and $\mathrm{E}[Y_1(0) \mid D = 0] = \mathrm{E}[Y_1(1) \mid D = 0]$. Recall that we are notation subsumes treatment paths so that $Y_2(1)$ denotes the potential outcome in the second time period of a unit under control in period 1 and under treatment in period 2 while $Y_1(1)$ denotes the first time period of unit under control in period 1 and under treatment in period 2. The first condition is thus not just a restriction on evolution of potential outcomes in a specific treatment state but also a restriction on treatment effects themselves which seems hard to motivate in realistic settings. As such, we follow the majority of the DiD literature in focusing on estimation of ATET.

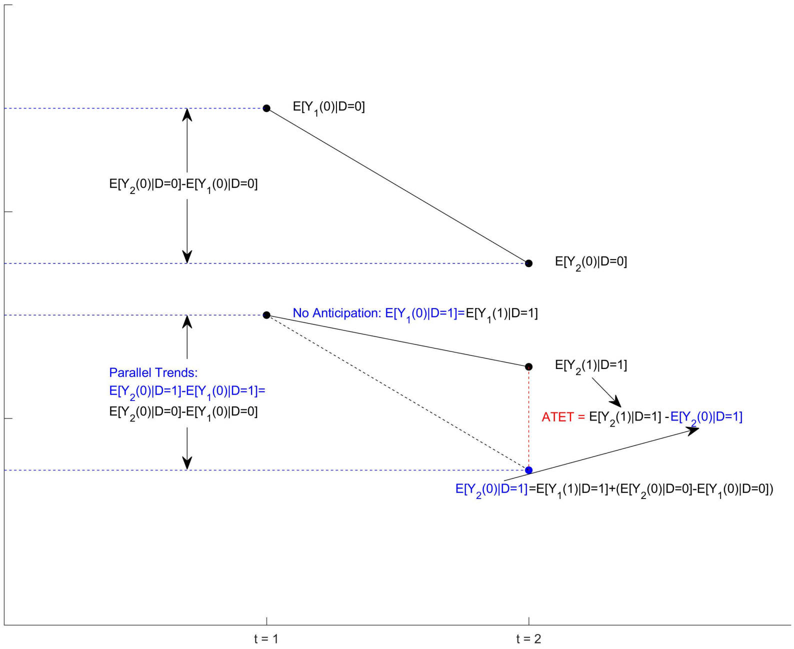
E[Y_1(0)|D=0]
E[Y_2(0)|D=0]-E[Y_1(0)|D=0]
E[Y_2(0)|D=0]
No Anticipation: E[Y_1(0)|D=1]=E[Y_1(1)|D=1]
Parallel Trends:
E[Y_2(0)|D=1]-E[Y_1(0)|D=1]=
E[Y_2(0)|D=0]-E[Y_1(0)|D=0]
E[Y_2(1)|D=1]
ATET = E[Y_2(1)|D=1] -E[Y_2(0)|D=1]
E[Y_2(0)|D=1]=E[Y_1(1)|D=1]+(E[Y_2(0)|D=0]-E[Y_1(0)|D=0])
t = 1
t = 2

Figure 16.2: DiD Identification. This figure illustrates identification of the ATET in the canonical DiD framework. Objects represented in black are observable. Objects in blue are unobserved and identified via Assumption 16.2.1. Visually we impute the unobserved $E[Y_2(0) \mid D = 1]$ by extrapolating from the observed $E[Y_1(1) \mid D = 1]$ using the observed "trend" between $E[Y_1(0) \mid D = 0]$ and $E[Y_2(0) \mid D = 0]$. The ATET is then the difference between the observed $E[Y_2(1) \mid D = 1]$ and the imputed $E[Y_2(0) \mid D = 1]$.

Estimation of the ATET in canonical DiD in a finite sample is straightforward by considering four group means:

$$\hat{\theta}_s(d) = \frac{\mathbb{E}_n[Y1(D = d, t = s)]}{\mathbb{E}_n[1(D = d, t = s)]}.$$

Defining the estimator of the ATET as $\widehat{\alpha}$, we have

$$\widehat{\alpha} = (\hat{\theta}_2(1) - \hat{\theta}_1(1)) - (\hat{\theta}_2(0) - \hat{\theta}_1(0)). \tag{16.2.4}$$

Asymptotic properties under independence follow in a fashion similar to difference-in-mean estimators for the ATE outlined in Chapter **??**.

We can also obtain a numerically equivalent estimator of the ATET via regression. Specifically, the ordinary least squares estimator of the parameter α in the linear model

$$Y = \beta_0 + \beta_1 D + \beta_2 P + \alpha DP + U, \tag{16.2.5}$$

where P is a binary variable with $P = 1$ indicating the post-treatment time period (t = 2), is numerically equivalent to $\widehat{\alpha}$ in (16.2.4). The regression formulation is especially convenient for obtaining standard errors under different dependence assumptions.

	1979	1981	Difference
Miami Unemployment	5.1	3.9	-1.2
	(1.1)	(0.9)	(1.4)
Comparison Unemployment	4.4	4.3	-0.1
	(0.3)	(0.3)	(0.4)
Difference (Miami - Comparison)	0.7	-0.4	-1.1
	(1.1)	(0.9)	(1.5)

Note: Unemployment rates among white individuals in Miami and four comparison cities – Atlanta, Los Angeles, Houston, and Tampa-St. Petersburg – reproduced from [5]. Standard errors assuming independence are in parentheses. The DiD estimate is provided in the entry in the last row and column.

Table 16.1: DiD Estimation of the Effect of the Mariel Boatlift on Unemployment

The Mariel Boatlift

Card's analysis of the impact of the Mariel Boatlift on the Miami labor market, [5], provides a prototypical application of DiD. For example, it is the example of DiD in Angrist and Krueger's *Handbook of Labor Economics* chapter on empirical methods [6]. The basic idea of the study was to use the Mariel Boatlift – a sudden and arguably unexpected inflow of immigrants that increased the Miami labor force by about 7% between May and September of 1980 – to understand the impact of immigration on low-skilled labor market outcomes.

A key component of the analysis was arguing that Atlanta, Los Angeles, Houston, and Tampa-St. Petersburg provide valid control cities in the sense that we might plausibly believe that the change in labor market outcomes in these cities from the late 1970's to the early 1980's is useful for inferring how the Miami labor market would have changed in the absence of the Mariel immigration. Part of the argument in [5] relies on evidence that the cities had similar characteristics in the pre-treatment period. Effectively, this argument relies on parallel trends holding *conditional* on these pre-treatment characteristics. We consider using DML to flexibly control for rich covariates in Section 16.3.

We illustrate canonical DiD in the Mariel Boatlift example in Table 16.1 which uses numbers taken from Table 4 in [5]. Here, we see the DiD estimate of the ATET on unemployment is -1.1 with standard error 1.5, which does not provide strong evidence of a large impact of the Mariel immigration on unemployment.

16.3 DML and Conditional Difference-in-Differences

In many empirical applications, researchers deviate from the canonical DiD framework by including additional control variables. The fundamental motivation is similar to that for including control variables in other causal contexts, e.g., as motivated in Chapter 5: it is easier to believe that parallel trends holds among units that are identical in terms of observed characteristics. In this section, we explore flexibly including control variables in a DiD framework leveraging DML methods.

We restate the canonical DiD assumptions so that they hold after conditioning on pre-treatment/strictly exogenous characteristics X.

Assumption 16.3.1 (Conditional DiD Assumptions) *Potential outcomes satisfy conditional parallel trends*

$$\begin{aligned} \mathrm{E}[Y_2(0) - Y_1(0) \mid D = 1, X] = \\ \mathrm{E}[Y_2(0) - Y_1(0) \mid D = 0, X] \ a.s. \end{aligned} \tag{16.3.1}$$

and no anticipation

$$\mathrm{E}[Y_1(0) \mid D = 1, X] = \mathrm{E}[Y_1(1) \mid D = 1, X] \ a.s. \tag{16.3.2}$$

In addition, there is a treatment group and its characteristics overlap with the control group

$$\exists\, \varepsilon > 0 : \mathrm{P}(D = 1) \geq \varepsilon \textit{ and } \mathrm{P}(D = 1 \mid X) \leq 1 - \varepsilon \ a.s. \tag{16.3.3}$$

The intuition for (16.3.1) and (16.3.2) is essentially identical to the intuition for Assumption 16.2.1 discussed in the previous section. The only difference is that these conditions are now imposed within observationally identical groups as defined by X. Condition (16.3.3) is a standard overlap condition for identifying ATET which essentially imposes that there are control observations available for every value of X. Under Assumption 16.3.1, it is straightforward to verify that the ATET is identified by repeating the argument in (16.2.3) conditional on X and averaging over the distribution of X in the $D = 1$ group.

We leave verification of identification of the ATET in the conditional DiD framework as an exercise.

Similar to estimating parameters in the partially linear model or average treatment effects under confounding as discussed in Chapter 9, obtaining estimates of the ATET in the conditional

DiD setting will require estimating high-dimensional nuisance objects. We thus exploit DML methods to accommodate the use of flexible methods in estimating these objects.

A key input into DML estimation is a Neyman orthogonal score. In the conditional DiD framework with panel data,[1]

$$\psi(W;\alpha,\eta) = \frac{D - m(X)}{p(1-m(X))}(\Delta Y - g(0,X)) - \frac{D}{p}\alpha \qquad (16.3.4)$$

provides an orthogonal score for the ATET, α, where $W = (Y_1, Y_2, D, X)$ denotes the observable variables; $\Delta Y = Y_2 - Y_1$; $\eta = (p, m, g)$ denotes nuisance parameters with true values $p_0 = \mathrm{E}[D]$, $m_0(X) = \mathrm{E}[D \mid X]$, and $g_0(0, X) = \mathrm{E}[\Delta Y \mid D = 0, X]$. See also [7], [8], [9]. Comparing to the score for the ATET provided in Chapter 9, we see that the score function in (16.3.4) is identical to that for learning the ATET under conditional ignorability where the outcome variable is simply defined as ΔY.

Given the Neyman orthogonal score (16.3.4), it is then straightforward to implement DML to estimate the ATET. $\sqrt{n}$-asymptotic normality of $\widehat{\alpha}$, the DML estimator of the ATET, follows from Theorem 9.4.1 in Chapter 9.

DML for ATET in Conditional DiD

Let $(W_i)_{i=1}^n = (Y_{1i}, Y_{2i}, D_i, X_i)_{i=1}^n$ be the observed data.

1. Partition sample indices into random folds of approximately equal size: $\{1, ..., n\} = \cup_{k=1}^K I_k$. For each $k = 1, ..., K$, compute estimators $\hat{p}_{[k]}$, $\hat{g}_{[k]}$, and $\hat{m}_{[k]}$ of $\mathrm{E}[D]$ and the conditional expectation functions $g_0(0, X) = \mathrm{E}[\Delta Y \mid D = 0, X]$ and $m_0(X) = \mathrm{E}[D \mid X]$ leaving out the k^{th} block of data and enforcing $\hat{m}_{[k]} \leq 1 - \epsilon$.
2. For each $i \in I_k$, let $k(i)$ denote the fold to which observation i belongs and

$$\begin{aligned}\hat{\psi}(W_i;\alpha) = {}& \frac{D_i - \hat{m}_{[k(i)]}(X_i)}{\hat{p}_{[k(i)]}(1 - \hat{m}_{[k(i)]}(X_i))}(\Delta Y_i - \hat{g}_{[k(i)]}(0, X_i)) \\ & - \frac{D_i}{\hat{p}_{[k(i)]}}\alpha.\end{aligned}$$

Compute the estimator $\widehat{\alpha}$ as the solution to

[1] We provide the Neyman orthogonal score and discuss DML estimation with repeated cross-sections in Section 16.A.

$\mathbb{E}_n[\hat\psi(W_i;\alpha)] = 0$ which yields

$$\widehat{\alpha} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{D_i - \hat m_{[k(i)]}(X_i)}{\hat p_{[k(i)]}(1-\hat m_{[k(i)]}(X_i))}(\Delta Y_i - \hat g_{[k(i)]}(0, X_i))}{\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat p_{[k(i)]}}}.$$

3. Let

$$\hat\varphi(W_i) = \frac{\hat\psi(W_i;\hat\alpha)}{\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat p_{[k(i)]}}}.$$

Construct standard errors via

$$\sqrt{\hat{\mathrm{V}}/n}, \quad \hat{\mathrm{V}} = \mathbb{E}_n[\hat\varphi(W_i)^2]$$

and use standard normal critical values for inference.

Comparison to Adding Regression Controls

The equivalence between the ATET estimator obtained by directly looking to the difference between the treatment and control differences in means and the ordinary least squares estimator of the coefficient α in the linear model (16.2.5) in the canonical DiD setting suggests a simple approach to incorporating control variables by augmenting the regression model to include controls linearly. That is, add $\beta' X$ to the model in (16.2.5). However, the coefficient on the DP-interaction term is not equivalent to the ATET and need not uncover any sensible causal effect without very strong functional form restrictions and restrictions on treatment effect heterogeneity. See, e.g., [10] for further discussion. In contrast, the DML estimator always targets the ATET under Assumption 16.3.1 and is relatively simple to implement.

The Notebooks 16.6.1 contain the code for the minimum wage example.

16.4 Example: Minimum Wage

In this section, we use DML for DiD to estimate the effect of minimum wage increases on teen employment. We use data from and roughly follow the approach of [11]. The data are annual county level data from the United States covering 2001 to 2007. The outcome variable is log county-level teen employment, and the treatment variable is an indicator for whether the county has a minimum wage above the federal minimum wage.[2] Note

[2] The federal minimum wage over 2001-2007 was constant at $5.15.

that this definition of the treatment variable makes the analysis straightforward but ignores the nuances of the exact value of the minimum wage in each county and how far those values are from the federal minimum.[3] The data also includes county population and county average annual pay. We follow [11] by removing observations with missing entries which produces a balanced panel with data from counties in 42 states. See [11], [12], and [13] for further details regarding the data.

We focus our analysis exclusively on the set of counties that had wage increases away from the federal minimum wage in 2004. That is, we treat 2003 and earlier as the pre-treatment period and the period 2004-2007 as the post-treatment period. We assume that parallel trends holds after conditioning on three pre-treatment variables – 2001 population, 2001 average pay, and 2001 teen employment – and the region to which each county belongs.[4]

We estimate dynamic effects by estimating the ATET in 2004-2007 which provide estimates of the effect in the year of treatment and one, two, and three years after the treatment. For control observations, we use the set of observations that still have minimum wage equal to the federal minimum in each year – the "as-yet not treated" – so the control group changes from period to period. For example, we use all observations that had minimum wage equal to the federal minimum in 2004 as control observations when estimating the ATET in 2004, but we use all observations that had minimum wage equal to the federal minimum in 2005 as control observations to estimate the 2005 ATET. These definitions yield 102 treatment observations for estimating each ATET and 2389, 2327, 2080, and 1417 control observations for 2004, 2005, 2006, and 2007 respectively.

Since our goal is to estimate the ATET of the county level minimum wage being larger than the federal minimum imposing that parallel trends holds after flexibly controlling for region and our pre-treatment variables, we employ DML using the algorithm from Section 16.3. We consider using an array of methods for learning the nuisance functions including several of the modern regression methods that we discussed in previous chapters. Specifically, we consider ten candidate learners for the high-dimensional nuisance functions $g_0(0, X) = \mathrm{E}[\Delta Y \mid D = 0, X]$ and $m_0(X) = \mathrm{E}[D \mid X]$. We consider using no control variables (No Controls) which corresponds to maintaining unconditional parallel trends. We consider linear index models using only the raw control variables (Basic) – the four region dummies and log of 2001 population, log of 2001 average pay, and log of 2001 employment – and using a full cubic expansion of the raw control

[3] 3: Under these definitions, this example is an example of *staggered adoption*. Staggered adoption refers to a setting with a binary, absorbing treatment variable. That is, once an observation becomes treated it remains treated thereafter. This setting is straightforward to analyze as treatment paths are completely characterized by the treatment date and controls can be constructed from observations that are not treated during the sample period (the never treated) or observations that are not treated prior to the treatment date and remain untreated in the period in which one wants to estimate the ATET (the as-yet not treated).

[4] 4: We follow [11] and categorize each observation as belonging to one of four U.S. census regions.

Table 16.2: RMSE for Learners in Minimum Wage example

	2004	2005	2006	2007
	A. $\mathrm{E}[\Delta Y \mid D = 0, X]$			
No Controls	0.1633	0.1882	0.2235	0.2302
Basic	0.1634	0.1854	0.2191	0.2216
Expansion	0.1887	0.2122	0.2445	0.2710
Lasso (CV)	0.1631	0.1851	0.2193	0.2214
Ridge (CV)	0.1631	0.1851	0.2191	0.2213
Random Forest	0.1716	0.1982	0.2330	0.2388
Deep Tree	0.1922	0.2250	0.2599	0.2708
Shallow Tree	0.1678	0.1924	0.2279	0.2290
Tree (CV)	0.1633	0.1889	0.2178	0.2227
	B. $\mathrm{E}[D \mid X]$			
No Controls	0.1983	0.2006	0.2111	0.2503
Basic	0.1986	0.2009	0.2113	0.2217
Expansion	0.1988	0.2007	0.2113	0.2217
Lasso (CV)	0.1968	0.1986	0.2083	0.2197
Ridge (CV)	0.1971	0.1989	0.2086	0.2198
Random Forest	0.2005	0.2051	0.2128	0.2355
Deep Tree	0.2207	0.2364	0.2303	0.2744
Shallow Tree	0.1921	0.1944	0.2029	0.2301
Tree (CV)	0.1937	0.1955	0.2039	0.2311

Note: Cross-fit RMSE for predicting ΔY and treatment status D in the minimum wage example. Row labels denote the method used to estimate the nuisance function, and column labels indicate the year for which we are calculating the ATET, with 2004, 2005, 2006, and 2007 respectively corresponding to the year of the treatment, one year after treatment, two years after treatment, and three years after treatment.

variables including all third order interactions (Expansion).[5] We consider Lasso and Ridge with the cubic expansion of the raw variables and penalty parameter chosen by cross-validation (Lasso (CV) and Ridge (CV)). We consider a random forest with no randomization over input variables and 1000 trees (Random Forest). Additionally, we consider three different tree models: a tree with depth 15 (Deep Tree), a tree with depth 3 (Shallow Tree), and a tree tuned using cross-validation (Tree (CV)). For random forest and the tree models, we use region, log of 2001 population, log of 2001 average pay, and log of 2001 employment as input variables. Finally, we consider estimation using the learner for $\mathrm{E}[\Delta Y \mid D = 0, X]$ and for $\mathrm{E}[D \mid X]$ that produce the lowest RMSE during cross-fitting (Best) allowing for a different learner to be selected for each task.[6]

We start by reporting the RMSE obtained during cross-fitting for each learner in each period in Table 16.2. Here we see that the Deep Tree systematically performs substantially worse in terms of cross-fit predictions than the other learners for both tasks and that Expansion performs similarly poorly for the outcome prediction. It also appears there is some signal in the regressors,

[5] We use a linear model estimated by OLS for $g_0(0, X)$ and a logistic model with linear index in the stated variables for $m_0(X)$.

[6] For any observation with estimated propensity score larger than 0.95, we replace the propensity score with 0.95. Applying this trimming, we replace 12, 10, 13, and 21 observations for the deep tree in 2004-2007 respectively and replace 2, 2, and 1 observation for Basic, Expansion, and Lasso (CV) in 2007.

Table 16.3: Estimated ATET in Minimum Wage example

	2004	2005	2006	2007
No Controls	-0.039	-0.076	-0.117	-0.131
	(0.019)	(0.021)	(0.023)	(0.026)
Basic	-0.037	-0.066	-0.088	-0.041
	(0.018)	(0.020)	(0.021)	(0.033)
Expansion	-0.022	-0.046	-0.061	0.303
	(0.025)	(0.030)	(0.033)	(0.227)
Lasso (CV)	-0.035	-0.062	-0.082	-0.049
	(0.018)	(0.020)	(0.021)	(0.031)
Ridge (CV)	-0.035	-0.062	-0.083	-0.061
	(0.018)	(0.020)	(0.021)	(0.025)
Random Forest	0.013	-0.056	-0.039	-0.071
	(0.029)	(0.024)	(0.028)	(0.038)
Deep Tree	0.077	0.007	0.100	-0.470
	(0.079)	(0.172)	(0.080)	(0.178)
Shallow Tree	-0.028	-0.040	-0.058	-0.065
	(0.019)	(0.021)	(0.021)	(0.026)
Tree (CV)	-0.027	-0.045	-0.060	-0.069
	(0.019)	(0.021)	(0.021)	(0.025)
Best	-0.028	-0.051	-0.055	-0.055
	(0.019)	(0.021)	(0.021)	(0.031)

Note: Estimated ATET and standard errors (in parentheses) in the minimum wage example. Row labels denote the method used to estimate the nuisance function, and column labels indicate the year for which we are calculating the ATET, with 2004, 2005, 2006, and 2007 respectively corresponding to the year of the treatment, one year after treatment, two years after treatment, and three years after treatment.

especially for the propensity score, as all methods outside of Deep Tree and Expansion produce notably smaller RMSEs than the No Controls baseline. The other methods all produce similar RMSEs, with a small edge going to Ridge and Lasso. While it would be hard to reliably conclude which of the relatively good performing methods is statistically best here, one could exclude Expansion and Deep Tree from further consideration on the basis of out-of-sample performance suggesting they are doing a poor job approximating the nuisance functions. Best (or a different ensemble) provides a good baseline that is principled in the sense that one could pre-commit to using the best learners without having first looked at the subsequent estimation results.

We report estimates of the ATET in each period in Table 16.3. Here, we see that the majority of methods provide point estimates that suggest the minimum wage increase leads to decreases in youth employment with small effects in the initial period that become larger in the years following the treatment. This pattern seems economically plausible as it may take time for firms to adjust employment and other input choices in response to the minimum wage change. The methods that produce estimates that are not consistent with this pattern are Deep Tree and Expansion which are both suspect as they systematically underperform in terms of having poor cross-fit prediction performance. In terms of point estimates, the other pattern that emerges is that all estimates that use the covariates produce ATET estimates that are systematically smaller in magnitude than the No Controls baseline, suggesting that failing to include the controls may lead to overstatement of treatment effects in this example.

Turning to inference, we would reject the hypothesis of no minimum wage effect in 2005 and 2006 at the 5% level, even after multiple testing correction, if we were to focus on the row "Best" (or many of the other individual rows). Focusing on "Best" is a reasonable ex ante strategy that could be committed to prior to conducting any analysis. It is, of course, reassuring that this broad conclusion is also obtained using many of the individual learners suggesting some robustness to the exact choice of learner made.

Because we have data for the period 2001-2007, we can perform a so-called *placebo* or *pre-trends* test to provide some evidence about the plausibility of the conditional DiD assumptions, Assumption 16.3.1. Specifically, we can continue to use 2003 as the reference period but now consider 2002 to be the treatment period. Sensible economic mechanisms underlying Assumption 16.3.1 would typically suggest that the ATET in 2002 – before the 2004 minimum wage change we are considering – should be zero. Finding evidence that the ATET in 2002 is non-zero then calls into question the validity of Assumption 16.3.1.

We repeat the exercise for obtaining our ATET estimates and standard error for 2004-2007 and report the results in Table 16.4. Here we see broad agreement across all methods in the sense of returning point estimates that are small in magnitude and small relative to standard errors. In no case would we reject the hypothesis that the pre-event effect in 2002 is different from zero at usual levels of significance. We note that failing to reject the hypothesis of no pre-event effects certainly does not imply that Assumption 16.3.1 is in fact satisfied. For example, confidence intervals include values that would be consistent with relatively large pre-event effects. Conditioning inference on the results of such an assessment is also generally a bad idea; see, e.g. [14] and [15] for a discussion specifically in the context of DiD. However, it is reassuring to see that there is not strong evidence of a violation of the underlying identifying assumption.

	RMSE Y	RMSE D	ATET	s.e.
No Controls	0.1543	0.1945	-0.0037	(0.0131)
Basic	0.1541	0.1949	-0.0044	(0.0130)
Expansion	0.1577	0.1949	0.0046	(0.0140)
Lasso (CV)	0.1544	0.1932	-0.0039	(0.0131)
Ridge (CV)	0.1544	0.1935	-0.0053	(0.0131)
Random Forest	0.1635	0.2265	0.0230	(0.0265)
Deep Tree	0.1822	0.2234	0.0080	(0.0276)
Shallow Tree	0.1620	0.1884	-0.0037	(0.0134)
Tree (CV)	0.1550	0.1905	-0.0056	(0.0133)
Best	0.1541	0.1884	-0.0031	(0.0134)

Note: Estimated pre-event (2002) ATET and standard errors (in parentheses) in the minimum wage example. Row labels denote the method used to estimate the nuisance function. RMSE Y and RMSE D give cross-fit RMSE for the outcome and treatment respectively. ATET provides the point estimate of the ATET based on the method in the row label with standard error given in column s.e.

Table 16.4: Pre-trends Assessment

16.5 Notes

There is a relatively large literature focusing on flexibly estimating ATET in DiD contexts. Much of this work has focused on potential failure of the usual practice of estimating homogeneous coefficient linear models with additive fixed effects for groups and time periods under heterogeneous treatment effects. Specifically, much of the work has noted that coefficients on a treatment variable in a homogeneous linear model with fixed effects need not be proper weighted averages of heterogeneous treatment effects but may place negative weights on some effects. The possibility of negative weights then leaves open the possibility of, for example, having uniformly positive treatment effects but obtaining negative and significant estimates of the coefficient on a treatment variable in a linear model. The DML approach we present in this chapter offers one solution to this problem that allows for flexibly accommodating control variables that can account for heterogeneity. See the excellent review papers [11], [16], [17] for more discussion.

16.6 Notebooks

Notebook 16.6.1 (Minimum Wage) Minimum Wage R Notebook and Minimum Wage Python Notebook contain the analysis of minimum wage example

16.7 Exercises

Exercise 16.7.1 (ATET) Verify that the ATET is identified under Assumption 16.3.1. Provide a short explanation of the intuition for the identification result. Give an intuitive example where the ATET would be identified after conditioning on covariates but where identification of the ATET would fail in the canonical DiD framework (i.e. without conditioning on additional covariates).

Exercise 16.7.2 (Minimum Wage I) Study the minimum wage empirical analysis notebook. Estimate the ATET for observations treated in a year different than 2004 – e.g. repeat the analysis doing the exercise for observations treated in 2005.

Exercise 16.7.3 (Minimum Wage II) Study the minimum wage empirical analysis notebook. Estimate the ATET using the never treated as opposed to the not-yet treated as the control group.

16.A Conditional Difference-in-Differences with Repeated Cross-Sections

Here we provide the Neyman orthogonal score for the ATET in the conditional DiD context with repeated cross-section data. For additional development including formal statement of additional assumptions for DiD with repeated cross sections, see [7], [8], [9].

The chief difference in this setting relative to when one has panel data is that we cannot directly construct the difference between outcomes in the first and second period as we do not see the same individuals across time periods. Rather, we revert to the analog of the canonical DiD estimator by directly working with the four conditional means defined by grouping the treated and control observations pre- and post-treatment. Specifically, we make use of the score function

$$\begin{aligned}\psi(W,\alpha,\eta) = &\left(\frac{DT}{p\lambda}(Y-g(1,2,X))\right.\\ &\left.-\frac{D(1-T)}{p(1-\lambda)}(Y-g(1,1,X))\right)\\ &-\left(\frac{m(X)(1-D)T}{p\lambda(1-m(X))}(Y-g(0,2,X))\right.\\ &\left.-\frac{m(X)(1-D)(1-T)}{p(1-\lambda)(1-m(X))}(Y-g(0,1,X))\right)\\ &+\frac{D}{p}\left(g(1,2,X)-g(1,1,X)\right)\\ &-\frac{D}{p}\left(g(0,2,X)-g(0,1,X)\right)-\frac{D}{p}\alpha\end{aligned} \tag{16.A.1}$$

where $W = (Y, T, D, X)$ denotes the observable variables for each observation with T an indicator which equals one if the observation is in the post-treatment period (period 2) and $\eta = (p, \lambda, m, g)$ denotes nuisance parameters with true values $p_0 = \mathrm{E}[D]$, $\lambda_0 = \mathrm{E}[T]$, $m_0(X) = \mathrm{E}[D \mid X]$, and $g_0(d,t,X) = \mathrm{E}[Y \mid D = d, T = t, X]$.

Under iid sampling, we can directly apply the generic cross-fitting approach to DML as in Section 9.4. In many DiD settings, researchers wish to allow for unmodeled dependence between observations corresponding to different groups such as cities or counties. As long as there are many such groups, it is straightforward to modify the DML algorithm to accommodate this dependence. The algorithm simply needs to be adjusted by forming the cross-fitting folds such that all observations within groups are included together in the same fold. Similarly, it is straightforward to adjust inference to account for this dependence by applying clustered standard errors with clustering done at the group level.

Bibliography

[1] Isaac Newton. *PhilosophiæNaturalis Principia Mathematica*. 1687 (cited on page 451).

[2] John Snow. *On the Mode of Communication of Cholera*. Edited by John Churchill, Second edition, London. 1855 (cited on page 452).

[3] Jonathan Roth and Pedro H. C. Sant'Anna. 'When is parallel trends sensitive to functional form?' In: *Econometrica* 91 (2 2023), pp. 737–747 (cited on page 454).

[4] Susan Athey and Guido W Imbens. 'Identification and inference in nonlinear difference-in-differences models'. In: *Econometrica* 74.2 (2006), pp. 431–497 (cited on page 454).

[5] David Card. 'The impact of the Mariel boatlift on the Miami labor market'. In: *Industrial and Labor Relations Review* 43 (1990), pp. 245–257 (cited on page 456).

[6] Joshua D. Angrist and Alan B. Krueger. 'Empirical Strategies in Labor Economics'. In: *Handbook of Labor Economics. Volume 3*. Ed. by O. Ashenfelter and D. Card. Elsevier: North-Holland, 1999 (cited on page 456).

[7] Michael Zimmert. 'Efficient Difference-in-Differences Estimation with High-Dimensional Common Trend Confounding'. In: *arXiv preprint arXiv:1809.01643* (2018) (cited on pages 458, 466).

[8] Neng-Chieh Chang. 'Double/Debiased Machine Learning for Difference-in-Differences Models'. In: *Econometrics Journal* 23 (2 2020), pp. 177–191 (cited on pages 458, 466).

[9] Pedro H. C. Sant'Anna and Jun Zhao. 'Doubly Robust Difference-in-Differences Estimators'. In: *Journal of Econometrics* 219 (1 2020), pp. 101–122 (cited on pages 458, 466).

[10] Carolina Caetano and Brantly Callaway. 'Difference-in-Differences with Time-Varying Covariates in the Parallel Trends Assumption'. In: *arXiv preprint arXiv:2202.02903* (2023) (cited on page 459).

[11] Brantly Callaway. 'Difference-in-Differences for Policy Evaluation'. In: *Handbook of Labor, Human Resources and Population Economics*. Ed. by K. F. Zimmermann. Springer Cham, 2023 (cited on pages 459, 460, 464).

[12] Brantly Callaway and Pedro H. C. Sant'Anna. 'Difference-in-differences with multiple time periods'. In: *Journal of Econometrics* 225 (2 2021), pp. 200–230 (cited on page 460).

[13] Arindrajit Dube, T. William Lester, and Michael Reich. 'Minimum wage shocks, employment flows, and labor market frictions'. In: *Journal of Labor Economics* 34 (3 2016), pp. 663–704 (cited on page 460).

[14] Jonathan Roth. 'Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends'. In: *American Economic Review: Insights* 4 (3 2022), pp. 305–322 (cited on page 463).

[15] Ashesh Rambachan and Jonathan Roth. 'A More Credible Approach to Parallel Trends'. In: (2023). forthcoming *Review of Economic Studies* (cited on page 463).

[16] Jonathan Roth, Pedro Sant'Anna, Alyssa Bilinski, and John Poe. 'What's trending in difference-in-differences? A synthesis of the recent econometrics literature'. In: *Journal of Econometrics* 235 (2 2023), pp. 2218–2244 (cited on page 464).

[17] Clément de Chaisemartin and Xavier D'Haultfœuille. 'Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey'. In: *Econometrics Journal* (June 2022), utac017. DOI: 10.1093/ectj/utac017 (cited on page 464).