A Complete Framework for Model-Free Difference-in-Differences Estimation

Daniel J. Henderson
Department of Economics, Finance and Legal Studies
University of Alabama, USA
djhenderson1@ua.edu

Stefan Sperlich
Geneva School of Economics and Management
University of Geneva, Switzerland
stefan.sperlich@unige.ch

Abstract

We propose a complete framework for data-driven difference-in-differences analysis with covariates, in particular nonparametric estimation and testing. We start with simultaneously choosing confounders and a scale of the outcome along identification conditions. We estimate first heterogeneous treatment effects stratified along the covariates, then the average effect(s) for the treated. We provide the asymptotic and finite sample behavior of our estimators and tests, bootstrap procedures for their standard errors and p-values, and an automatic bandwidth choice. The pertinence of our methods is shown with a study of the impact of the Deferred Action for Childhood Arrivals program on educational outcomes for non-citizen immigrants in the US.

1 Introduction

Arguably the most popular estimation technique to study treatment effects in a Rubin-Causal-Model (Holland, 1986) is the so-called difference-in-differences (DiD) approach. Today, the literature on this and related approaches is quite abundant.[1] As with many methods for studying causality, it originates from biometrics, in this case it is attributed to the epidemiologist John Snow (*1813–†1858) who applied DiD for finding the cause of the cholera outbreak of 1854 in London. In economics it was made popular by Card and Krueger (1992) who employed this method for studying the causal effect of a minimum wage rise in New Jersey (of almost 20%) in 1992, comparing the developments of the labor

[1] For example, change-in-changes by Athey and Imbens (2006) shifts the problem from mean to quantile regression which has many advantages like scale-independence, but is less popular due to practical complications.

markets of New Jersey and Pennsylvania, concentrating on the low-income sector (we call such intervention or similar event a ‘treatment’).

In our opinion, the most intensive and extensive discussion on this topic was provided by (Lechner, 2011). He showed that the basic concept for identifying the causal effect via DiD applies to more complex situations than previously considered. In this article we limit our considerations to the case of a single treatment and two groups (treatment group, $D = 1$ and control group, $D = 0$); extensions as discussed by him work in principle the same way.

The DiD concept is feasible when a panel or repeated cross-sections of observations of individuals are provided both before and after an intervention has taken place. Although more often studied for panels, we outline all methods first for the more general case of repeated cross-sections (cohorts); but we show afterwards that the methods apply equally well to balanced panels and actually give much more simplified statistics and asymptotics. Notice that in its basic form, i.e., without imposing further non-testable assumptions, the DiD approach identifies the treatment effect on the treated. The primary assumption behind this identification is that without such intervention (i.e., the treatment), the outcome of interest Y experienced in both groups (treated and control group) would have developed ‘similarly’ over time, where ‘similarly’ for mean-regression refers to ‘in-the-mean’ but in quantile regression (change-in-changes) refers to the quantiles. This is also known as the ‘common trend’ or ‘parallel path’ condition. This insinuates that there had been only a constant difference between the two groups without the treatment under consideration.

Often it is unlikely that this difference is independent of other factors like age distribution or infrastructure. The fear is that, for instance, differences in age structure predict different developments of Y, or that certain infrastructure changes impact, while neither originate from treatment itself. In the former case you can think of an interaction between a (pre-)condition and time, and in the latter of an exogenous change of conditions over time. These fears can be mitigated by proper conditioning, say by including confounders X. While for identification a common trend, conditional or unconditional, is only required for a given period around treatment, it seems reasonable to assume that this should also hold for the period(s) before the intervention. The same could be said about periods after treatment only if the treatment simply shifts the development of $Y|X$ by a constant (an unnecessary assumption). Again, as in practice we typically look at means (or say, are interested in average treatment effects), all statements about the development of Y or its conditional version $Y|X$ refer simply to the mean. If we are interested in changes in higher-order moments like variance, skewness or kurtosis, we would either directly compare the entire distributions of both groups before and after the treatment or estimate these higher moments. As for the latter, the general procedure and ideas remain basically the same; we concentrate here on the estimation of the average treatment effect on the treated.

1.1 Central Equation

For the considerations above, we focus on the DiD of conditional means

$$\{E[Y_t|x, d_1] - E[Y_t|x, d_0]\} - \{E[Y_{t-1}|x, d_1] - E[Y_{t-1}|x, d_0]\}, \tag{1.1}$$

where we define $E[Y_s|x, d] := E[Y_s|X_s = x, D_s = d]$ for the outcome Y in time s, given conditions x, and belonging to treatment group d. In the literature you may see different notations and orders of terms (taking first the differences inside the same groups and afterwards between). The idea is usually to condition the expectation of Y on the set of confounders X and treatment status D in period s. For simplicity we consider $d \in \{0, 1\}$, i.e., treatment group ($d = 1$) and control group ($d = 0$).

When treatment takes place between periods $t - 1$ and t, expression (1.1) gives the conditional treatment effect on the treated from which we can obtain average effects. Identification of a causal impact of treatment on Y is based on the assumption that without treatment, (1.1) had been zero almost surely for all x of the *common support* defined below (Section 2.1.1). To identify a causal effect, we work with a scale for Y and a set of covariates X such that (1.1) is zero (noting that both choices have consequences for interpretation). Using this statistic can turn a bane into a boon: while it may be difficult to convince others that this assumption is fulfilled, an appropriate statistic can guide you data-adaptively.

For simplicity, we will mostly assume we have data on three time periods $t = -1, 0$ and 1 and consider the case where the treatment occurs between periods 0 and 1. Given that we have data in an additional period to treatment ($t = -1$), we can check if (1.1) is zero for a given X prior to treatment (the development between $t = -1$ and $t = 0$). We emphasize that while this is not the (non-testable) identification condition needed, it empirically supports its credibility.

The DiD expression (1.1) is far more useful than being used to estimate an average treatment effect on the treated (TT). We study its estimation, including heterogeneous TT, its sample average (i.e., the average TT itself), and the analogue of its squares (i.e., test statistics). In each case, we study the asymptotic and finite sample properties. In practice, it is likely preferable to rely on bootstrap methods than on estimates of complex asymptotics, but the latter help to better understand the performance of the statistics. For approximating the p-value of the tests that we will introduce, a challenge is to find procedures that generate data under the null hypothesis.

1.2 What Does it Mean to be Model-Free?

Without covariates, the nonparametric TT estimator reduces to the classic DiD estimator which simply subtracts averages of the observed Y. In this situation, the four means can be estimated without a statistical model; the only model we use is the causality model (i.e., the supposition that the difference of differences would identify the TT). However, when including covariates, which is unavoidable in the presence of confounders, the specification of the mean functions matters. This is also true if we are only interested in the average (over all x) of (1.1) (see also Meyer (1995) for more discussion). Then, in order to avoid a bias due to misspecification, we would prefer avoiding the specification of a statistical model for the mean functions, and use nonparametric estimation instead.[2] The only model we use is the causality model (i.e., the supposition that (1.1) would identify the causal effect). Our procedure is certainly not model-free regarding the causality model; we are only model-free regarding the estimation of (1.1). This way of thinking is somewhat different from the classical econometrics literature on identification as there the identification was largely or fully interwoven with the parametric specification of the structural equations. Here we distinguish between the causality model for identification, and the statistical model for estimation and testing. When the latter is done nonparametrically, we speak of nonparametric identification of causality since it does not depend on the parametric specification.

Nonparametric estimation is often avoided for fear of the curse of dimensionality, its interpretation, implementation or the complex inference (like non-standard calculus of standard errors and p-values). Although the provision of user-friendly software has improved a lot, it is true that in many situations the latter can still be the bottleneck. This is why in this article we also describe the implementation, explain and provide our `R`-code, and discuss issues the practitioner is confronted with. Interpretation will become more involved when exploring heterogeneity of the treatment effects along several covariates simultaneously. Lastly, while the curse of dimensionality can be real, in many situations, it is not an issue. For example, in the presence of only discrete regressors, Ouyang *et al.* (2009) show that the nonparametric conditional expectation estimator is estimated at the parametric (i.e., root-n) rate without asymptotic bias. Unless the number of variables increases with the sample size (and then it is also an issue for parametric estimation), only continuous confounders count for the curse. If the unconditional treatment effect is of interest, you need to have more than three continuous variables to be affected asymptotically. Even then, imposing higher smoothness conditions allows for bias reduction such that we end up with the parametric rate again.[3]

In practice, many variables can be discrete, and many continuous variables are measured or recorded discretely (e.g., years of education). For this reason it is often argued in the

[2] In the econometrics literature, Heckman *et al.* (1997) were perhaps the first who mentioned the non- or semiparametric extension of DiD to include covariates.

[3] Even though this is standard practice in econometric theory, one may criticize that these conditions impose non-testable restrictions. However, they simply exclude discontinuities in derivatives of higher-orders, and it is not clear to what extent a potential oversmoothing of them would affect the final estimates. In any case, those smoothness conditions are far milder than any parametric approach would require.

applied economics literature that parametric methods would be sufficient almost always, as we could construct a saturated model. We will later discuss why this is rarely the case (Section B.1). Therefore we argue that if most applications contain a continuous covariate or discrete ones with many values, nonparametric methods are the better option for causal analysis in most cases. In fact, nonparametric regression is at least as reasonable as a parametric one even when all confounders are discretely measured. Moreover, in our application we show that this also holds true computationally. We should mention here that it is relatively straightforward to employ parametric or semiparametric versions of our methods if desired. However, those strict parametric assumptions may or may not be justified by prior knowledge like economic theory, and a misspecification of functional forms easily leads to biased and inconsistent estimates.

As said, we are not much concerned about a potential curse of dimensionality (see also Section B.1) because the case of facing mainly (or only) discrete regressors is indeed quite common in economics. For example, solely looking at the *American Economic Journal: Applied Economics*, examples include Ang (2019), who looks at the impact of the Supreme Court in 2013 striking down parts of the Voting Rights Act on long-run voter turnout. His model regresses voter turnout (a continuous variable) on year indicators interacted with treatment group dummies, county and state-by-year fixed-effects as well as a dummy for elections that were subject to bilingual requirements in a given year. Panhans (2019) looks for adverse selection in the Affordable Care Act health insurance exchanges. A supplemental section of his paper uses DiD with a set of fixed effects which are not exhaustive and hence are not identical to nonparametric estimates. McKenzie *et al.* (2014) look at migration patterns of Filipinos when there is a binding minimum wage change in the country of origin. They use a host of fixed effects and an indicator for whether or not the individual was a domestic helper. Jayachandran *et al.* (2010) use a host of specifications solely with discrete right-hand-side variables to study the impact of surfa drugs on mortality rates. Regarding our data analysis, Kuka *et al.* (2020) examine human capital responses to the availability of the Deferred Action for Childhood Arrivals (DACA) program. In addition to having all binary right-hand-side variables (some are discrete information transformed to dummies), their outcome variables are binary. Nonetheless, as authors usually have a mix of discrete and continuous variables, we consider this rather general setting, and argue that empirical researchers should be more concerned about systematic biases and inconsistency due to model specification than potential issues with model-free estimation.

1.3 Structure of the Article

The plan is to introduce a complete framework for model-free DiD based causal analysis under the potential presence of confounders. To do so we start by presenting a data-driven procedure to find an appropriate scale of Y with a set of confounders compounded in a vector X that (both together) prove to have some credibility to identify the treatment effects via the 'parallel path'. As this cannot be done for the period of interest itself, we will study the parallel path for previous periods (i.e., not the actual assumption but an indicator for its plausibility). Then we estimate the identified effects on the treated. The procedure is concluded by the introduction of nonparametric tests for significant treatment effects. Modified versions of the simultaneous test for significance of conditional effects can be used for testing heterogeneity of effects or the credibility of identification assumptions.

The next section will provide the analytical developments with technical details which are afterwards completed by simulations. These show the usefulness of all methods even for moderate and small samples. As it is uncommon for nonparametric estimators to be estimated at parametric rates,[4] it is particularly interesting to see their performance with very small samples. We will see that, a bit surprisingly, the performance of our scale and covariate selector, our estimators and tests are admirable, even in these small sample settings. The simulations will be followed by the various issues in practice, namely the discussion of implementation, bandwidth choice, details on bootstrap procedures, presentation of `R` functions, and further miscellaneous.

To highlight usefulness and relevance of our approach, we re-examine the results of Kuka *et al.* (2020) in the last section. We find mixed evidence that their set of confounders satisfy the 'parallel path' assumption. Regarding their treatment effect estimates, their models underestimate the positive impact that DACA had on the rate at which 14-18 year old students stayed in school and the positive impact of DACA on high school completion (either via graduation or obtaining a GED). Moreover, they fail to identify the negative impact of DACA on school attendance of college aged individuals (19-22). With respect to enrolling in college, we can confirm that these effects are insignificant.

Beyond the replication, we also look at hetereogeneity in treatment effects. For example, we find that DACA had a positive and significant impact on the rate at which 14-18 year old male students stayed in school, but an insignificant impact on female students. We also find significant effects only for Hispanic, Black and White students. The impact also increased by age. There was no economically or statistically significant impact for 14 or 15 year olds, but statistically significant and monotonically increasing impacts with age for 16, 17 and 18 year olds. We conclude our application by stating that there are far more questions that should be addressed in this literature beyond an average treatment effect on the treated.

We conclude this introduction with a remark on a recently much discussed inference problem. You could ask about post-selection (or pretesting) inference as we propose a

[4] The logic here is similar to that for (kernel estimated) average derivative estimators (Härdle and Stoker, 1989).

procedure that allows you to select between different covariates and scales of Y, or to test for bias stability before treatment started. However, our problem differs from the post-selection inference typically considered (cf. Rolling and Yang (2014) for the treatment effect estimation context and Kuchibhotla *et al.* (2022) for a general recent review). Intuitively, Taylor and Tibshirani (2015) describe the standard problem as follows: "Having mined a set of data to find potential associations, how do we properly assess the strength of these associations? The fact that we have cherry-picked, i.e., searched for the strongest associations means that we must set a higher bar for declaring significant the associations that we see."

Our criterion is not the covariates contribution to a regression, but the maximization of bias stability (i.e., checking the identifying assumptions necessary for causal conclusions). However, as this is infeasible for the period of interest, it has to be done for a prior period. That is, there is no cherry-picking for significance or finding the strongest treatment effect; we rather try to maximize the conditional independence. Moreover, doing this for periods prior to the one of interest suggests that we apply a strategy similar to sample splitting. Notice also that standard literature on post-selection inference recommends to condition on the applied pretests (calling it selective inference), whereas the literature related to our context advises against such conditioning (Roth, 2022).

2 Model-Free Approach

Our main contribution is the provision of a complete framework for DiD-based causal analysis. As already indicated, the literature on DiD estimation is abundant; when we cite some of the existing econometric literature we limit us mostly to contributions which provide methods with its corresponding asymptotic theory. For a general discussion, recalling ideas, definitions and assumptions of DiD with confounders we refer to Frölich and Sperlich (2019) and Lechner (2011).

For a linear parametric DiD with confounders we would like to recommend Sant'Anna and Zhao (2020) who consider a so-called double robust version (i.e., using propensity score weighting and regression). This is not to be confused with double machine learning or double debiased methods as these are completely different concepts, both designed to tackle problems we don't have. In a fully parametric context, the so-called double robust estimator provides a consistent estimator of the treatment effect if either the propensity score or the regression function is correctly specified. In nonparametric estimation, both are 'correctly specified', and it is not clear if doing both would result in an improvement in practice. It is possible that for some nonparametric estimators (e.g., series estimators), the approximation bias could be substantial (this is why we suggest kernels). In that case mixing propensity score weighting and matching could perhaps result in an improvement. Kennedy *et al.*

(2017) introduced a special nonparametric doubly robust matching estimator for continuous treatment whose extension to DiD might be interesting. However, we are unaware of papers supporting the guess that double robust estimators are more efficient when propensity score and regression function are poorly approximated. We therefore primarily focus our attention on the regression setting.

2.1 Nonparametric Difference-in-Differences with Confounders

Assuming that two groups have an *unconditional* common trend in their responses over a certain period of time might be too strong of a restriction. Abadie (2005) and Qin and Zhang (2008) proposed DiD with nonparametric and semiparametric propensity score weighting, respectively. More recently, Chan *et al.* (2016) proposed a more general weighting scheme for matching, but not explicitly for the DiD estimator. We will see below, for propensity score weighting, the asymptotics for DiD estimation based on nonparametric matching follows from the asymptotics of nonparametric matching without the second difference. As has been discussed in many papers, there is no general superiority of propensity score weighting over matching, and therefore there isn't one for the DiD context either.[1] We stick to the latter (i.e., a DiD regression approach) for different reasons: the first is that practitioners relate conditional treatment effects rather to regression than weighting. Further, we also avoid numerical problems that occur when dividing by nonparametric estimates of potentially small propensities which is a nontrivial advantage in practice. Finally, we prefer not to jump between nonparametric propensity estimation and nonparametric regression. The last point is linked to the inclusion of further covariates.

More specifically, one may also be interested in exploring potential heterogeneity of effects beyond confounders. Recall that confounders are only those variables that partly predict both D and Y. We can obviously regress on those additional covariates (subsumed in X together with the confounders), and may find that (1.1) varies over them; in such a case we call them *solely effect modifiers*. In contrast, conditional on the confounders, the propensity score does not exhibit variation over *solely effect modifiers*. In other words, propensity score weighting alone would not be sufficient to explore effect heterogeneity over effect modifiers that are not confounders.[2] For simplicity, we will not treat here confounders differently from those *solely effect modifiers*; see Benini and Sperlich (2022) for an explicit

[1] Some papers stated the superiority of propensity score weighting as it would avoid the curse of dimensionality since one could estimate the propensity quite well with parametric methods. However, as we divide by its estimates, any error has a leverage effect giving huge errors for the treatment effect estimator even if the error in the propensity estimate was small (Drake, 1993). People have tried to steadily improve on the propensity score estimator, more recently by using machine learning methods (McCaffrey *et al.*, 2004; Lee *et al.*, 2010).

[2] You also may have variables that impact the propensity score but have no further impact on Y. Those, however, you typically do not want to include.

separated modeling approach. In some literature, those additional covariates are called moderators; others define moderators in a way that includes confounders, and may call our X the vector of moderators. For matching, Heckman *et al.* (1997) suggest in their Section 4 to put all confounders in the propensity score, the solely effect modifiers in the regression, and then mix propensity score weighting and regression (not to be mixed up with double robust regression which puts the entire X everywhere). We could construct something similar for nonparametric DiD estimation, but do not see any gain in this, rather the risk of confusion.

As Frölich and Sperlich (2019) discuss, there are further reasons you might want to condition on certain covariates. One is to measure a direct or a partial impact of D on Y, controlling for certain covariates that are impacted by D; another is to include covariates that are not impacted by D, but have predictive power for Y. Their inclusion can improve the statistical analysis by reducing noise. Which covariates to include is seemingly the researcher's choice, but this has implications for both interpretation and assumptions. As we condition on both, confounders and additional covariates, we will henceforth speak of 'covariates' in general. Different from most of the existing econometrics literature, we allow these covariates X to vary over time. Further, we use notation appropriate for cohorts in which t stands for the indicator of the time period when the observation is made. Where appropriate, we will discuss the much simpler case of balanced panel data explicitly.

2.1.1 Difference-in-Differences with Covariates

Before estimating the treatment effects, we should have a closer look at the identification conditions. We want to include a set of covariates X, and need to know the scale of Y, such that stochastically speaking, both the parallel path and a common support condition hold. As we need to assume the common trend for the period in which the treatment takes place, we have to introduce the notion of 'potential outcomes' for Y, where Y^d represents the response that would be obtained if treatment $D = d$ had taken place. We further need to define the domain $\mathcal{X} \subset supp(X)$ which is implicitly determined by the so-called common support condition (CSC) which says that

$$\textbf{CSC} \qquad P(T = 1 \cap D = 1 | X = x, (T, D) \in \{(t, d), (1, 1)\}) > 0$$
$$\forall x \in \mathcal{X}, \forall (t, d) \in \{(0, 0), (1, 0), (0, 1)\}$$

where for the sake of notation, time T is dealt with like a random variable. To be specific, CSC says that domain $\mathcal{X}$ contains only values of X that can be found in each group in each time period. There should be no value in $\mathcal{X}$ whereby we cannot find a counterfactual match.

At first glance, our CSC seems to be more restrictive than other common support conditions given in the literature. This, however, is not the case. It typically collapses to a specific case. This somewhat cumbersome notation has been chosen to allow for unbalanced panels and cohorts. While we will define these more formally after Equation (2.3), we may be interested in the average treatment effect on the treated for the treatment group in time period 1 (TT_a) or the average treatment effect on the treated for the treatment group in time periods 0 and 1 (TT_b).[3] The domain $\mathcal{X}$ does not change over time. It is only related to the fact that we allow for time-varying covariates in the following sense: the CSC we wish to have for estimating TT_a , say $\mathcal{X}_a$, contains exactly all x-values observed in the treatment group at $t = 1$, whereas its counterpart $\mathcal{X}_b$ would in addition contain all x-values observed in the treatment cohort at $t = 0$.

The CSC says little about the underlying distribution of X within each group in each time period. For the confounders these are actually supposed to be different between the treatment and control groups. Generally, one should interpret CSC neither as a restriction nor as an assumption; it simply defines the domain for which you can identify counterfactual treatment effects. The population of interest is then to be (re-)defined such that its support of X is in domain $\mathcal{X}$. For the ease of presentation we will consider as our populations of interest those represented by the people that are in the treatment group (for instance at time point $t = 1$ for TT_a); or we condition on a specific $x \in \mathcal{X}$ when the conditional effect TT_x is of interest. Since in Section 3.1.3 we extend our method to data with sampling weights, you can calculate the average treatment effects for many different (treated) populations by applying weights that correspond to their distributions. All you need is that the support of X is in domain $\mathcal{X}$.

Specifically, for identification of the counterfactual treatment effect of the treated (TT), we need

Assumption I For all $x \in \mathcal{X}$ the difference in potential outcomes under no treatment (Y^0) between the treatment and control group is the same before and after treatment:

$$E\left[Y^0_{t=1}|x,1\right] - E\left[Y^0_{t=1}|x,0\right] = E\left[Y^0_{t=0}|x,1\right] - E\left[Y^0_{t=0}|x,0\right] \ , \tag{2.1}$$

recall expression (1.1) with the explicit definition of our conditional expectations. We would like to emphasize that we condition here, in (1.1) and also in the following, on $X_s = x$ with s corresponding to the respective time index of Y^0, where x are from support $\mathcal{X}$ defined via our CSC above. This implies that identification does not hinge on the potential time-variance of X. The difference to identification conditions you typically find in the literature, lies in (i) that we allow for time-varying confounders and (ii) for other moderators, consequently have (iii) more general conditional expectations in (1.1) and (2.1), and (iv) need a more restrictive CSC when estimating TT_b (or TT_x outside of $\mathcal{X}_a$).

[3] Certainly, both cohorts are expected to be taken from the same population, but their observed values x need not be the same in practice as the distribution of X may change over time. If this is true, TT_a and TT_b have a slightly different interpretations (as in stochastic terms this means the population has changed over time).

We are no longer looking for a parallel path of the potential outcomes Y^0, but of $Y^0|x$, an important distinction when switching from unconditional to conditional DiD. Moreover, (2.1) highlights the link to matching estimators based on a conditional comparison of treatment versus control groups after treatment ($t = 1$). In the matching setting, we assume that the vector X accounts for all differences in Y^0 such that the left-hand-side of (2.1) is zero, but if not, its average over all x is the bias of the well-known TT matching estimator. In the DiD setting we only assume that this difference is the same before treatment, suggesting that we can use pre-treatment data for bias correction. Therefore, calling Assumption I 'bias stability' is perhaps more appropriate as it does not deceptively insinuate a parallel path of Y^0.

We also must be concerned about spillover effects (Lechner, 2011). It is feasible that the outcome of a particular individual is affected by not only their treatment status, but also the treatment status of other individuals. To avoid this possibility of spillover or general equilibrium effects we must employ the stable unit treatment value assumption (SUTVA). This assumption states that the observed outcome of an individual only depends upon the treatment status of that individual and not the allocation of other individuals (Rubin, 1977). This assumption is potentially violated if individuals interact directly or indirectly.

Assumption I is the usual 'non-testable identification condition'. However, as said earlier, it is not very credible if it does not hold (shortly) before treatment as well. Consequently, we could apply this assumption to periods prior to treatment ($t = -1$ and 0) and use data from those periods to evaluate its credibility, which is feasible because for $t < 1$, $Y_t^0 = Y_t$.

For simplifying the notation further, denote the conditional expectations for each year and treatment group by

$$m_{dt}(x) = E[Y|X = x, D = d] \ , \quad d = 0, 1, \ t = -1, 0, 1 \ . \tag{2.2}$$

Obviously, under Assumption I, SUTVA and CSC, the conditional TT for a given x is identified by

$$TT_x = \{m_{11}(x) - m_{01}(x)\} - \{m_{10}(x) - m_{00}(x)\}, \tag{2.3}$$

and consequently also the unconditional TT for any (sub-)population, by integrating out x accordingly. Let n_{dt} denote the number of observations in group d at time t, and suppose that all n_{dt} converge at the same rate to infinity. Further, denote TT_a as the TT that results from integrating TT_x over the distribution of x in the group containing the members of the treatment group $D = 1$ observed in $t = 1$. We will also provide most of the formulae and asymptotics for the TT that results from integrating over all individuals of the treatment groups, no matter if observed in $t = 1$ or in $t = 0$, and denote this parameter by TT_b.

We will speak of unconditional TT when we refer to both: TT_a and TT_b. Recall that we do not require a balanced panel. We therefore would not typically encounter all X for all people at all time points. The observed x_{it} refers to time point t. The often applied strategy to use only X-values from $t = 0$ to avoid that treatment effects get mediated via such a covariate is typically not feasible when cohorts replace balanced panels. This is the main reason why we allow for time varying covariates but do not require them. All our methods and results are applicable to the simpler case of balanced or unbalanced panels, as well as to the case of time invariant X. Unfortunately, assuming a balanced panel and/or time-invariant X from the onset does not lead to equivalent results for repeated cross-sections, but it simplifies the asymptotics (as will be shown). For balanced panels with time-invariant X (or say, always using the X values observed before treatment), TT_a and TT_b are the same.

2.1.2 Causality Graphs

It is often helpful to visualize the problem. Here we consider three Directed Acyclic Graphs (DAGs). The first graph will be to illustrate the general idea and model. The latter show identification and generality of DiD estimators for more complex structures, respectively.

To set the stage, recall that treatment occurs between $t-1$ and t (between periods 0 and 1 in our setting). Similar to above, our outcome is given as Y and recorded covariates as X. We add the unobserved random terms U and E acting on Y and on treatment D. In these graphs, D indicates the treatment itself and not the affiliation to the treatment vs control group. Since the treatment happens only once, there is no need for a time index. In addition, we introduce characteristics C that could cause a bias in matching which is supposed to remain stable over time (Assumption I) and can therefore be controlled for by differencing. For this reason C is typically thought to be time invariant. Finally, we introduce potentially unobserved variables W when highlighting the generality and power of the method. Regarding the time indices of Y, X and U we follow the standard notation in panel data econometrics. This does not exclude the possibility that X_t contains or is composed of values already observed before t. Moreover, as is common in panel econometrics, while Y, X, U show the same time index, the regression equations as well as our graph suggests that X and U come first (same for D), impacting Y.

Consider our first causality graph (Figure 2.1). In this graph the arrow from X_t to D is painted in gray to show that not all X are supposed to be confounders. At the same time we are not saying that no further arrows are allowed or could be skipped (like for instance those between the Y). While this graph illustrates the general idea and model, it is not sufficient for identification (what DAGs are commonly used for). For instance, it tells us that C acts as a confounder. But since we don't observe C, at least not completely, we can't condition on it. Instead, thanks to the time dimension we can apply differencing which in turn is only sufficient for identification under bias stability, our Assumption I (2.1).

What would a DAG look like to be sufficient for identification? Recall that we use DAGs to show in the simple matching context that $Y^d|X = x$ is mean-independent

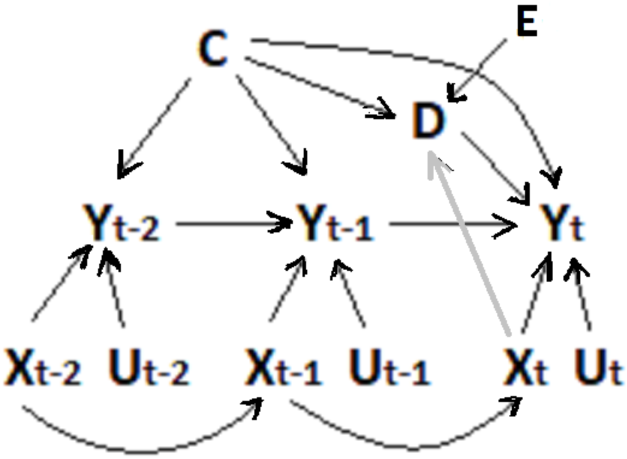

Figure 2.1: A dynamic DAG for illustration.

of D, where for estimating the TT this is sufficient to hold for Y^0. This allows us to identify the potential mean of $Y^0|X = x$ for the treated, i.e., $E[Y^0|x, D = 1]$ for all x of the common support. For a moment let us consider the case where X is time invariant. Then, demanding mean-independence of $\Delta Y_t^0|X = x$ from D with $\Delta Y_t^0 := Y_t^0 - Y_{t-1}^0$ in our context, allows us to identify the analogue to matching, i.e., $E[\Delta Y_t^0|x, D = 1]$. Like in the simpler matching case where the conditional treatment effect is given by $E[Y|x, D = 1] - E[Y|x, D = 0]$, it is now given by $E[\Delta Y_t|x, D = 1] - E[\Delta Y_t|X, D = 0]$. If simple matching was already sufficient, the bias correction was redundant and we get the same result, as then $E[Y_{t-1}|x, D = 1] - E[Y_{t-1}|x, D = 0] = 0$, i.e., we just subtracted zero.

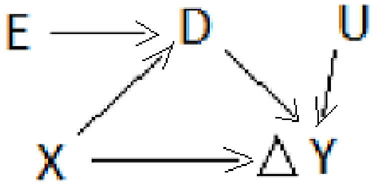

Figure 2.2: A reduced DAG for identification.

Conditional mean-independence of $\Delta Y_t^0|x$ from D gives our Assumption I, and the corresponding DAG is given in Figure 2.2. There we skipped the time-index of ΔY because we only considered the difference between Y_t and Y_{t-1}. What happens if we switch from cases where X is time invariant to those where (at least part of) X is time variant? We then ask D to be (mean-)independent from the difference $(Y_t^0|X_t = x) - (Y_{t-1}^0|X_{t-1} = x)$ which is the needed extension of $\Delta Y_t^0|X = x$ from above. As in our graph X can also stand for (X_t, X_{t-1}), we only need to adjust the meaning of ΔY in our DAG.

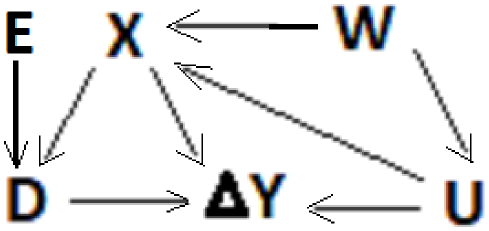

Figure 2.3: Example of a possible extension of our DAG in Figure 2.2.

The third graph (Figure 2.3) is only included in order to demonstrate that a DiD estimator allows for much more complex structures, including endogenous confounders X. It is not hard to see that deleting all arrows emerging from D, node X blocks all paths between D and ΔY giving the required conditional independence to obtain identification.

We conclude this section with the reminder that no matter which x you condition on, or over what distribution of X you integrate, the DiD as it is introduced above only identifies the (conditional) treatment effect of the treated; identification of treatment effects for the non-treated needs additional non-testable assumptions that are not attractive in typical DiD applications (Lechner, 2011).

2.1.3 Nonparametric Conditional Expectations

Most empirical papers use linear panel data methods to estimate the TT. While the linear specification without covariates is equivalent to the method derived via conditional expectations, there is no such result here (Meyer, 1995). Even if we had only discrete X which we could decompose into dummies; a saturated linear model would require to include all these dummies together with all interactions of different orders, see also Section B in the Appendix.[4] Most practical work doesn't do this properly; instead, such inclusion is usually arbitrary, guided by numerical convenience. Moreover, if at least one covariate is continuous or discrete with many values, this problem is heavily aggravated. Nonparametric methods remove these concerns. Practitioners often ignore the use of these methods and use the curse of dimensionality as their argument against them. Yet, in most common settings, the curse of dimensionality is not an issue as only very few of the covariates are actually continuously measured. This is even more so for DiD compared to all competitors for nonexperimental data, as the differencing already accounts for many counfounders.

[4] The common practice of splitting the sample to obtain heterogeneous estimates in the parametric world is valid assuming the functional form is correct and there is a sufficient number of observations in each cluster. This practice addresses parameter heterogeneity; it does not cure functional form misspecification.

Now, suppose the scale of Y and the set of covariates X are given. In a first step, for each group d and each time point t, we can estimate their mean functions $m_{dt}(x)$ from the data set $\{Y_{it}, X_{it}\}_{i=1}^{n_{dt}} | D_{it} = d$. Let us split the vector of covariates X_{it} into a vector with p continuous variables entering the smoother, say $X_{it}^s = (X_{it,1}^s, ..., X_{it,p}^s)$ and another vector with k categorical variables $X_{it}^c = (X_{it,1}^c, ..., X_{it,k}^c)$. We use a multiplicative kernel $K(X_i, x, h, \lambda) = W(X_i^s, x^s, h) \cdot \lambda^{d_{X_i,x}}$ where $d_{X_i,x} := \sum_{l=1}^{k} 1\!\!1\{X_{it,l}^c \neq x_l^c\}$ and W a product of p univariate continuous kernels $w\{(X_{it,l}^s - x_l^s)h^{-1}\}h^{-1}$, $l = 1, ..., p$, where $h \geq 0$ and $\lambda \in [0, 1]$ are our bandwidths.[5] Function $1\!\!1\{A\}$ is equal to 1 if the event A is true, and zero otherwise. Under standard regularization conditions outlined in Ouyang *et al.* (2009) (cf. also Li *et al.* (2009) for propensity score weighting), namely on the smoothness of $m_{dt}(\cdot)$ and density $f_{dt}(\cdot)$ of X^s in group d at time t, for $\lambda, h \to 0$ when $n_{dt} \to \infty$, we have

$$\sqrt{n_{dt}h^p}\,\{\widehat{m}_{dt}(x) - m_{dt}(x) - B_{dt}(x, h, \lambda)\} \to N(0, \Omega_{dt}(x)) \tag{2.4}$$

where the conditional mean estimator, given by

$$\widehat{m}_{dt}(x) = \sum_{i=1}^{n_{dt}} K(X_i, x, h, \lambda) Y_i \;/\; \sum_{i=1}^{n_{dt}} K(X_i, x, h, \lambda) \tag{2.5}$$

is the local-constant least-squares estimator where x^s is an interior point of X^s. For boundary points, we need to take boundary kernels to achieve this rate.

The convergence rate, and thereby the curse of dimensionality, is only affected by the continuous covariates (even though we smooth discrete covariates), without imposing any separability structure between continuous and discrete covariates. Unless $\lambda = 0$, this does not correspond to sample splitting, but can be more efficient in practice. For the univariate kernels $w(\cdot)$ defined above, the resulting bias of this mixed smoothing equals

$$\begin{aligned} B_{dt}(x, h, \lambda) &= h^2 \left[\nabla^t m_{dt}(x) \nabla f_{dt}(x) f_{dt}^{-1}(x) + tr\{\nabla^2 m_{dt}(x)\}\right] \\ &\qquad \times \int w(u) u^2 du \\ &\quad + \lambda \sum_{\tilde{x}, d_{\tilde{x},x}=1} \{m_{dt}(x^s, \tilde{x}^c) - m_{dt}(x)\}\, f(x^s, \tilde{x}^c) f_{dt}^{-1}(x) \end{aligned} \tag{2.6}$$

$$\text{and} \quad \Omega_{dt}(x) = Var(Y|x, D = d, T = t) \int w^2(v) dv \; f_{dt}^{-1}(x), \tag{2.7}$$

where $\nabla\mu(x)$ denotes the p-dimensional vector of first derivatives of the function $\mu(\cdot)$ with respect to the continuous covariates x^s, and ∇^2 is the corresponding Hessian matrix.

[5] The notation for the bandwidths h and λ are distinct because of the asymptotic properties for continuous vs discrete variables. Note that $\lambda = 0$ leads to an indicator function and $\lambda = 1$ to a uniform weight function. We do not have a second set of bandwidths (just one per covariate). In practice, we can use a separate bandwidth (i.e., h_l, λ_k) for each covariate. For notational convenience we treat them as equal in our formulae (h_l $\forall l$, and $\lambda_k = \lambda$ $\forall k$). More general extensions are straightforward.

Equation (2.4) shows that only the number of continuous covariates (p) impedes the parametric rate (root-n) of convergence. By using local-polynomials for the continuous covariates, we could achieve a faster rate for the bias (h^2) as long as we are willing to accept higher smoothness conditions on $m_{dt}(\cdot)$ and the densities of the continuous covariates. Although this is standard in the semiparametric econometrics literature, especially in the context of sieve estimators, we abstain from those tricks and concentrate on practical issues. In our application, all of our covariates are discrete and hence a local-polynomial estimator is infeasible.

2.2 Covariates and Scale

Before estimating the TT, we first need to decide on the set of covariates, and the scale of the response Y. Even if prior knowledge like intuition or economic theory (often called 'expert' or 'domain' knowledge) may tell you assuming a common trend is sensible, it does not necessarily tell you the right scale of Y nor the right set of covariates $X^S \subseteq X$ for which it holds. What is often criticized as the bane of DiD estimation, we suggest here to turn into a boon. We use such prior knowledge to help specify the causality model, but allow the data to drive the set of confounders, scale of Y and the form of the conditional expectations. For the ease of presentation, we only consider TT_a; modifications for TT_b and TT_x are mostly evident.

The scale of Y matters a lot because, if the common trend (2.1) holds for one scale of Y, it can hold for affine, but not for nonlinear transformations (e.g., for convex and concave transformation this becomes evident by Jensen's Inequality). Yet, the scale of Y is clearly irrelevant for Assumption I if there is no trend or if there is no selection bias (i.e., both sides of (2.1) are zero); see also Roth and Sant'Anna (2023) for a formal proof and a discussion of time invariant mixtures of the latter cases. For all other situations, the scale is important. Unless the researcher has a strong opinion about it, this could be chosen data-adaptively. The covariates are often driven by reasons of total versus direct (or partial) TT estimation (i.e., filtering out certain indirect effects), the reduction of noise, and Assumption I. While the first is fully up to the researcher's interest, the second should be limited to a few cases (due to its implications for interpretation), the third could be done data-adaptively.

Although we propose a feasible, computationally inexpensive procedure, both choice problems are theoretically intertwined. Therefore, the data-adaptive choices should be based on the same objective function and be considered as a simultaneous problem. Note first that different sets of covariates may define different common supports, and the data-adaptiveness of the proper CSC is straightforward. The objective is to comply with Assumption I. As all non-treatment outcomes Y^0 are observed only prior to the treatment, we consider periods prior to treatment, i.e.

$$\frac{1}{n_{1\bullet}} \sum_{i:D_{i\bullet}=1}^{n_{1\bullet}} \Big\{ m_{1t}(x_{i\bullet}) - m_{0t}(x_{i\bullet}) - m_{1(t-1)}(x_{i\bullet}) + m_{0(t-1)}(x_{i\bullet}) \Big\}^2 \tag{2.8}$$

for $t < 1$, where the summation is over treated individuals in time period t (i.e., $n_{1\bullet} = n_{dt}$, $D_{i\bullet} = D_{it}$ and $x_{i\bullet} = x_{it}$ for a fixed t) if we are interested in TT_a (similarly for TT_b, $n_{1\bullet} = n_{1t} + n_{1(t-1)}$, with $D_{i\bullet} = D_i$ and $x_{i\bullet} = x_i$ running over both periods). Here $m(\cdot)$ refers to the conditional expectation of a potential transformation of Y, conditioned on different subsets x^S of the potential set of covariates. We could choose a transformation and covariates that minimize (2.8). Alternatively, we could likewise integrate (2.8) over the x_{i1} of the treated in $t = 1$.

As discussed, (2.8) does not fully correspond to Assumption I, it only gives credibility to it. This is why we speak of evaluation, not testing. It also has little to do with the typical variable selection problem, especially popular in linear treatment effect estimation with LASSO. The target in that literature is efficient estimation in linear high-dimensional models, while identification is already taken as granted, and the objective function is a penalized least squares or moment condition for estimation. It has nothing to do with our objective or procedure. Moreover, our above objective function is different from the one we use for estimation, and as such, popular procedures for debiasing or post-selection inference have no meaning here. Following Kuchibhotla *et al.* (2022), the only feasible way we see here for addressing the post-selection problem is to perform an analogue to sample-splitting; either to use $t < 0$ in (2.8), or to split the samples of time point 0 when using $t = 0$ in (2.8). In the case of facing panel data one still needs some orthogonality assumption for the residuals. Potential auto-correlation in the residuals destroys selective inference in this situation (cf. Roth, 2022). In our conclusions we discuss bootstrap based inference that could account for the variability of our entire procedure. In practice, instead of doing post-selection or selective inference, we could do a robustness check by performing estimation and testing not only for the best scale-and-covariates combination found in the prior-to-treatment periods, but also for the second and third best.

2.2.1 Data-Driven Evaluation of Potential Scales

Finding a strictly monotone transformation of Y that fulfills (2.1) corresponds to finding a proper scale. Consequently, this scale should provide a reasonable interpretation. As mentioned above, unless you face one of the trivial solutions (no trend or no difference between groups before treatment) in which the scale of Y is irrelevant for (2.1), asymptotically that transformation is unique. We say ‘asymptotically’ as for finite samples this does not need to be the case. In practice, this is not an issue as for interpretation as we would only compare two to four clearly different scales. You may think of the Box-Cox transformation which depends on a parameter θ giving $Y(\theta)$ but you only consider $\theta \in \{0, 0.5, 1\}$, from a set Θ. For each set x^S of covariates, there exists a parameter value θ^S_{opt} that optimizes the common trend condition. Clearly, (2.8) looks at the squared deviations from Assumption I in a prior period, and can thus be understood as a measure of variation. Since variations are scale dependent, we need to adapt the criterion by accounting for the variance of $Y(\theta)$, and define for any given S and a fixed $t < 1$, the optimal transformation parameter for Y by

$$\begin{aligned}\theta^S_{opt} \quad &= \quad \underset{\theta\,\in\,\Theta}{argmin}\, \frac{1}{n_{1\bullet}} \sum_{i:D_{i\bullet}=1}^{n_{1\bullet}} \Big\{ \widehat{m}_{1t}(x^S_{i\bullet}) - \widehat{m}_{0t}(x^S_{i\bullet}) \\ &\qquad - \widehat{m}_{1(t-1)}(x^S_{i\bullet}) + \widehat{m}_{0(t-1)}(x^S_{i\bullet}) \Big\}^2 \widehat{Var_{\bullet}}^{-1}[Y(\theta)], \qquad (2.9)\end{aligned}$$

where $\widehat{Var_{\bullet}}[Y(\theta)]$ is a standard estimator of the unconditional variance of the transformed responses. As for $n_{1\bullet}$ and $D_{i\bullet}$, the $\bullet$ indicates if this variance refers to the (sub-)population of all subjects belonging to the treatment group or only the treated in t.

As nonparametric conditional expectation estimators depend on bandwidths, it is worth mentioning that for this step, we do not need optimal bandwidths for each θ. It is sufficient to have a bandwidth for which the selection outcome along the above criterion does not importantly change compared to the outcome based on an optimal bandwidth. This statement can hardly be defined more precisely due to different uncertainties we face, including the variance of various estimators, and the question of how we define 'optimal bandwidth' in our context. In practice, we ask that for the grid of values over which we search for θ, our working bandwidth picks the same θ^S_{opt} (or a very similar one) as the optimal bandwidth would. We suggest using computationally attractive plug-in bandwidths (see Henderson and Parmeter, 2015 and Chu *et al.*, 2015). For small samples, these tend to slightly oversmooth what would stabilize the numerical performance of the selection procedure. You should not search for the optimal bandwidth using a criterion like (2.8) or (2.9) as these criteria are supposed to be based on reasonable estimates of $\widehat{m}(\cdot)$, but not vice-versa.

2.2.2 Data-Driven Evaluation of Confounder Sets

While prior knowledge helps clarify which covariates to include, data-driven methods can help guide us by choosing credible sets. In practice, the choice of X in an academic paper typically relies on many dummy variables to attempt to control for all the biases a referee would consider feasible. This means, their inclusion is neither due to a clear data generating or causal model, nor on considerations of total versus partial impact measurement.

It is often argued that the covariates should not be impacted themselves by the treatment, and therefore, only time invariant covariates are considered, or only values of X observed before treatment. In other fields, people are interested in direct effects (or to control for some specific indirect effects), and therefore include certain covariates because they are affected by treatment. So both, the set of covariates you want to include, as well as the set of potential confounders you allow for, depend on the parameter of interest. The correct interpretation hinges on your assumptions. These must be consistent with your data, and your interpretation with these assumptions (Kahn-Lang and Lang, 2019). This implies you may not want to allow for any combination of covariates; instead you prefix a set $\mathcal{S}$ of covariate sets S from which you wish to choose the most appropriate one(s). We should not think here of a step-wise elimination of covariates but of a ranking of all eligible sets regarding credibility. Then, for the θ^S_{opt} from above,

$$\begin{aligned} S_{opt} \quad = \quad & \underset{S\in\mathcal{S}}{argmin}\, \frac{1}{n_{1\bullet}} \sum_{i:D_{i\bullet}=1}^{n_{1\bullet}} \Big\{ \widehat{m}_{1t}(x^S_{i\bullet}) - \widehat{m}_{0t}(x^S_{i\bullet}) \\ & -\widehat{m}_{1(t-1)}(x^S_{i\bullet}) + \widehat{m}_{0(t-1)}(x^S_{i\bullet}) \Big\}^2 \widehat{Var_\bullet}^{-1}[Y(\theta^S_{opt})] \end{aligned} \qquad (2.10)$$

defines the optimal set along the analogue to (2.9), i.e., you jointly calculate the same criterion for all (θ, S) combinations to obtain $(\theta^{S_{opt}}_{opt}, S_{opt})$ which is the most credible regarding the DiD identifiability assumption. In practice these may not be unique for a given data set; then the practitioner may try all those optimal pairs but should keep in mind that they may define somewhat different treatment effects.

If initially there are too many sets, we can even perform pre-selection procedures. A simple method is a visual check to see to what extent a covariate could be a confounder. When plotting the distribution of a potential confounder per group and time period, these should exhibit different features between groups; otherwise they are not confounders. Certainly, pre-selection could also be based on variable selection in regression; if they exhibit absolutely no impact on Y, they are of no use. In the context of nonparametric estimation, however, those procedures are more complex than directly applying (2.10) (Hall *et al.*, 2007). Moreover, these selection procedures might be based on objective functions different from minimizing the deviations in (2.8). Generally we would advise against mixing different objective functions, especially where the objective is essentially the same.

In practice, we suggest using a penalty factor to account for too many covariates. We tried several alternatives, but found that a simple AIC factor worked well in simulations. Considering our criterion in (2.10), we propose to add

$$\Big(2(k+p)^2 + 2(k+p)\Big) \,/\, (n_{1\bullet} - (k+p)) \,, \qquad (2.11)$$

to penalize against including too many covariates. In our simulations, this factor helps to correctly identify models with irrelevant covariates even for small samples.

It is possible to formally conduct a nonparametric significance test to see if Assumption I is rejected for any given pair (θ, S) for the period before treatment. In practice, you would test this for the “optimal set” or the one you favor (e.g., for interpretational reasons). This can be done by taking (2.8) as a test statistic (for $t = 0$) as will be shown in Section 2.4. However, recall that you can only test the credibility of Assumption I, not the assumption itself. Note also that applying this idea to post-treatment periods is questionable (cf., Kahn-Lang and Lang, 2019).

2.3 Treatment Effect Estimators

In this section we first consider heterogeneous (conditional) treatment effects on the treated for both repeated cross-sections and balanced panels. We then turn to average (unconditional) treatment effects on the treated. Asymptotic results are discussed in each setting.

2.3.1 Conditional Treatment Effect on the Treated

To keep notation simple, let henceforth Y and X denote the adequately scaled response and the chosen covariates. Define the DiD estimators of the *conditional* TT (also known as CATET) for $x \in \mathcal{X}$

$$\widehat{TT}_x = \{\widehat{m}_{11}(x) - \widehat{m}_{01}(x)\} - \{\widehat{m}_{10}(x) - \widehat{m}_{00}(x)\}\,. \tag{2.12}$$

Recalling Section 2.1.3, we immediately obtain for this estimator

Proposition 2.3.1. Under the assumptions (A1) and (A2) of Racine and Li (2004), extended to the four groups, and assuming independence of errors $u_{it} := Y_{it} - m_{dt}(X_{it})$ for all groups, for all x being interior points for each group, estimators of $\widehat{m}_{dt}(x)$, for each combination of $d = 0, 1$ and $t = 0, 1$ will be independent for any x from the common support as well. $\widehat{TT}_x$ has a smoothing bias which is the difference of differences of the corresponding individual biases given in (2.6), i.e.,

$$\{B_{11}(x) - B_{01}(x)\} - \{B_{10}(x) - B_{00}(x)\}\,. \tag{2.13}$$

Similarly, its asymptotic variances are the sum of their asymptotic variances, i.e.,

$$\Omega_{11}(x)/(n_{11}h_{11}^p) + \Omega_{01}(x)/(n_{01}h_{01}^p) + \Omega_{10}(x)/(n_{10}h_{10}^p) + \Omega_{00}(x)/(n_{00}h_{00}^p).$$

The biases and variances resulting from the smallest n_{dt} will dominate the others. Following (2.4), $\widehat{TT}_x$ converges at this rate to a normal distribution.

It is well known that the assumptions (kernel, smoothness or other regularity) could be modified, but for simplicity, we stick with the work of Racine and Li (2004). We allow each bias term to have its own set of bandwidths (h_{dt}, λ_{dt}). As sign and smoothness of the $m_{dt}(\cdot)$ should not change over d and t, equation (2.6) suggests that the differencing has not only a bias reducing effect regarding identification (i.e., a potential specification bias), but also regarding smoothing.

In the popular setting of balanced panels and conditioning only on covariates from $t = 0$ with notation $n^d = n_{d1} = n_{d0}$, assuming the independence $u_{i0} \perp u_{i1}$ for all i becomes less credible.[6] The asymptotics simplify nonetheless, as now we have for $d = 0, 1$

$$\widehat{m}_{d1}(x) - \widehat{m}_{d0}(x) = \frac{\sum_{D_i=d:i=1}^{n^d} K(X_{i0}, x, h_d, \lambda_d) \ (Y_{i1} - Y_{i0})}{\sum_{D_i=d:i=1}^{n^d} K(X_{i0}, x, h_d, \lambda_d)}. \tag{2.14}$$

Corollary 2.3.1. For balanced panels with $\tilde{\sigma}_d^2(x) = Var(u_{i1} - u_{i0}|X_{i0} = x, D = d)$, and conditioning only on covariate values observed in $t = 0$, allowing for heteroskedastic autocorrelation in u_{it}, but else the same assumptions as in Proposition 2.3.1, the bias expression for $\widehat{TT}_x$ remains the same, whereas its variance simplifies to

$$\tilde{\sigma}_1^2(x)/(n^1 h_1^p) \int w^2(v)dv \ f_{10}^{-1}(x) + \tilde{\sigma}_0^2(x)/(n^0 h_0^p) \int w^2(v)dv \ f_{00}^{-1}(x) \ .$$

As we will see below, the asymptotics of the unconditional TT are not straightforward. But even for this simpler case of conditional TT estimation, in practice, no one would try to estimate the bias and variance of $\widehat{TT}_x$, especially not for all potential x. Even the estimation of the variance of $\widehat{TT}_a$ or $\widehat{TT}_b$ can hardly be recommended. Instead, we recommend to use a wild bootstrap procedure. Furthermore, we would suggest to use slightly undersmoothing bandwidths to reduce the smoothing bias(es) and concentrate on the estimation of the variance(s).

Before we turn to the unconditional treatment effects, it is worth recalling two points. First, looking at conditional treatment effects may be the most insightful way to study (potential) heterogeneity of treatment effects. Consequently, the above results are not just an intermediate step for the 'popular result'. Second, in the next subsection, we directly integrate over the vector of covariates x to obtain TT_a and TT_b. To further explore the heterogeneity of treatment effects, you may integrate only over a subvector of x, say x_1 with $x := (x_1, x_2)$, to study the heterogeneity over different groups defined by x_2. For example, if x_2 is a binary variable for sex assigned at birth, you obtain TT_{x_2} to study the respective TT for males and females separately (as we will do in Section 4.3).

2.3.2 Unconditional Treatment Effect on the Treated

Given the estimator in (2.12), it is straightforward to obtain a model-free DiD estimator for the unconditional TT by integrating (i.e., averaging over the estimates of) $\widehat{TT}_x$. For the sake of brevity, we consider TT_a, estimated by averaging over the n_{11} observations in group $d = 1$ at time period $t = 1$: supposing that all $x_{i1} \in \mathcal{X}$, i.e., assuming $\mathcal{X}_a$, we have

$$\widehat{TT}_a = \frac{1}{n_{11}} \sum_{i:D_{i1}=1}^{n_{11}} \Big\{ \widehat{m}_{11}(x_{i1}) - \widehat{m}_{01}(x_{i1}) - \widehat{m}_{10}(x_{i1}) + \widehat{m}_{00}(x_{i1}) \Big\}. \tag{2.15}$$

[6] In the case of repeated cross-sections, we typically observe u_{i0} and u_{j1}, where $i \neq j$. In other words, dependencies in errors over time are unlikely.

In a balanced panel when all covariates X_{it} are kept fixed over time the distinction between TT_a and TT_b becomes meaningless. The analogue to TT_b would then be to average over each treated x_i treated twice, which does not make much sense. Recall that this situation does not imply that these characteristics X_i are indeed time invariant, but that one only considers x-values observed at $t = 0$ (i.e., before treatment).

At this stage, it is worthwhile recalling the common support condition. In practice, this is achieved for the continuous covariates by redefining the population of interest such that CSC is fulfilled, which typically corresponds to trimming at the boundaries. This is convenient for other reasons, like avoiding the necessity of boundary corrections for the estimator $\widehat{m}_{dt}(x)$. To avoid complicating our formulas, and in abuse of the above notation let us suppose that in (2.15), we only average over interior points. We are aware that taking it strictly, this supposition together with the notation contains a contradiction as at least for estimating $\widehat{m}_{11}(\cdot)$ some of them will be boundary points. In practice there are at least three alternative options: (i) using boundary correction (via boundary kernels or local polynomials) for the continuous boundary points, (ii) skipping the boundary points from $\mathcal{X}$, (iii) ignoring the boundary problem as the wild bootstrap samples will contain exactly the same boundary points and you stick to bootstrap inference anyway.

For the asymptotics, we refer to the fact that in case of independent residuals, statistic (2.15) can be viewed as an extension of the kernel based matching estimator. It is feasible then to replicate the calculations for nonparametric matching estimators in the existing literature to obtain the bias and variance, and invoke the central limit theorem. The convergence of $\widehat{m}_{dt}(x)$ implies we can choose λ_{dt} and h_{dt} for $dim(X^s) = p \leq 3$ such that $B = o(n_{dt}^{-1/2})$ and $\sqrt{n_{dt}h_{dt}^p} = o(1)$. To achieve this for more than three continuous covariates, we could invoke higher-order kernels or local-polynomial estimators, both based on higher-order smoothness assumptions for $m_{dt}(\cdot)$ and the distributions of X.[7] Asymptotically, for $dim(X^c) = k$, we have no such restriction unless k increases with the sample size.

Proposition 2.3.2. For $p \leq 3$ such that h_{dt}^4 and $n_{dt}^{-2}h_{dt}^{-p}$ are of order $o(n_{11}^{-1})$ we obtain for the TT_a estimator $\frac{1}{n_{11}}\sum_{i:D_{i1}=1}^{n_{11}} \widehat{TT}_{X_{i1}}$ (to emphasize also the randomness of the sample) and common support $\mathcal{X}_a$

$$\sqrt{n_{11}}\left\{\frac{1}{n_{11}}\sum_{i:D_{i1}=1}^{n_{11}} \widehat{TT}_{X_{i1}} - TT_a\right\} \to N(0, V_a), \tag{2.16}$$

where for $\kappa_{dt} = \lim(n_{dt}/n_{11})$ and $\sigma_{dt}^2(x) = Var[Y|x, D = d, T = t]$

$$V_a = E\left[\{m_{11}(X) - m_{10}(X) - m_{01}(X) + m_{00}(X) - TT_a\}^2 | D = T = 1\right]$$

[7] As we mentioned in the introduction, this is not a restrictive assumption for many data sets. In econometrics, when sieve estimators are used, those higher-order smoothness assumptions are standard.

$$
\begin{aligned}
&+E\left[\sigma^2_{11}(X)|D=T=1\right]+E\left[\frac{\sigma^2_{10}(X)f^2_{11}(X)}{\kappa_{10}f^2_{10}(X)}|D=1-T=1\right]\\
&+E\left[\frac{\sigma^2_{01}(X)f^2_{11}(X)}{\kappa_{01}f^2_{01}(X)}|D=1-T=0\right]\\
&+E\left[\frac{\sigma^2_{00}(X)f^2_{11}(X)}{\kappa_{00}f^2_{00}(X)}|D=T=0\right],
\end{aligned}
\tag{2.17}
$$

where each $f_{dt}(\cdot)$ stands for the density of X in group d at time t.

Uniform rates of convergence could be obtained by following results similar to Racine and Li (2004). In the Appendix A.2 we give the influence functions (IF) for $\widehat{TT}_a$ and $\widehat{TT}_b$ to derive a seemingly simpler though equivalent presentation of the variance of $\widehat{TT}_a$ and to provide the asymptotic variance of $\widehat{TT}_b$. The variance expression in Proposition 2.3.2 we find, however, more informative as it explicitly shows where the five parts of the variance come from. Notice that V_a is not the asymptotic first-order variance of the estimator itself but of $\sqrt{n_{11}}\widehat{TT}_a$. In Appendix A.2 we will see that for $D \perp T|X$ which is true for the case where X does not change over time,[8] the resulting simplified variances meet the efficiency bounds given in Sant'Anna and Zhao (2020), though in a quite different context (fully parametric doubly robust DiD estimation for time invariant X, where $D \perp T$, and $D \perp T|X$).

We can also give a simpler variance expression for balanced panels with time invariant X_i using the same notation as in Corollary 2.3.1. Define $\tilde{\kappa}_0 = \lim(n^0/n^1)$ with $n^d = n_{d0} = n_{d1}$ as above, and denote our unconditional TT in balanced panels with covariate values only taken from $t=0$ by $\widetilde{TT}$, then we have

Corollary 2.3.2. For balanced panels with $\tilde{\sigma}^2_d(x)$ as in Corollary 2.3.1, conditioning only on covariate values observed in $t=0$, allowing for heteroskedastic autocorrelation in u_{it}, but else the same assumptions as in Proposition 2.3.2, the variance of the TT estimator simplifies to

$$
\begin{aligned}
&\frac{1}{n^1}\Big\{E\left[\{m_{11}(X)-m_{10}(X)-m_{01}(X)+m_{00}(X)-\widetilde{TT}\}^2|D=1\right]\\
&+E\left[\tilde{\sigma}^2_1(X)|D=1\right]+E\left[\frac{\tilde{\sigma}^2_0(X)f^2_{10}(X)}{\tilde{\kappa}_0 f^2_{00}(X)}|D=0\right]\Big\}.
\end{aligned}
\tag{2.18}
$$

In practice we prefer to estimate the variance of these TT estimators via a wild bootstrap procedure. Given the existing literature, the bootstrap inference for these estimators is relatively easy, so we defer it to the next section. For simplicity, we subsequently assume CSC holds, i.e., we have at least $\mathcal{X}_a$ when speaking of TT_a, etc.

[8] Taking group assignment or time as given, this notation could be confusing. What $D \perp T|X$ means is that even if you know the distributions of X for all periods and groups, D is not helpful for guessing the current time period. If all X values are from period $t=0$, then this is true for sure, but else we don't know.

2.4 Testing

To complete the cycle of a DiD analysis, we consider several testing problems that can be of interest in this context. We first discuss how to test in general for significance of an unconditional or a particular conditional treatment effect. Then we introduce a nonparametric test statistic that can be used for checking different interesting questions: first we use it to jointly test for the significance of conditional treatment effects, then we propose it for checking if treatment effect heterogeneity is large, and finally we discuss how it could also be used for supporting bias stability: Assumption I (recall Section 2.2.2).

2.4.1 Significance of Treatment Effects

Consider the null hypothesis

$$H_0^z : \ TT_z = 0 \qquad vs. \qquad H_1^z : \ TT_z \neq 0 \ . \tag{2.19}$$

This can either refer to a test for significant unconditional treatment effects TT_z of type $z = a$ or b, or it can refer to a significance test for a specific conditional treatment effect TT_x for which x is given. In either case we suggest to construct a wild bootstrap $(1-\alpha)\%$ confidence interval for the respective estimate $\widehat{TT}_z$ to see if zero is included or not; if not, the H_0 from (2.19) can be rejected at level α. You may alternatively consult Proposition 2.3.2 and Proposition 2.3.1 respectively to evaluate the option to estimate a confidence interval directly without a bootstrap. This is only recommended for large data sets.

2.4.2 Composite Significance Testing

The following three testing problems are all based on the same general statistic of squared difference-in-differences. Specifically, for a given time t we define

$$\mathcal{T}_t := \frac{1}{n_{1t}} \sum_{i:D_{1t}=1}^{n_{1t}} \left\{ \widehat{m}_{1t}(x_{it}) - \widehat{m}_{0t}(x_{it}) - \widehat{m}_{1(t-1)}(x_{it}) + \widehat{m}_{0(t-1)}(x_{it}) \right\}^2 , \tag{2.20}$$

which can be used to test several hypotheses of the general form:

$$\begin{aligned} H_0^t \quad : \quad & m_{1t}(x) - m_{0t}(x) - m_{1(t-1)}(x) + m_{0(t-1)}(x) = 0 \\ & \forall x \in supp(X|D=1, T=t) \ . \end{aligned}$$

Notice that if we now want to use a (wild) bootstrap to approximate the p-value via the distribution of $\mathcal{T}_t$ under H_0, then we need to resample the data under this null hypothesis which is a nontrivial task and will therefore be discussed in detail below.

Joint Significance of Heterogeneous Effects

When heterogeneity in treatment effects is important, it is more sensible (from a statistical point of view) and interesting (from an interpretation point of view) to test all TT_x jointly over the sample of interest. In the above terms that means to test H_0^1 (supposing treatment took place between $t = 0$ and $t = 1$) by checking if $\mathcal{T}_1$ under H_0^1 is significantly different from this statistic obtained from our sample.[9]

Homogeneous Treatment Effects

You can extend the above idea to test the null $H_0^1(c)$: $TT_x = c$ *for either all or an interesting subrange of* x, with c being a given constant. An interesting case is when you apply this to test all TT_x jointly over the sample of interest with $c := \widehat{TT}_a$. The resulting test statistic

$$\mathcal{T}_1^H := \frac{1}{n_{11}} \sum_{i:D_{1t}=1}^{n_{11}} \{\widehat{m}_{11}(x_{i1}) - \widehat{m}_{01}(x_{i1}) - \widehat{m}_{10}(x_{i1}) + \widehat{m}_{00}(x_{i1}) - c\}^2$$

would check if treatment effects are homogeneous over the observed x_{i1}. If $dim(x) = 1$, we could alternatively construct bootstrap confidence intervals and bands around TT_x for all x.

Bias Stability Condition

While Assumption I cannot be directly tested, its credibility can on periods prior to treatment. We use essentially the same statistic as we did for the selection procedures, namely (2.8), though a rescaling by the response variance estimate is not needed here. As for the selection of covariates and scale, the statistic $\mathcal{T}_t$ is applied to the pre-treatment period (from $t = -1$ to 0), where by definition, $Y_i = Y_i^0$ for all subjects i. Then, to test if the bias stability in the period(s) prior to treatment (e.g., from $t = -1$ to $t = 0$) held (H_0^0), consider the statistic $\mathcal{T}_0$. Not rejecting H_0^0 would strengthen Assumption I's (2.1) credibility.

2.4.3 Asymptotic Behavior

Here we study the asymptotic behavior of $\mathcal{T}_1$. For $\mathcal{T}_0$ and $\mathcal{T}_1^H$, the derivations follow analogously. In case you are concerned about setting, e.g., $c := \widehat{TT}_a$, note that $\widehat{TT}_a$ converges faster than $\widehat{m}_{dt}(\cdot)$ such that its randomness is negligible in the first order asymptotics of $\mathcal{T}_1^H$. To simplify notation, consider the case of a single continuous covariate $x \in [0, 1]$. We later discuss the case of $p = dim(x) > 1$, the inclusion of discrete covariates and the much simplified statistical behavior of this test statistic when we are provided with a balanced panel and covariates fixed over time are discussed afterwards.

Theorem 2.4.1. Define the four one-dimensional densities $f_{dt}(x)$ implicitly by $\int_0^{x_{it}} f_{dt}(x)dx = i/n_{dt}$ for all observed x_{it} with $D_{it} = d$.[10] Assume all $m_{dt}(\cdot)$ and $f_{dt}(\cdot)$ are $r \geq 2$ times

[9] As you may prefer TT_b over TT_a, you can also average in (2.20) over all treated ($n_{11} + n_{10}$).

[10] We could alternatively assume that all samples have asymptotically regular designs with respect to their density $f_{dt}(\cdot)$.

continuously differentiable on $[0,1]$, and the kernel $W(X,x,h)$ is of order r.

For the optimal testing rate $h = O(n_{11}^{-2/(4r+1)})$ with $n_{11}h^2 \to \infty$, and κ_{dt} as defined after (2.16), we have under H_0

$$n_{11}\sqrt{h}\left\{\mathcal{T}_1 - \frac{1}{n_{11}h}\int W^2 \sum_{d,t=0}^{1}\int \frac{\sigma_{dt}^2(x) f_{11}^2(x)}{\kappa_{dt} f_{dt}(x)}dx\right\} \longrightarrow N(0,\mathcal{V}) \tag{2.21}$$

as all $n_{dt} \to \infty$, where the variance $\mathcal{V}/(n_{11}^2 h)$ of statistic $\mathcal{T}_1$ is

$$\begin{aligned} \frac{2}{n_{11}^2 h}\int (W*W)^2 \Bigg(&\sum_{d,t=0}^{1}\int \frac{\sigma_{dt}^4(x) f_{11}^2(x)}{\kappa_{dt}^2 f_{dt}^2(x)}dx \\ &+2\sum_{mix(dt,ks)}\int \frac{\sigma_{dt}^2(x)\sigma_{ks}^2(x) f_{11}^2(x)}{\kappa_{dt}\kappa_{ks} f_{dt}(x) f_{ks}(x)}dx\Bigg), \end{aligned} \tag{2.22}$$

for which $\sum_{mix(dt,ks)}$ runs over the six combinations of $(dt) \neq (ks)$, $d,t,k,s \in \{0,1\}$.

For the case where the statistic $\mathcal{T}_1$ averages over the $n_1 = n_{11} + n_{10}$ treated, replace n_1 for n_{11} and $f_1(\cdot)$ for $f_{11}(\cdot)$ in (2.21), (2.22), and in the definition of κ_{dt}. Its extensions to allow for the inclusion of weights is discussed in the next section.

Similar statements can be made for higher dimensions ($p = dim(x) > 1$) using multivariate kernels. For simplicity, assume we take the same bandwidth h for all covariates; we only have to replace h by h^p in (2.21) and adjust its rate accordingly. Again, for $p > 3$, this requires bias reducing methods like the use of higher-order kernels or local-polynomials. Similarly, the inclusion of discrete covariates with smoothing parameter λ does not change our result, but renders the expressions more complex. Asymptotically, like in estimation, their inclusion does not change the rate. Due to (2.14), for balanced panels we have

Corollary 2.4.2. Consider a balanced panel taking all covariate values from $t = 0$ with $\tilde{\sigma}_d^2(x) = Var(u_{i1} - u_{i0}|x, D = d)$, and let $f_1(\cdot)$ define the density of X for the treated, $f_0(\cdot)$ for the controls. Then, along with the assumptions from Theorem 2.4.1,

$$n^1\sqrt{h}\left\{\mathcal{T}_1 - \frac{\int W^2}{h}\int \frac{\tilde{\sigma}_1^2(x) f_1(x)}{n^1} + \frac{\tilde{\sigma}_0^2(x) f_1^2(x)}{n^0 f_0(x)}dx\right\} \longrightarrow N(0,\tilde{\mathcal{V}}), \tag{2.23}$$

under H_0, for $\kappa_d = \lim(n^d/n^1)$, and with

$$\tilde{\mathcal{V}} = 2\int (W*W)^2\left(\sum_{d=0}^{1}\int \frac{\tilde{\sigma}_d^4(x) f_1^2(x)}{\kappa_d^2 f_d^2(x)}dx + 2\int \frac{\tilde{\sigma}_1^2(x)\tilde{\sigma}_0^2(x) f_1(x)}{\kappa_1\kappa_0 f_0(x)}dx\right). \tag{2.24}$$

2.4.4 Feasible Bootstrap Test

Arguments in favor of using a bootstrap for testing are as strong as for estimation. We need large samples before the first-order terms fully dominate the second and third-order terms.[11] Even if the samples were large enough to trust the normal approximation, and supposing that we could neglect higher-order terms, the estimation of the first-order terms for the variance(s) we saw above would still be a non-trivial problem. As said, the challenge is to simulate the distribution of the statistic $\mathcal{T}_t$ under the null hypothesis. We need to produce bootstrap samples that come from a data generating process similar to the observed data, but under which H_0^t holds.

Our proposal follows ideas of the related literature, namely Dette and Neumeyer (2001) and Vilar and Vilar (2012). The latter provide a consistency proof for our procedure. Their context is more complex regarding the correlation structure of the errors as they test several differences at a time. However, they only check differences of pairs of nonparametric functions whereas we are looking at the difference of differences. Only the latter has a consequence for the bootstrap. Different scenarios are conceivable to comply with H_0^t. For that reason, we need to take the residuals from the alternative (as proposed by Vilar and Vilar, 2012) instead of taking them from the null model (as proposed by Dette and Neumeyer, 2001).[12] This has consequences for the calibration (Sperlich, 2014). In the following we outline the procedure for $\mathcal{T}_1$ under H_0^1, i.e., under the hypothesis that

$$m_{11}(x) - m_{01}(x) - m_{10}(x) + m_{00}(x) = 0 \quad \forall x \in \mathit{supp}(X|D=1, T=0) \ .$$

The steps are:

1 Pool data (over treated and control groups) within each year $t, (t-1)$, and estimate $m_{t=1}(x) := E[Y|T=1, X=x]$ for all x observed in $t=1$. Analogously, $m_{t=0}(x) := E[Y|T=0, X=x]$ for all x observed in $t=0$.

2 Generate a large number (B) of bootstrap samples $\{Y_{it}^{*b}, (D_{it}=d), t, X_{it}\}_{i=1}^{n_{dt}}$, $b = 1, ..., B$, for each of the four (d,t) groups by setting $Y_{it}^{*b} = \widehat{m}_t(X_{it}) + u_{it}^{*b}$, for given $d, t,\ i = 1, ..., n_{dt}$, where u_{it}^{*b} might be generated by $\widehat{u}_{it} := Y_{it} - \widehat{m}_{dt}(X_{it})$ times an independent $N(0,1)$ variable.[13]

[11] For smaller samples sizes, the convergence rates observed in simulations are even faster than theory suggests (see e.g., Roca-Pardiñas and Sperlich, 2010).

[12] More specifically: in the literature on nonparametric testing, people either advertise to take the (larger) residuals from the null hypothesis, or they advertise the (smaller) residuals from the alternative as this makes the test more powerful. A problem with the former is that in case the null hypothesis is far from correct, the residuals can become quite large and destroy the power of the bootstrap test. In contrast, the latter strategy can lead to over-rejection in practice; see Sperlich (2014). Here we chose nonetheless the latter, mainly because different scenarios can result in the null hypothesis, and it is not clear from which to take the residuals then.

[13] Note that we define residuals $\widehat{u}_{it}$ to be taken from the alternative, i.e., estimating the conditional expectations $m_{dt}(\cdot)$ for all four groups separately, i.e., without pooling.

3 From these samples, calculate B estimators $\mathcal{T}_1^{*b}$ which are calculated as in (2.20), but with the $\widehat{m}_{dt}(\cdot)$ replaced by their bootstrap analogues $\widehat{m}_{dt}^{*b}(\cdot)$ estimated at $\{x_{it}\}_{i:D=1}^{n_{11}}$.

4 From the B bootstrap estimates $\mathcal{T}_1^{*b}$, obtain the p-value for the test statistic by counting how often the bootstrap statistics are larger than $\mathcal{T}_1$.

The key idea is the pooling in step 1, which guarantees that the null hypothesis (2.1) will be fulfilled in the bootstrap samples. It is, however, possible that within a year, the differences between groups are so severe that the pooling seriously diminishes power. For a robustness check, we could then switch the pooling and consider $m_d(\cdot)$, $d = 0, 1$. This has the tendency to suffer from size distortions in the sense of over-rejection. A possible reason why our former pooling proposal outperforms the latter is the following: D is definitely a function of X (by the definition of confounders), T should be much less so (for time invariant X it is definitely not). Consequently, under the null hypothesis of no treatment effect, a response prediction based on $m_t(x)$, ignoring d, should outperform a prediction based on $m_d(x)$, ignoring t. It may not always be true, but it likely occurs more often than not. This is confirmed by our simulations. Certainly, this bootstrap test can have poor power when the true data generating process is too far from the bootstrapped one.

It is obvious how to modify this procedure for $\mathcal{T}_0$ in order to test the credibility of Assumption I: you simply shift the entire procedure by one period to only compare cohorts before treatment started. Moreover, there exists an extensive literature on how to adapt the wild bootstrap to situations with correlated errors, may it be by given clusters (inside the groups) or autocorrelation (in panel data). Many of these modifications can be applied to our bootstrap in a straightforward way.

For more complex modifications or generalizations of our test, one should check the validity of the bootstrap procedure (at least with simulations) since its consistency does not necessarily carry over to all kinds of estimators or complex generalizations. For instance, Neumeyer and Sperlich (2006) studied a test in which they compared marginal (though not necessarily causal) effects from different cohorts with a similar statistic. For their context and estimator the wild bootstrap procedure was inconsistent and even divergent.

3 Nuts and Bolts

The basics of our methodology are given, but in practice there are many issues which can arise. In order for this approach to be useful in applications, we dig a bit deeper into the mechanics. In Section 3.1 we discuss practical issues (e.g., bandwidth selection). In Section 3.2 we outline a suggested algorithm as well as the `R` functions that can be used to carry out an analysis. In the Section 3.3 we show the finite sample performance via simulations.

3.1 Practical Issues

In this section, we discuss four critical issues surrounding the practical use of our procedures, namely the data-driven choice of bandwidths, bootstrap inference, how to incorporate sample weights, and potentially useful alternatives to kernel smoothing.

3.1.1 Bandwidth Selection

Bandwidth selection has a long history in nonparametric econometrics and it is a common view that they should be selected automatically via the data. Cross-validation (CV) routines are commonly performed and can be found in many texts (e.g., Henderson and Parmeter, 2015). Plug-in bandwidth selectors for both continuous (Silverman, 1986) and discrete (Chu *et al.*, 2015) data are feasible and less computationally intensive. See Köhler *et al.* (2014) for a review.

Data driven methods are attractive, but it is unclear what objective function the CV procedure should attempt to minimize. It can be argued that the final objective is not the optimal estimation of the TT_x, but of TT_a or TT_b. From a theoretical, asymptotic point of view, for those kind of semiparametric estimators, the optimal bandwidth must be of a faster rate than the usual optimal one or else its choice has only higher-order effects. This is in line with the findings of Frölich (2005) whose simulations show that CV bandwidths perform well in this respect. This occurs because CV bandwidths tend to undersmooth, but still keep the variance under control. For the matching context, Galdo *et al.* (2008) proposed a modified nonparametrically weighted CV method, Häggström and Luna (2014) a complex plug-in method based on nonparametric prior estimators, and Barbeito *et al.* (2023) a smoothed bootstrap method.

In our settings, we need bandwidths for at least four different nonparametric estimators. A computationally intensive method would be to use CV on each of the conditional expectations.[1] As most averages will only be made over the treated in $t = 1$, we propose to use least-squares cross-validation (LSCV) to estimate the bandwidths for the first conditional expectation, i.e.,

$$LSCV(h, \lambda) = \sum_{i:D_{i1}=1}^{n_{11}} \left(Y_i - \widehat{E}_{-i} \left[Y_i | X = x_i \right] \right)^2, \tag{3.1}$$

where $\widehat{E}_{-i} \left[Y_i | X = x_i \right]$ is the leave-one-out estimator of $E \left[Y_i | X = x_i \right]$ for the treatment group in time period 1 (i.e., $m_{11}(\cdot)$). The CV procedure picks the bandwidths (h, λ) which lead to the best out-of-sample prediction of the data (i.e., minimize the CV criterion). The bandwidths for the other conditional expectations can then be corrected by the sample size (the other three conditional expectations are expected to share the same smoothness as the first).

[1] We also tried this when conducting our application and found similar results.

If the set of potential sets of covariates, the number of potential transformations of Y, or sample size is too large for running the CV for all potential models, we can first resort to plug-in methods, and apply (3.1) once the selection of covariates and transformation is concluded. This is based on the assumption that the ranking of models along the selection criterion is robust within a reasonable range of bandwidths. For the continuous covariates, we may take a simple plug-in bandwidth developed only for densities because (i) it does not depend on the transformation θ and (ii) depends on the set of further covariates only via the rate. For discrete covariates, we could choose λ such that about $\sqrt{n_{dt}}$ observations are included in each estimation.

As we explain in more detail in Section 3.2.1, for estimation, as we have done in our application, we suggest the method above. We use CV to select the bandwidths for $m_{11}(\cdot)$ and modify that bandwidth (via the relevant sample size) for the other three cases. For testing, given the results in Parmeter *et al.* (2009) that suggest employing CV in nonparametric tests causes size distortions, we use plug-in bandwidths to calculate the relevant test statistics.

3.1.2 Bootstrap Inference

Asymptotic results for nonparametric statistics are rarely used directly for inference. Estimating any of the above variances is a nontrivial task that involves several bandwidth choices, with the challenge that there hardly exist bandwidth selectors for such variance estimators. Even if you succeed to estimate these expressions, in practice, the suppressed remainder terms may still play a role, not to mention the slow convergence to normality. In cases such as ours, bootstrapping is a widely accepted remedy. It is well known (Mammen, 1992), for nonparametric methods, that the naive bootstrap is insufficient (yields inconsistent estimators for most situations), while the wild bootstrap works. Abadie and Imbens (2008) confirmed the failure of naive bootstrap for kNN matching. Politis (2013) emphasized the superiority of nonparametric (which can be seen as a particular version of the wild) bootstrap for model-free prediction. This is common practice in matching and conditional DiD (Sperlich, 2013). Bodory *et al.* (2020) studied explicitly the consistency of the wild bootstrap for nonparametric matching estimators.

The distinction between a wild and nonparametric bootstrap often reduces to the question of how many moments are asymptotically matched. While asymptotic theory tells us that, the higher the bootstrap residuals match the moments of the original residuals, the more efficient the procedure, Davidson and Flachaire (2008) argue that you need quite large samples before this finding becomes effective. Following their recommendations, we propose a simple version (modifications towards higher-moment matching bootstraps are straightforward), first for continuous responses, then for discrete ones.

Given consistent nonparametric estimators for $m_{dt}(x)$, our residuals are given by

$$\widehat{u}_{it} = Y_{it} - \widehat{m}_{dt}(X_{it}) \ , \quad i = 1, ..., n_{dt}, \ d = 0, 1, \ t = 0, 1. \tag{3.2}$$

Generate B bootstrap samples $\{Y^{*b}_{it}, (D_{it} = d), t, X_{it}\}_{i=1}^{n_{dt}}$, $b = 1, ..., B$,[2] for all groups

$$Y^{*b}_{it} = \widehat{m}_{dt}(X_{it}) + u^{*b}_{it} \ , d = 0, 1, \ t = 0, 1, \ i = 1, ..., n_{dt}, \tag{3.3}$$

where u^{*b}_{it} can be generated by $\widehat{u}_{it}$ multiplied by an independent $N(0,1)$ variable (which performed best in our simulations).[3] From these B tuples of the four samples, we calculate B estimators of $\widehat{TT}^{*b}_x$, $\widehat{TT}^{*b}_a$ and/or $\widehat{TT}^{*b}_b$, which are calculated as described in Section 2.3.1, except that $\widehat{m}_{dt}(\cdot)$ is replaced with their bootstrap analogues $\widehat{m}^{*b}_{dt}(\cdot)$. From the B bootstrap estimates $\widehat{TT}^{*b}_z$ (for $z = x, a, b$), we obtain the bootstrap variance and confidence interval estimates for the corresponding $\widehat{TT}_z$.

For discrete Y, several scenarios are feasible. If you use the local-constant version and face binary responses, as we do in our application, you can generate bootstrap replicates

$$Y^{*b}_{it} := 1\!\!1\{\widehat{m}_{dt}(X_{it}) > v^b\} \quad , \ b = 1, ..., B \tag{3.4}$$

with randomly drawn $v^b \sim U[0,1]$. In our application, we received essentially the same standard errors when applying bootstrap versions of (3.3) and (3.4). In more complex cases, a link function is recommended. Then a semiparametric bootstrap can be applied to draw from the conditional distribution defined by this link: define a distribution with $Y|X = x \sim \mathcal{G}\{\eta(x)\}$; estimate the index function $\eta(x)$ and its conditional expectation by local-likelihood, and draw the bootstrap responses Y^*_{it} from $\mathcal{G}\{\widehat{\eta}(X_{it})\}$.

There exists an extensive literature on how to adapt a wild bootstrap to situations with correlated errors, say by given clusters (inside the groups) or autocorrelation (in panel data). Most of these modifications can be applied to our bootstrap in a straightforward way. For nonparametric analysis of continuous covariates, Faraway (1990) and Härdle and Marron (1991) noticed that these bootstrap procedures based on nonparametric or semiparametric models do not capture the smoothing bias well. This can lead to size distortions in nonparametric testing based on a wild or nonparametric bootstrap. They proposed to fix this problem by using different bandwidths for estimation (bandwidth h) and bootstrap sample generation (call this bandwidth g), see Sperlich (2014) for details. The same occurs for the smoothed bootstrap of Cao-Abad and González-Manteiga (1993). A less commonly used alternative is to explicitly correct for the smoothing bias, may it be by bias estimates (Xia, 1998; Cornillon *et al.*, 2017), bias reduction (for instance via higher

[2] As typically people look at confidence bands or intervals with $\alpha = 0.05$ or 0.10, we recommend to take a large number B (with 100 being an absolute minimum). We have not derived a theoretical justification for particular recommendations.

[3] Some favor the Rademacher distribution, though in a quite different context.

order kernels or higher order local polynomials) or a (double) bootstrap (Hall and Horowitz, 2013). Neumann and Polzehl (1998) show that asymptotically, using local-polynomials with undersmoothing h works as well, as the bias converges faster.

For testing with continuous covariates, we suggest an approach following the discussion of Vilar and Vilar (2012). Specifically, we search for the bandwidth over the set of scaled covariates X for each regression problem and apply then the largest h in all steps.[4] This is simple and works well in simulations. Regarding our choice of g (for generating the bootstrap residuals), our procedure is preferable in practice to the common recommendation of multiplying h by a fixed constant (e.g., $g = 1.5h$). While it is clear in simulations that you can find a constant such that the latter procedure performs better, in practice you don't know this constant *a priori*. As we only have discrete data in our application, we will not face the choice of g.

3.1.3 Sampling Weights

In our application, sample weights are used. This can be implemented in the most generic setting of our estimator $\widehat{m}_{dt}(x)$. Our objective function for a given conditional expectation can be written as

$$\sum_{i=1}^{n_{dt}} w_i \widehat{u}_i^2 K(X_i, x, h, \lambda) = \sum_{i=1}^{n_{dt}} w_i [Y_i - \widehat{m}_{dt}(x)]^2 K(X_i, x, h, \lambda),$$

where w_i is the sample weight for observation i. This leads to the (weighted) estimator

$$\widehat{m}_{dt}(x) = \frac{\sum_{i=1}^{n} Y_i w_i K(X_i, x, h, \lambda)}{\sum_{i=1}^{n} w_i K(X_i, x, h, \lambda)},$$

which, unfortunately, is not common in canned statistical packages. One way to implement this is via the `npksum` tool in the `np package` in `R` (Hayfield and Racine, 2008). This allows us to calculate

$$\sum_{i=1}^{n} Y_i w_i K(X_i, x, h, \lambda) \quad \text{and/or} \quad \sum_{i=1}^{n} w_i K(X_i, x, h, \lambda)$$

and taking the ratio of these two sums gives us the local-constant estimator. Certainly, the same approach works with other weighting schemes researchers may want to include (e.g., for scenario predictions).

3.1.4 Parametric and Semiparametric Alternatives

It is feasible to use parametric or semiparametric methods with our approach. We could replace the conditional expectations with parametric or semiparametric versions. However,

[4] This includes the bootstrap samples (i.e., estimating conditional means for pooled data under the null, and generating residuals under the alternative). See Section 2.4.4.

we still suggest that our method be first. Our methods do not have to be the last step, instead, they can guide the practitioner to find appropriate specifications and avoid wrong conclusions based on results which are strongly model-dependent. A compromise could be the use of splines which simplify modeling, but still provide important flexibility.[5]

3.2 Implementation

In this section we suggest an algorithm and highlight three procedure codes which can implement many of the methods discussed above.

3.2.1 Algorithm

We have produced three procedures that can be implemented in R (http://www.r-project.org). There are three separate procedures, namely covariate/scale selection, evaluating bias stability, and estimation. To employ the full set of procedures here we will need at least three periods of data. We need at least two periods prior to treatment to check the bias stability condition and at least one period after treatment to estimate the conditional and unconditional TT.

If there are more than two periods after treatment (say, $t = 1$ and 2), the code will collapse the additional periods (i.e., the TT estimate will be interpreted as an average of the TT over all post periods). Similarly, if there are more than two periods prior to the treatment (say, $t = -1$ and -2), the code will collapse the additional periods.

The algorithm is as follows:

1 Use both intuition and statistical analysis to suggest sets of potential confounders. It is important to pick the set of confounders that minimize the bias stability condition Assumption I. Possible suggestions include plotting the densities separately between groups and either visually confirming or statistically confirming the difference between densities.

2 Suggest possible strictly monotone transformations of the outcome variable Y. Two common cases in the continuous setting are in levels and logs.[6]

3 For each combination of transformations of Y and sets X^S of covariates for X, use plug-in bandwidths to calculate the conditional expectation $m_{dt}(\cdot)$ for the setting $d = 1$ and $t = 0$. Use the scale factors from this setting to select the plug-in bandwidths for

[5] Typically splines do not include all possible interactions among the covariates. This would be analogous to an additively separable nonparametric (kernel estimated) model, which would not be subject to the $p \leq 3$ restriction.

[6] In our application, Y is binary and hence is our only suggestion.

the conditional expectations for the other three cases ($d = 1, t = -1$, $d = 0, t = 0$ and $d = 0, t = -1$).[7]

4 For each combination listed in the previous step, calculate the bias stability condition in Assumption I, but for the period before treatment started. The combination that makes this condition closest to zero is our candidate set.

5 Run the bias stability test for the set X^S identified in step (4). If you reject the null, consider adding additional confounders and running steps (3) and (4) again.

6 For the combination of (transformation of) Y and (set of covariates) X^S that minimizes the bias stability condition, use a CV routine to best estimate the conditional expectation $m_{dt}(\cdot)$ for the setting $d = 1$ and $t = 1$. Use the scale factors from this setting to select the bandwidths for the conditional expectations for the other three cases ($d = 1, t = 0$, $d = 0, t = 1$ and $d = 0, t = 0$).[8]

7 Estimate each of the four conditional expectations and evaluate each TT of interest.

8 Obtain the standard errors via the bootstrap procedure outlined in Section 3.1.2 and perform the tests of interest.

3.2.2 Procedure Code

In this section, we detail three procedures that can be implemented in the programming language `R`. We decided to present them as three separate procedures as it may be desirable to disentangle them in an application. Note that the first two procedures require two periods of data prior to the treatment whereas the third only requires one period before and one after. The first procedure `bsc.choice()`, identifies the set of covariates including all confounders, and the scale of the outcome variable that minimize the objective function(s). Different from above, in the following of this Section we will speak of ‘confounders’ instead of ‘covariates’ simply to emphasize that this is very different from standard variable selection in regression. The second procedure `bsc.test()`, checks if the bias stability condition is violated via the test statistic $\mathcal{T}_t$. The final procedure `npdid.estimation()`, estimates the

[7] For the continuous founders, we suggest using the Silverman (1986) rule-of-thumb and for the discrete confounders we suggest using the methods discussed in Chu *et al.* (2015). These were designed for density estimation, but avoid the large computational burden with multiple combinations and CV (in the fifth step, we use CV to obtain more accurate estimates).

[8] For continuous variables, $h_j = c_j \widehat{\sigma}_{x_j} n^{-1/(4+q)}$, where c_j is the scale factor and $\widehat{\sigma}_{x_j}$ is the sample standard deviation of the jth continuous covariate. For discrete variables, $\lambda_j = c_j n^{-2/(4+q)}$, where c_j is the scale factor for the jth discrete covariate.

treatment effect $\widehat{TT}_b$.[9] All R code is available from the authors' upon request.[10]

Description of the Function bsc.choice()

The main purpose of this function, `bsc.choice()`, is to suggest a set of confounders amongst a set of potential confounders.[11] The `bsc.choice()` function can be called with,

```
bsc.choice(y,sx,d,t,w,ycont)
```

The function has six main arguments where the first four are obligatory. These are

`y`: The outcome variable, which is a $n \times 1$ matrix. It contains the outcome variable for each individual in each time period. It may be discrete or continuous.

`sx`: The sets of potential confounders, which is a list. It requires multiple data frames, each consisting of sets of potential confounders. The number of rows of each confounder must be of dimension n. The number of confounders and types of variables (discrete or continuous) can vary with each data frame. It is feasible to have some of the confounders in competing sets.

`d`: The treatment status. This is a binary variable of dimension $n \times 1$.

`t`: The time period. This is a discrete variable which must be equal to zero in the period immediately before the treatment was administered.[12] This variable is of dimension $n \times 1$.

`w`: These are the sample weights. It must be a $n \times 1$ matrix. If no sample weights are needed, it should be set equal to a column of ones.

`ycont`: This asks whether or not the outcome variable (y) is continuous. If set equal to "continuous", it will evaluate the function for both the level and the log of the outcome variable.[13]

The function consists of several steps. It first determines the type of variable (ordered, factor or continuous) from each data frame. It scales each continuous variable by its respective standard deviation. It then calculates plug-in bandwidths for each regressor

[9] It is feasible to extract $\widehat{TT}_a$ from the code for $\widehat{TT}_b$.

[10] We are currently in development of both R and Stata packages to perform all of the results in the application, including heterogenous TT estimates. The present versions of the packages are available at https://olegbadunenko.github.io/didnp/ and https://olegbadunenko.github.io/didnp/stata, respectively.

[11] It also checks for the level versus the log of Y if the outcome variable is continuous.

[12] We only consider treatment occurring in a single period. Extensions to treatments conducted in different time periods for different individuals is left for future research.

[13] Note that you must ensure that the outcome variable can be logged. Also, it is feasible to include alternative transformations of the outcome variable within the section of the code as desired.

type. For continuous variables it uses the Silverman (1986) bandwidth and for the discrete variables it uses the plug-in bandwidths from Chu *et al.* (2015). To equate the amount of smoothing across each functional, it calculates the scale factors for the treatment group in period 0 and then adjusts for the rate of convergence of the other three groups (treated before period 0, control in period 0 and control before period 0). Once these are obtained for each set of confounders, recall (2.8), i.e.

$$\frac{1}{n_{1\bullet}} \sum_{i:D_{i\bullet}=1} \left\{ m_{1t}(x_{i\bullet}) - m_{0t}(x_{i\bullet}) - m_{1(t-1)}(x_{i\bullet}) + m_{0(t-1)}(x_{i\bullet}) \right\}^2$$

we divide by the relative variance of the outcome variable as well as penalize for the number of confounders. This is calculated for each set of confounders and scale of the outcome variable. The procedure then determines the set which minimize (2.8).

The function then returns six objects. Each object of interest can be called via `$`:

`y`: The outcome variable associated with the smallest value for (2.8).

`x`: The set of confounders that minimize the objective function.[14]

`bsc.store`: The value produced for each set of confounders of (2.8).

`min.bsc.store`: The minimum value of produced amongst the set of confounders of (2.8).

`qt`: The number of discrete regressors in the chosen set of confounders.

`qc`: The number of continuous regressors in the chosen set of confounders (should be three or less).

At this point, the user could take the resulting outcome variable and set of confounders and conduct the `bsc.test()` with those variables.

Description of the Function bsc.test()

The main purpose of this function, `bsc.test()`, is to check if there is a violation of the bias stability condition.[15] The `bsc.test()` function can be called with,

```
bsc.test(y,x,d,t,w,nb)
```

The function has six main arguments where the first four are obligatory. These are

[14] The function scales each of the continuous variables to have variance 1. This improves estimation in practice and does not impact the ranking of sets of confounders nor does it impact the estimated treatment effect.

[15] It is feasible to modify this procedure to conduct the significance test.

`y`: The outcome variable, which is a $n \times 1$ matrix. It contains the outcome variable for each individual in each time period. It may be discrete or continuous.

`x`: The set of confounders, which is a data frame. This is a $n \times q$ matrix where q refers to the total number of confounders.

`d`: The treatment status. This is a binary variable of dimension $n \times 1$.

`t`: The time period. This is a discrete variable which must be equal to zero in the period immediately before the treatment was administered. This variable is of dimension $n \times 1$.

`w`: These are the sample weights. It must be a $n \times 1$ matrix. If no sample weights are needed, it should be set equal to a column of ones.

`nb`: The number of bootstrap replications. This must be an integer value. If not specified, 399 bootstrap replications will be run.

The function consists of several steps. It first determines the type of variable (ordered, factor or continuous) from the data frame. It then calculates plug-in bandwidths for each regressor type. For continuous variables it uses the Silverman (1986) bandwidth and for the discrete variables it uses the plug-in bandwidths from Chu *et al.* (2015). To equate the amount of smoothing across each functional, it calculates the scale factors for the treatment group in period 0 and then adjusts for the rate of convergence of the other three groups (treated before period 0, control in period 0 and control before period 0). Once this is obtained for the set of confounders, $\mathcal{T}_t$ is calculated. A bootstrap[16] is used to approximate the sampling distribution of the test statistic.

The function then returns four objects. The first object, a figure, will automatically be produced. The remaining three objects of interest can be called via `$`:

`bsc.stat`: The value produced by $\mathcal{T}_t$.

`sd.bsc`: The standard deviation (standard error of the test statistic) of the bootstrapped estimates of the test statistic.

`p.value`: The p-value associated with the test statistic. This is calculated as the percentage of bootstrapped test statistics which are larger than the original test statistic.

[16] The code can automatically detect if the outcome variable is binary. If so, then a bootstrap procedure which ensures the bootstrap outcome is binary, is applied.

The figure plots the estimated density of the bootstrapped test statistics[17] along with the value of the test statistic itself as a vertical line. If the vertical line does not appear present in the figure, it is likely far to the right which would suggest rejecting the null hypothesis (i.e., a p-value near zero).

Description of the Function npdid.estimation()

The final function, `npdid.estimation()`, is designed to estimate the treatment effect and its standard error. The `npdid.estimation()` function can be called with,

```
npdid.estimation(y,x,d,t,w,nb)
```

The function has six main arguments where the first four are obligatory. These are

`y`: The outcome variable, which is a $n \times 1$ matrix. It contains the outcome variable for each individual in each time period. It may be discrete or continuous.

`x`: The set of confounders, which is a data frame. This is a $n \times q$ matrix where q refers to the total number of confounders.

`d`: The treatment status. This is a binary variable of dimension $n \times 1$.

`t`: The time period. This is a discrete variable which must be equal to zero in the period immediately before the treatment was administered. This variable is of dimension $n \times 1$.

`w`: These are the sample weights. It must be a $n \times 1$ matrix. If no sample weights are needed, it should be set equal to a column of ones.

`nb`: The number of bootstrap replications. This must be an integer value. If not specified, 399 bootstrap replications will be run.

The function consists of several steps. It first determines the type of variable (ordered, factor or continuous) from the data frame. It then calculates plug-in bandwidths for each regressor type to be used as starting values for the cross-validation function. Again, for continuous variables it uses the Silverman (1986) bandwidth and for the discrete variables it uses the plug-in bandwidths from Chu *et al.* (2015). To equate the amount of smoothing across each functional, it calculates the scale factors for the treatment group in period 1 and then adjusts for the rate of convergence of the other three groups (treated before period 0, control in period 0 and control before period 0). The LSCV procedure defined in (3.1) is minimized using the `bobyqa()` function in the `minqa` package in `R`. We calculate the scale factors from the CV function for the treatment group in period 1 and then adjust for the

[17] The Sheather and Jones (1991) bandwidth is used to produce this kernel density. It is available in the base package of `R` via `density(x,bw="sj")`.

rate of convergence of the other three groups (treated in period 0, control in period 1 and control in period 0).

We are interested in both, the effect heterogeneity (i.e., estimating first TT_x) and the average treatment effect (here we will integrate the TT_x only over the second cohort of treated giving us $\widehat{TT}_a$). A bootstrap[18] is used to produce the sampling distribution of the TT. We use the sample standard deviation of the bootstrapped values of TT as the standard error of the treatment effect.

The function then returns six objects. Each object of interest can be called via `$`:

`bw11`: The cross-validated bandwidths for the treatment group in period 1.

`bw10`: The convergence rate adjusted bandwidths for the treatment group in period 0.

`bw01`: The convergence rate adjusted bandwidths for the control group in period 1.

`bw00`: The convergence rate adjusted bandwidths for the control group in period 0.

`atet`: The estimated value of the TT

`sd.atet`: The estimated standard error of the TT

These three functions together can be used to reproduce nonparametric results in the paper. They can be used to replicate the simulations or the empirical application. The `R` files that we used to construct any of these results are also available upon request.

3.3 Simulations

In this section, we study the finite sample performances, and show our theoretical results hold with simulated data. We focus our attention on three sets of simulations. First, we see how well our method picks the correct set of covariates (i.e., confounders). Second, we examine the nominal size and power of our test for violation of the bias stability condition. Finally, we examine the performance of our estimate of the TT and its variance.

We begin with this basic data generating process and specifically mention where it is modified below. We keep it simple and only look at two covariates, no time correlation, continuous Y, and no interactions. We generate our two covariates via $X_{it} \sim U[0,2]^2$, and our random errors via $\epsilon_{it} \sim N(0, 1.5)$ (c.f. E in our causality graphs), and $u_{it} \sim N(0, \sigma_u^2)$ for $t = -1, 0, 1$. We obtain the treatment status and outcome values as

$$\begin{aligned} D_{it} &= \mathbb{1}\{0.75X_{it,1} - 0.5X_{it,2}^2 > \epsilon_{it}\} \qquad (3.5) \\ Y_{it} &= 1 + t(2 + X_{it,1} + X_{it,2}^2) + D_{it} + D_{it}\mathbb{1}\{t \geq 1\} + u_{it} \qquad (3.6) \end{aligned}$$

[18] The code can automatically detect if the outcome variable is binary. If so, then a bootstrap procedure which ensures the bootstrap outcome is binary, is applied.

where the treatment effect on the treated is the coefficient on the interaction term (i.e., $TT = 1.0$) in (3.6).[19] In (3.6) this starts from period $t = 1$ onward. We consider samples of size $n = \Sigma_{t=-1}^{1} n_t = \Sigma_{d=0}^{1}\Sigma_{t=-1}^{1} n_{dt} = 100, 200, 400$ and 800 where n is the total number of observations of all individuals in all time periods, n_t is the number of individuals in time period t (3 total time periods are observed) and n_{dt} is the number of individuals in group d in time period t. We are creating a repeated cross-section whereby each sample produces roughly an equal number of treated and controlled observations.

While we choose $n = 100, 200, 400$ and 800, the effective sample sizes are much smaller. The last two columns of numbers in Table 3.1 give the average sample size (to the nearest integer) for n_{10} (the number of observations we sum over in our criterion function), and the smallest sample size over all n_{dt} ($d \in \{0, 1\}, t \in \{-1, 0, 1\}$).[20] For example, for $n = \Sigma_{t=-1}^{1} n_t = \Sigma_{d=0}^{1}\Sigma_{t=-1}^{1} n_{dt} = 100$, the average number of observations in $n_{10} = 18$ and $\min(n_{dt}) = 12$. This is unheard of in nonparametric kernel estimation, yet our methods perform admirably.

Given that we only consider continuous outcome variables and covariates, we use Gaussian kernel functions. Adding additional discrete covariates or having a binary outcome variable does not significantly impact the results of the simulations. In each exercise, we use 999 Monte Carlo simulations. For cases that require bootstrap replications, we use $B = 999$ bootstrap replications.

We do not consider linear parametric models as our data are generated nonlinearly and standard linear models will produce biased estimates here (i.e., stickman comparison models). Even if we had correctly specified parametric models, we would expect similar results from both approaches. Given our theoretical results and potential parametric functional form misspecification, we feel the comparison is unnecessary in this simulated setting.[21] By a *linear parametric model* we refer to a model in which X enters linearly. Equivalences to linear models without X were considered by others (Lechner, 2011; Frölich and Sperlich, 2019). We do not consider the two-way fixed effects model or its problems (Chaisemartin and D'Haultfoeuille, 2020) as this generally refers to potentially unbalanced panel data regression with a purely linear specification, fixed effects for all subjects or groups and time periods, facing more than two time points in which it is not required that controls were observed over the entire period nor that all included treated subjects were also observed before treatment. In sum, it allows for many violations which renders a comparison beyond the scope of this paper. We plan to address this more formally in future research.

[19] While our simulations have to be generated by a specific parametric model, our nonparametric model does not include a treatment times post-time variable as our estimation strategy focuses on four conditional expectations.

[20] The effective sample sizes are nearly identical in the remaining tables.

[21] We compare our methods to linear parametric methods in our application.

3.3.1 Choice of the Confounder Set

To see if our method appropriately picks the correct set of covariates, we generate our data as in (3.6). However, we also generate irrelevant covariates (from the same distributions as our relevant covariates). In each case, we include both the correct covariates and then add either all irrelevant or some irrelevant covariates to determine if we can identify the correct set. We present the results for moderate ($\sigma_u^2 = 1.0$) and a low signal-to-noise ratio ($\sigma_u^2 = 2.0$). In each case, all our (three separately simulated) irrelevant covariates come from a uniform distribution from zero to two. In other words, we generate each $X_{it,j} \sim U[0, 2]$ separately for $j = 1, 2, \ldots, 5$. We consider the following sets:

1. $S_{1,2} = \{X_{it,1}, X_{it,2}\}$,
2. $S_{1,3} = \{X_{it,1}, X_{it,3}\}$,
3. $S_{2,4} = \{X_{it,2}, X_{it,4}\}$,
4. $S_{3,4} = \{X_{it,3}, X_{it,4}\}$,
5. $S_{4,5} = \{X_{it,4}, X_{it,5}\}$,
6. $S_{1,3,4} = \{X_{it,1}, X_{it,3}, X_{it,4}\}$,
7. $S_{2,4,5} = \{X_{it,2}, X_{it,4}, X_{it,5}\}$,
8. $S_{1,2,3} = \{X_{it,1}, X_{it,2}, X_{it,3}\}$,
9. $S_{1,2,4} = \{X_{it,1}, X_{it,2}, X_{it,4}\}$,
10. $S_{1,2,3,4} = \{X_{it,1}, X_{it,2}, X_{it,3}, X_{it,4}\}$, and
11. $S_{1,2,3,4,5} = \{X_{it,1}, X_{it,2}, X_{it,3}, X_{it,4}, X_{it,5}\}$.

We consider the following comparisons against $S_{1,2}$ (i.e., the correct set of covariates): versus $S_{1,3}$ and $S_{2,4}$, versus $S_{3,4}$ and $S_{4,5}$, versus $S_{1,3,4}$ and $S_{2,4,5}$, versus $S_{1,2,3}$ and $S_{1,2,4}$, and finally, versus $S_{1,2,3,4}$ and $S_{1,2,3,4,5}$. The first comparison is the hardest as each time just one relevant covariate was replaced. We do not know in advance which is the second most difficult, as this depends on how well the penalty factor $\left(2(k+p)^2 + 2(k+p)\right) / \left(n_{1\bullet} - (k+p)\right)$ does its job.

If we choose at random, then the fraction correctly specified should be approximately 1/3 and if we choose correctly each time, then it should be 1. Table 3.1 gives the results of our simulations. The top panel is for the moderate signal-to-noise ratio and the lower panel is for the low signal-to-noise ratio. As expected, we perform better when the signal-to-noise ratio is higher. It is clear that larger sample sizes are needed when more noise is present in the model.

The first case represents the hardest one. With $n = 100$ (i.e., some n_{dt} just about 10), we are roughly at or above random choice. For $n > 100$, it improves even for low signal-to-noise ratios.[22] If we move to the second column, the procedure already works for $n = 100$, and quite rapidly improves for increasing samples or higher signal-to-noise ratios.

The third column of numbers add an additional irrelevant covariate. Here, with help of the penalty factor, we easily distinguish the correct set of covariates from those with one relevant covariate. For a more fair comparison, we include both relevant covariates and one irrelevant covariate in the fourth column of numbers. Here we actually do better. Even for sample sizes as small as $n = 100$, we correctly predict over 0.97 for both the low and moderate signal-to-noise settings. Finally, we add two and three irrelevant covariates to the two correct covariates in the fifth column. These fractions are near one in every setting.

In summary, we were generally able to identify the correct set of covariates. In practice, we expect a mix of relevant and irrelevant covariates in each set. Given that we have very small sample sizes here, we have faith in practice that our method will choose the correct set of covariates with standard sample sizes in the applied literature.

[22] We continued to raise the sample size to see if these fractions tended towards 1. When doubling the sample size, this occurred by $n = 3200$ ($n_{10} \approx 575$) for $\sigma_u^2 = 2.0$.

Table 3.1: Fraction correctly choosing $S_{1,2}$ versus alternative sets of covariates: AIC penalty factor included, average sample size (to the nearest integer) for n_{10} and $\min(n_{dt})$ given for each overall sample size ($n = \Sigma_{t=-1}^{1} n_t = \Sigma_{d=0}^{1}\Sigma_{t=-1}^{1} n_{dt}$)

σ_u^2	n	$S_{1,3}, S_{2,4}$	$S_{3,4}, S_{4,5}$	$S_{1,3,4}, S_{2,4,5}$	$S_{1,2,3}, S_{1,2,4}$	$S_{1,2,3,4}, S_{1,2,3,4,5}$	n_{10}	$\min(n_{dt})$
1.0	100	0.376	0.593	0.893	0.975	0.998	18	12
	200	0.411	0.654	0.897	0.976	0.999	35	27
	400	0.491	0.812	0.907	0.985	0.999	71	57
	800	0.577	0.912	0.921	0.990	0.999	140	119
2.0	100	0.320	0.493	0.864	0.971	0.996	18	12
	200	0.381	0.541	0.888	0.973	0.999	35	27
	400	0.403	0.713	0.896	0.982	1.000	70	57
	800	0.522	0.804	0.918	0.987	1.000	141	119

3.3.2 Testing

Here we check the performance of our second primary contribution, nonparametric tests for the credibility of bias stability, joint significance of heterogeneous effects, and homogeneous treatment effects, respectively. Recall that studying the unconditional TT is much easier (Section 3.1.2). We conduct our simulations along the problem of studying the bias stability ('parallel path') condition.[23] We generate our data as in (3.6) to determine the size of the test. To determine the power, we change the indicator function to $\mathbb{1}\{t \geq 0\}$ in (3.6) as this will generate a situation in which the bias stability condition is violated. We again use $n = \Sigma_{t=-1}^{1} n_t = \Sigma_{d=0}^{1} \Sigma_{t=-1}^{1} n_{dt} = 100, 200, 400$, and 800 total observations and estimate the size (and power) of the test at each of the common (arbitrary) values $(1, 5$, and $10\%)$.

Inference with nonparametric estimation methods can be notoriously difficult. Using the asymptotic variances of tests are often useless and bootstrap procedures can bring large improvements. That being said, it is common to oversmooth with such tests when using the bootstrap. As we mentioned in the main document, we recommend a common approach of oversmoothing when calculating the residuals which are used in the bootstrap procedure (Vilar and Vilar, 2012). We calculate the test statistic $(\mathcal{T}_0)$ as outlined above, but calculate the residuals using the bandwidth procedure of Vilar and Vilar (2012).[24] In short, we obtain the bootstrap residuals by adding the fitted values (using the standard bandwidth) to the resampled residuals (using the larger bandwidth). Using the smaller bandwidth leads to too little variation in the data (and would result in an improperly sized test).

[23] We focus our attention on this particular test statistic as it is the most difficult and maybe most interesting one.

[24] We tried the generic approach of multiplying the bandwidth by a constant (Härdle and Marron, 1991, pp. 791). Specifically, we set $g = 1.5h$, where h is obtained from plug-in methods (only necessary for continuous variables). The size of the test for this approach is better than what we present. As the multiple (1.5) is arbitrary, we preferred the automated approach in Vilar and Vilar (2012). The results are available upon request.

Table 3.2: Size and power of our bias stability ('parallel path') condition test ($\mathcal{T}_0$): The probability of rejection at each significance level (1, 5 and 10%) using $B = 999$ bootstrap replications in each of our 999 simulations, average sample size (to the nearest integer) for n_{10} and $\min(n_{dt})$ given for each overall sample size ($n = \Sigma_{t=-1}^{1} n_t = \Sigma_{d=0}^{1}\Sigma_{t=-1}^{1} n_{dt}$)

		size			power				
σ_u^2	n	1%	5%	10%	1%	5%	10%	n_{10}	$\min(n_{dt})$
1.0	100	0.006	0.036	0.070	0.067	0.212	0.326	18	13
	200	0.009	0.055	0.086	0.190	0.401	0.531	35	28
	400	0.011	0.054	0.121	0.488	0.743	0.837	71	58
	800	0.010	0.049	0.109	0.870	0.971	0.987	140	119
2.0	100	0.006	0.026	0.083	0.030	0.127	0.213	18	12
	200	0.012	0.042	0.128	0.074	0.222	0.332	35	27
	400	0.009	0.056	0.125	0.212	0.420	0.573	71	58
	800	0.011	0.053	0.110	0.517	0.724	0.823	140	119

The results for both the size and power of our test ($\mathcal{T}_0$) can be found in Table 3.2. The test seems to be correctly sized starting with relatively small samples (say $n > 200$). As expected, the size of the test improves with the number of observations and is better in the moderate signal-to-noise ratio. This is impressive given the history of nonparametric kernel based tests. We do feel the need to mention that the oversmoothing here is necessary. When we perform the test without a bandwidth g, the test is not properly sized (even for relatively large samples).

As for the power of the test (again in Table 3.2), the power is relatively low for small sample sizes, but improves quickly as n increases. For example, when $\sigma_u^2 = 1.0$, by the time $n = 800$, the percent of time the test correctly rejects the null is in excess of 85% at the 1% level and in excess of 97% at the 5 and 10% levels. The results for $\sigma_u^2 = 2.0$ are also strong, but require about twice as many observations when compared to the moderate signal-to-noise ratio.

In conclusion, the test is easy to use and works well. Power decreases for increasing dimensions (especially when bias reducing techniques are needed: $p > 3$). We also studied in detail the effect when the true data generating process deviates from the bootstrap generating process in different ways. While certainly the p-value estimate is affected, the test generally detected violations of the parallel path.

3.3.3 Treatment Effect Estimator

Finally, we move to estimates of the TT and its variance. Our estimators are consistent, but we provide a brief set of results here for TT_b to confirm (i.e., integrate TT_x over all treated individuals).[25] While consistency should not be in question, the ability of nonparametric estimators to produce correct results for the variance are less reliable. The asymptotic results are not useful for finite sample sizes and so we employ our bootstrap procedure outlined in Section 3.1.2. We do not require a bandwidth g and use h for both estimation and in our bootstrap.[26]

It should be emphasized again, with TT_b, we are integrating over all treated individuals. In other words, we are summing over n_{11} and n_{10}. What this implies is that we are using roughly twice the number of observations as compared to the previous two sub-sections. The results for TT_a would use roughly half as many observations (i.e., solely n_{11}).

Table 3.3 gives our simulation results. We choose a moderate (upper panel) and a low (lower panel) signal-to-noise ratio. In each case, the finite sample bias exists and tends

[25] The results for each of our treatment effect estimators are similar. Simulations for TT_a or TT_x (at a given x) are available upon request.

[26] We again use plug-in methods here, but note that the bias is much smaller for cross-validated bandwidths as plug-in methods tend to oversmooth. Specifically, for the cross-validated bandwidths, by the time $n = 400$ ($n_{11} = 71, n_{10} = 84$), our average (over the 999 simulations) biases are zero to two decimal places.

towards zero as n increases. Again, larger biases are functions of using plug-in bandwidths which tend to oversmooth (LSCV bandwidths lead to much smaller average biases).[27]

The average mean square error (AMSE) also tends towards zero (evidence that our estimator is consistent). As expected, the moderate signal-to-noise ratio results in smaller AMSE values for any given sample size (it does not significantly impact the bias). The third column of numbers gives the average variance of the TT_b estimator over each of the 999 simulations. Recall that we calculate the variance in each of those 999 simulations via 999 bootstrap replications. We are able to see the variance of the estimator converges as the sample size increases.

The performance of our estimator is impressive given its nonparametric nature. Overall, our simulations suggest that our covariate selector, test and estimator are reliable and match our asymptotic developments. Next, we discuss the use of these methods with empirical data.

[27] We advocate for using cross-validated bandwidths in practice when estimating the TT. The sign of the bias is not random, but if it is negative can only be deduced from the average over the linear combinations of individual biases $B_{dt}(x, \lambda, h)$, which in turn depends on the particular bandwidth choices, true densities and functions. Importantly, it is minor in size and rapidly converges to zero.

Table 3.3: Performance of nonparametric TT_b estimator: Average bias and MSE over the simulations and average variance (calculated via $B = 999$ bootstraps over each of the 999 simulations), average sample size (to nearest integer) for n_{11} and $\min(n_{dt})$ given for each overall sample size ($n = \Sigma_{t=-1}^{1} n_t = \Sigma_{d=0}^{1} \Sigma_{t=-1}^{1} n_{dt}$)

n	Bias	AMSE	Var(TT_b)	n_{11}	$\min(n_{dt})$	Bias	AMSE	Var(TT_b)	n_{11}	$\min(n_{dt})$
			$\sigma_u^2 = 1.0$					$\sigma_u^2 = 2.0$		
100	-0.190	0.434	0.182	18	12	-0.190	0.782	0.353	18	12
200	-0.140	0.190	0.111	35	27	-0.140	0.351	0.218	35	28
400	-0.106	0.100	0.065	71	57	-0.102	0.185	0.128	71	58
800	-0.089	0.049	0.036	140	119	-0.087	0.088	0.071	140	119

4 Application: Human Capital Responses to DACA

On June 18th, 2020, the Supreme Court of the United States ruled that the president could not immediately end DACA. As any attempts to strike down the program will need additional study, it is important to carefully examine the evidence both for and against the program. One potential benefit is that the rules in place to qualify for DACA require schooling. Additional units of education should lead to increased human capital and benefits to society. Kuka *et al.* (2020) examine human capital responses to the availability of the DACA program and (using a DiD approach) find that DACA significantly increased high school attendance and completion rates. They further find positive, but insignificant, impacts on college attendance. Even though they had only discrete covariates which they all decomposed to dummies, the results still rely on restrictive parametric assumptions and hence are subject to misspecification bias and potential inconsistency, cf. our discussion on (allegedly) saturated models in Appendix B.1. Moreover, we show that for their set of covariates, there are issues with the underlying identification assumption.

4.1 Data

The data come directly from Kuka *et al.* (2020) and we only discuss them briefly[1]. Kuka *et al.* (2020) use the Integrated Public Use Microdata Series (IPUMS) American Community Survey (ACS) (Ruggles *et al.*, 2018) over the period 2005–2015. They focus on (a sample of) immigrant youth aged 14 to 22 during the time of the survey such that they arrived on US soil by the age of 10 in 2007. The sample from 14–18 is used to study high school attendance, while the sample from age 19–22 is used to study high school completion (including those who graduated from high school as well as those who earned a passing grade on the General Educational Development test) and post-secondary attendance (three different binary left-hand-side variables). Recall that with a binary outcome, linear DiD estimators do not guarantee the predicted outcome lies between zero and one. Our nonparametric estimator guarantees this support condition.

The ACS includes a large amount of demographic variables which are exploited by Kuka *et al.* (2020) to attempt to make Assumption I hold. Specifically, they account for fixed individual characteristics by including controls for sex, year of immigration and birth region. Given the nature of parametric models, they also include interactive dummies for age of immigration-by-eligibility and age-by-eligibility fixed effects.[2] They include state-by-year, race-by-year and age-by-year fixed effects. Our nonparametric methodology does not require arbitrary interactions (even if based on sound logic), but accounts for them automatically.

[1] The data are freely available at `doi.org/10.1257/pol.20180352`.

[2] In econometrics, those fixed effects are used to control for unobserved time-invariant heterogeneity which may be correlated with the error term.

We have seven different potential variables for X in each regression. Each are discretely measured. The potential unordered variables include sex, race, birthplace and current U.S. state, while the potential ordered variables include age, year, year of immigration and age at time of immigration.[3]

It is important to note that the ACS is a representative sample of those living in the United States, regardless of their citizenship or legal status. The Census Bureau encourages responses to ACS and is not allowed to share the personal information with other government agencies, and it also makes the survey available in Spanish.

Kuka *et al.* (2020) note that their measure of eligibility is measured with noise as it includes non-citizens who may have green cards or may be temporary visa holders (i.e., not eligible for DACA). The estimated effect of DACA is likely a "scaled-down" estimate of the true intent-to-treat effect. Their Appendix B estimates that their estimated effects are likely to underestimate the true effect by roughly 45 percent.

4.2 Average Treatment Effects

The replicated parametric results can be found in Tables 4.1 and 4.2. These correspond to their model

$$\begin{aligned} Y_{idast} &= \alpha_0 + \alpha_1 Eligible_d + \alpha_2 \left(Eligible_d \times Post_t\right) + \alpha_3 X_{id} \\ &\quad + \gamma_{st} + \gamma_{rt} + \gamma_{at} + u_{idast}, \end{aligned}$$

where Y is the outcome of interest (in school, completed high school or some college) for individual i, who has eligibility status d, who is aged a and living in state s at time t. Given the sample selection (age and year of immigration), *Eligible* is a dummy variable that equals 1 if the immigrant is not a citizen and zero otherwise. The variable *Post* is a dummy variable that equals 1 on or after 2012. X_{id} includes the dummies for sex, year of immigration and birth region, while each of the γ terms represent the interactive fixed effects. The treatment effect estimate is captured by α_2. It is interpreted as the average effect of DACA after 2012 (the analysis covers four "treated" years: 2012–2015).

Parametric estimation is performed via least-squares dummy-variable techniques and requires a relatively large memory to construct (not to mention invert) such a data matrix. The authors cluster their standard errors at the state level. The nonparametric estimates (TT_b) are listed below their parametric counterparts. Estimation of our treatment effect is described above (Section 2.3), we use cross-validated bandwidths and use our bootstrap procedure (with $B = 999$) to calculate our standard errors.

[3] In the Hispanic sample and in the high-take up sample, we exclude the variable for race and region of origin. In the cases where we only examine 19 year olds, we remove the variable for age.

The final three values associated with each sample in Tables 4.1 and 4.2 are the sample size, the mean of the outcome variable and the p-value associated with our bias stability test. The latter shows mixed results.[4] In Table 4.1, we firmly reject the null that the BSC ("parallel path") holds in our sample for 14-18 year olds, but are unable to reject it for each case for 19-22 year olds. Table 4.2 shows four cases where we fail to reject the null and five cases where we reject the null. As we are simply looking to replicate the results of their paper, we proceed as if we were unable to reject the null hypothesis in each scenario.[5] We therefore should be careful about the interpretation of each treatment effect as identification is in question for several of them. In practice, we would suggest that more potential covariates be tracked down in order to satisfy the identification condition.

4.2.1 School Attendance

The results for school attendance are found in Table 4.1. For individuals aged 14-18, the parametric models show positive and significant estimates for each grouping (all, Hispanic and high take-up sample). These results suggest that DACA led to an increase in school attendance of 1.2 percentage points among all immigrants with 2.2 and 2.9 percentage point increases for Hispanic and high take-up sample immigrants.

If we look to the nonparametric results for those aged 14-18, they are larger (albeit not statistically larger). The nonparametric point estimates are 0.022, 0.033 and 0.034 and the standard errors are similar (0.005, 0.008 and 0.008 versus 0.007, 0.012 and 0.012 for the parametric and nonparametric models, respectively). This bodes well for the results in Kuka *et al.* (2020). The nonparametric models relax restrictive assumptions and the conclusions are statistically similar. Ignoring other potential issues, these results should be considered to be robust.

[4] As all variables are discrete, there is no need to oversmooth bandwidths in the bootstrap routine.

[5] The usual caveat applies: a failure to reject the null hypothesis is not an acceptance of the null.

Table 4.1: Effect of DACA on school attendance

	All	Hispanic	High take-up	All	Hispanic	High take-up
		Age 14-18			Age 19-22	
Parametric	0.012	0.022	0.029	0.019	0.020	0.005
	(0.005)	(0.008)	(0.008)	(0.012)	(0.014)	(0.012)
Nonparametric	0.022	0.033	0.034	-0.047	-0.034	-0.051
	(0.008)	(0.012)	(0.012)	(0.015)	(0.021)	(0.021)
Average Y	0.921	0.891	0.889	0.5467	0.405	0.401
Sample size n	114,453	54,015	48,359	82,077	38,704	34,768
BSC p-value	0.000	0.000	0.000	0.317	0.191	0.524

Table 4.1 also gives the results for 19-22 year olds. While this group was primarily used to examine later schooling outcomes, it is interesting to see these impacts. The parametric model gives positive, but insignificant estimates. The nonparametric model gives negative and significant estimates for each sample. There is substantial evidence in the literature to suggest that the impact of DACA on college-age enrollment is in fact negative. Hsin and Ortega (2018) found that DACA increased dropout rates by 7.3% in 2018. Amuedo-Dorantes and Antman (2017) found that DACA reduced the probability of school enrollment of eligible higher-educated individuals as it increased the likelihood of employment of men. The lack of authorization led individuals to enroll in school when working legally was not feasible. While the differences in point estimates with respect to 14-18 year olds is interesting, the ability of our method to identify the negative impact on college-aged individuals shows the downsides of relying on parametric assumptions.

4.2.2 High School Completion and College Enrollment

The effects of DACA on high school completion and college enrollment can be found in Table 4.2. The first three columns represent the effect on high school completion (GED or diploma) for all immigrants, Hispanic immigrants and immigrants from high take-up countries, respectively. These results are broken down by age (19, 19-22 and 23-30). Similarly, the fourth through sixth columns give the impact of DACA on the completion of some college (more than 12 years of education completed) for each of the groups (all, Hispanic, and high take-up) for each age group.

Beginning with the parametric high school completion regressions, completion rates for all 19 year old immigrants increased by 4.6 percentage points. The effects for 19 year old Hispanics and immigrants from high take-up countries experienced increases of 6.5 and 8.5 percentage points, respectively. The impact for 19-22 year olds is smaller: 3.8, 5.9 and 6.4 percentage point increases for all, Hispanic and high take-up sample immigrants, respectively. For those individuals 23-30 years old, the impacts are either marginally significant or insignificant. The impact appears to be stronger for younger individuals.

Table 4.2: Effect of DACA on high school completion and college enrollment

			High-School			College	
Age		All	Hispanic	High take-up	All	Hispanic	High take-up
19	Parametric	0.046	0.065	0.085	0.003	0.034	0.057
		(0.016)	(0.026)	(0.027)	(0.025)	(0.029)	(0.028)
	Nonparametric	0.096	0.128	0.152	0.010	0.046	0.077
		(0.022)	(0.031)	(0.032)	(0.028)	(0.040)	(0.040)
	Average Y	0.824	0.747	0.741	0.468	0.350	0.343
	Sample size n	22,153	10.252	9,173	22,153	10,252	9,173
	BSC p-value	0.000	0.007	0.000	0.000	0.232	0.288
19-22	Parametric	0.038	0.059	0.074	0.017	0.013	0.011
		(0.007)	(0.010)	(0.011)	(0.009)	(0.010)	(0.011)
	Nonparametric	0.013	0.020	0.019	-0.012	-0.022	-0.015
		(0.011)	(0.016)	(0.016)	(0.015)	(0.021)	(0.021)
	Average Y	0.858	0.781	0.775	0.544	0.407	0.399
	Sample size n	82,077	38,704	34,768	82,077	38,704	34,768
	BSC p-value	0.000	0.000	0.000	0.181	0.000	0.000
23-30	Parametric	0.013	0.015	0.013	0.008	-0.001	-0.000
		(0.005)	(0.008)	(0.008)	(0.009)	(0.010)	(0.010)
	Nonparametric	-0.008	-0.007	-0.014	0.005	-0.007	-0.009
		(0.009)	(0.011)	(0.011)	(0.011)	(0.016)	(0.015)
	Average Y	0.862	0.0767	0.761	0.613	0.443	0.435
	Sample size n	133,576	61,210	54,110	133,576	61,210	54,110
	BSC p-value	0.000	0.000	0.996	0.000	0.000	0.000

The nonparametric results are equally interesting. Here we find the impact of DACA on high school completion to be larger than that found in Kuka *et al.* (2020). For 19 year olds, the nonparametric model suggests that the increase was 9.6 percentage points for all immigrants, 12.8 percentage points for Hispanic immigrants and 15.2 percentage points for immigrants from high take-up countries. That being said, these point estimates are not statistically different from their corresponding parametric counterparts.

While the point estimates for 19 year olds were larger for the nonparametric model, those same results for 19-22 and 23-30 years olds are often smaller in the nonparametric model. The parametric model appears to underestimate the impact of DACA for 19 year olds, but exaggerates it for older individuals.

A similar patter occurs for the impact of DACA on some college. The fourth through sixth columns of Table 4.2 show higher impacts of DACA in the nonparametric setting (except for the high take-up sample) for 19 and 19-22 year olds and lower impacts of DACA for 23-30 year olds. However, the majority of point estimates here are insignificant. While the nonparametric model removes restrictive assumptions, it is unable to conclude that DACA has a significant impact on college enrollment.

In summary, our average effects were able to confirm the parametric result of increased schooling in individuals aged 14-18. This result is important as we can have more faith in the impact of such policies on high school aged students. As for completion of high school, the impact was stronger than previously thought for individuals aged 19-22. This result suggests the program is more effective than previously thought. However, high school completion is defined as earning a GED or a diploma and we are unable to disentangle the two.[6] At the same time, our nonparametric model was able to accurately uncover the negative impact of DACA on school attendance of college aged immigrants, which the parametric model could not (positive and insignificant).

4.3 Heterogeneous Treatment Effects

While our replication results above are interesting, our methods are far more general. We are able to look at TT_x and in this section will delve more into heterogeneous treatment effects. As the number of possible groupings is seemingly endless, we will focus our attention on the first result (the effect of DACA on school attendance for 14-18 year olds). In what follows, we will look at the average effect for male vs female non-citizen immigrants, the average effect by race, the average effect by age and the average effect by age immigrated. Additional results are easily tabulated via minor changes to our code.

We present our main findings in Figures 4.1-4.4. In each figure we plot the average treatment effect on the treated for each group TT_x along with confidence bounds. For

[6] Pope (2016) finds suggestive evidence that DACA pushed individuals to obtain their GED certificate.

simplicity, these symmetric bounds are calculated via the average effect plus or minus two times the (bootstrapped) standard error of the average treatment effect estimate (alternative methods produced similar results). We used 999 bootstrapped resamples for the calculation of each standard error.

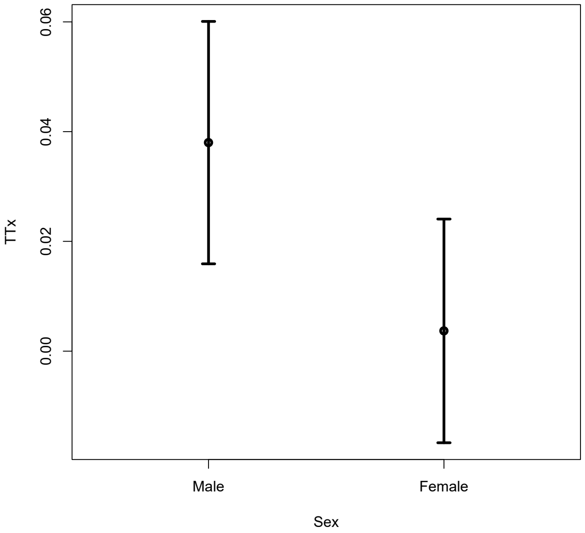

Figure 4.1: Effect of DACA on school attendance (age 14-18) by sex

The results via sex are striking. Figure 4.1 show a positive and significant effect for male non-citizen immigrants ($TT_{male} = 0.038$, $se\left(TT_{male}\right) = 0.011$), but an insignificant effect for female non-citizen immigrants ($TT_{female} = 0.004$, $se\left(TT_{female}\right) = 0.010$). The average treatment effect for the full population ($TT = 0.022$, $se\left(TT\right) = 0.008$) is roughly half the effect between the two groups (we note a modest increase in the standard errors as the respective group sample sizes are smaller than that of the full sample). Although the $2\hat{\sigma}-$confidence intervals overlap a bit, there does appear to be a much larger impact of DACA on male relative to female non-citizen immigrants.

The results by race are also very interesting. Figure 4.2 shows positive and significant average treatment effects for Hispanic ($TT_{Hispanic} = 0.030$, $se\left(TT_{Hispanic}\right) = 0.022$), White ($TT_{White} = 0.009$, $se\left(TT_{White}\right) = 0.004$) and Black ($TT_{Black} = 0.037$, $se\left(TT_{Black}\right) = 0.007$) non-citizen immigrants. Those for Asian and Other non-citizen immigrants are insignificant. The largest percentage of treated observations are Hispanic ($37,659$), but they also have the most variation in their estimates. It seems likely that this result should be examined deeper. One lesson learned from the previous figure is that perhaps it should be further broken down by sex or perhaps by country of origin.

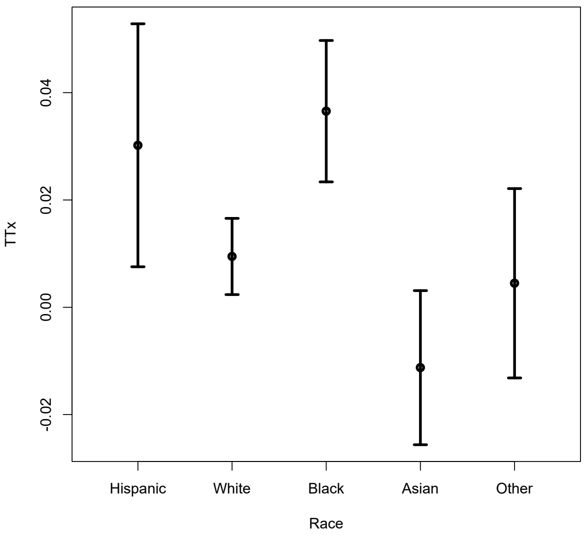

Figure 4.2: Effect of DACA on school attendance (age 14-18) by race

The results broken down by age are intuitive. Figure 4.3 shows that the effect increases with age. For 14 and 15 year olds, the effects are insignificant ($TT_{14} = -0.012$, $se\,(TT_{14}) = 0.008$ and $TT_{15} = -0.009$, $se\,(TT_{15}) = 0.006$). For 16, 17 and 18 year olds, the effects are significant and monotonically increasing with age ($TT_{15} = 0.019$, $se\,(TT_{16}) = 0.006$, $TT_{17} = 0.033$, $se\,(TT_{17}) = 0.009$) and $TT_{18} = 0.082$, $se\,(TT_{18}) = 0.019$)). For most US states, schooling is compulsory until the age of 16, 17 or 18. This likely explains the insignificant impact for ages 14 and 15 and the increasing average effect by age afterward (for 16-18). We note here that the effect for 18 year olds is the largest we have seen in this paper. Most students graduate high school at age 18 and perhaps DACA encourages them to complete the final step.

The final set of results we consider here are looking at the average effects for the age at which the individual immigrated. Recall that eligibility required that they arrived on US soil by the age of 10 (in 2007). We therefore looked at the average effect with respect to the age immigrated (0 to 10). These results can be found in Figure 4.4. With the exception of age 4 (which we have no explanation for), all the results are insignificant until age 8. The average effects for non-citizens who immigrated at the age of 8, 9 or 10 are not statistically or economically different from one another, but they are each significantly different from zero. Without further knowledge it is unclear why this may be true, but we conjecture that it may have to do with English language skills (it is common knowledge that it is more difficult to learn new languages with age).

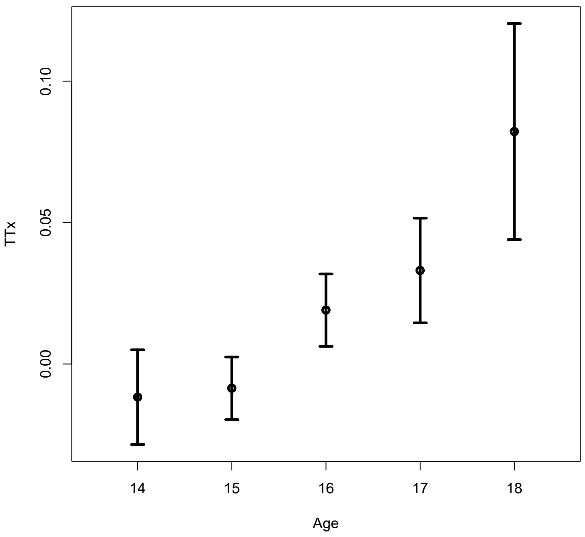

Figure 4.3: Effect of DACA on school attendance (age 14-18) by age

While these results shed new light, they also create more questions. We argue that looking at average treatment effects on the treated for the entire population likely masks many important results. While this is not a controversial statement, we believe that our methodology is desirable for this type of analysis.

4.4 Caveats and Directions for Future Research

Before this application is to be taken seriously for policy analysis, a few caveats remain. First, we want to emphasize that the bias stability condition was rejected for several of our cases. We should look into whether we can find additional confounders or different samples for which this is not the case. Second, the authors of the previous study treat 2012 as the year in which treatment occurred. This may or may not be true. While DACA was signed in 2012 and was put in place in 2012, the first set of recipients did not become aware until later in the year (the Department of Homeland Security first began accepting applications in August of 2012). We are unable to determine when the individual filled out their Census form (this is unknown even to the Census as they have private firms collect and transmit this data). It is likely the case that this adds additional noise, but could also impact the point estimates.[7] Third, we treated all previous years and post years equally. There is no

[7] It is also true that military veterans were eligible for DACA. This omission, however, likely played no role for younger individuals.

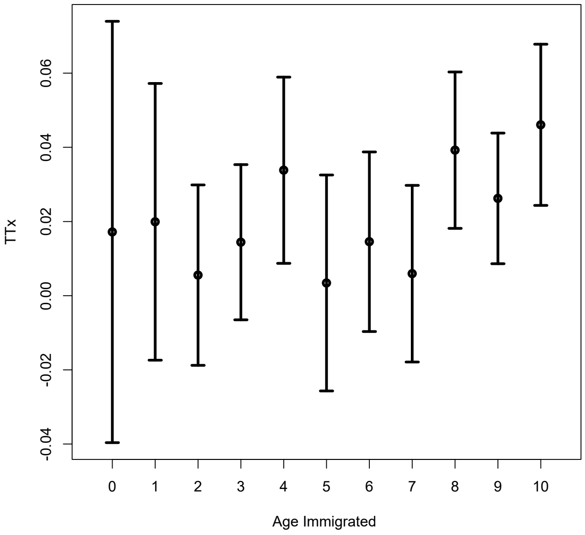

Figure 4.4: Effect of DACA on school attendance (age 14-18) by age immigrated

reason that need be the case.

So where do we go from here? We believe that it would make sense to examine the sample more carefully. We should search for additional confounders so that we fail to reject the bias stability condition. We should also look at effects for more homogeneous groups. For example, we may want to look at Hispanic males who immigrated to the US after the age of 7. Given the root-n consistency of our TT_x estimator, we should get relatively trustworthy results for these types of breakdowns.

Acknowledgements

The authors wish to thank Brant Callaway, Traviss Cassidy, Bill Greene, Sukjin Han, Guido Imbens, Andrew Johnston, James MacKinnon, Chris Parmeter, Jeff Racine, Pedro Sant'Anna, Gonzalo Vazquez-Bare, Le Wang and Jun Zhao for useful comments and suggestions. The authors would also like to thank participants of the following (1) seminars: McMaster University, University Carlos III of Madrid, University of Alabama, University of California, Riverside, and University of Passau and (2) conferences: North American Summer Meetings of the Econometric Society (Montreal), Asian Meeting of the Econometric Society (Miri), International Association of Applied Econometrics Conference (Rotterdam), International Conference on Computational and Financial Econometrics (London), Meeting of the Economics of Education Association (Porto), New York Camp Econometrics (Lake

Placid), and West Indies Economic Conference (Mona). All R code is available from the authors' upon request and can be used with the understanding that those who use the code will cite this article in any publication which uses the code.

A Proofs

This appendix includes the main proofs of the paper. It begins with the asymptotics for the test statistics and ends with giving the influence functions for our treatment effect estimators.

A.1 Asymptotics of the Test Statistics

Here we give all the main steps of the technical proof. For calculation of the bias and variance, we partly follow Vilar-Fernández and González-Manteiga (2004) and Dette and Neumeyer (2001). They consider the problem of nonparametric comparisons of regression curves, say $H_0 : \ m_1 = m_2 = \cdots = m_K$ for $m_k(x) = E[Y|X = x]$, $k = 1, \ldots, K$ which correspond to different populations. The former considered this for autocorrelated data, while the latter considered this for independent data, but with different statistics. We decompose

$$\mathcal{T}_1 = \sum_{d,t=0}^{1} \Gamma_{dt} + 2 \sum_{mix(dt,ks)} (-1)^{d+k+t+s} \Gamma_{dt,ks} + o_P\left(\frac{1}{n_{11}\sqrt{h}}\right), \tag{A.1}$$

where for $W_{dt}(x_{it}) := \frac{1}{n_{dt}h} W\{(x_{it} - x)/h\}/f_{dt}(x)$

$$\begin{aligned}
\Gamma_{dt} &= \sum_{D_i=d:i=1}^{n_{dt}} \sum_{D_j=d:j=1}^{n_{dt}} \int W_{dt}(x_{it}) W_{dt}(x_{jt}) dF_{11}(x) \ u_{it} u_{jt} \\
\Gamma_{dt,ks} &= \sum_{D_i=d:i=1}^{n_{dt}} \sum_{D_j=k:j=1}^{n_{ks}} \int W_{dt}(x_{it}) W_{ks}(x_{js}) dF_{11}(x) \ u_{it} u_{js},
\end{aligned}$$

where we first interchanged the sums, and then approximated the average $\frac{1}{n_{11}} \sum_{D_i=1:i=1}^{n_{11}}$ by $\int dF_{11}(x)$. Due to the independence of the u_{it}, an assumption we relaxed for balanced panels (for repeated cross sections it is less problematic), the expectation of $\Gamma_{dt,ks}$ is zero, and so is the expectation of all mixed terms of Γ_{dt}. Taking the expectation of the remaining $\sum_{D_i=d:i=1}^{n_{dt}} \int W_{dt}^2(x_{it}) dF_{11}(x) \ u_{it}^2$ leads us (after some calculations that are standard in kernel regression) to the stated bias.

To obtain the variance, we need to consider the expectation of the square (A.1), but suppressing in Γ_{dt} the $\sum_{D_i=d:i=1}^{n_{dt}} \int W_{dt}^2(x_{it})dF_{11}(x)\ u_{it}^2$. That is, we consider the $\Gamma_{dt,ks}$ and

$$\Gamma'_{dt} = 2 \sum_{D_i=d:i=1}^{n_{dt}} \sum_{D_j=d:j<i} \int W_{dt}(x_{it})W_{dt}(x_{jt})dF_{11}(x)\ u_{it}u_{jt}.$$

The independence of these terms follows from the independence of the u_{it} (as we consider cohorts of independent observations), so that we can calculate the variance of each term separately. From the related literature on nonparametric testing, it is well known that the variance of the Γ'_{dt} gives the first part of $\mathcal{V}/(n_{11}^2 h)$ with the sum over the four groups. The errors u_{it} belonging to group (dt) are independent not only within this group, but also from those of any other group (ks); all additive terms in $\Gamma_{dt,ks}$ are independent from each other. Taking expectation, the second part of $\mathcal{V}/(n_{11}^2 h)$ containing all mixtures $mix(dt, ks)$ is

$$\begin{aligned}
&E[\Gamma_{dt,ks}^2] \\
&= \frac{1}{n_{dt}^2 n_{ks}^2 h^4} E\left[\sum_{D_i=d:i=1}^{n_{dt}} \sum_{D_j=k:j=1}^{n_{ks}} \left\{ \int W_{dt}(x_{it})W_{ks}(x_{js})dF_{11}(x) \right\}^2 u_{it}^2 u_{js}^2 \right] \\
&= \frac{1}{n_{dt}^2 n_{ks}^2 h^2} E\Bigg[\sum_{D_i=d:i=1}^{n_{dt}} \sum_{D_j=k:j=1}^{n_{ks}} \left(K * K\left(\frac{x_{it}-x_{js}}{h} \right) \right)^2 \times \\
&\qquad\qquad \frac{f_{11}(x_{it})f_{11}(x_{js})u_{it}^2 u_{js}^2}{f_{dt}^2(x_{it})f_{ks}^2(x_{js})} \Bigg] \\
&= \frac{1}{n_{dt} n_{ks} h^2} E\Bigg[\left(W * W\left(\frac{x_{it}-x_{js}}{h} \right) \right)^2 \times \\
&\qquad\qquad \frac{f_{11}(x_{it})f_{11}(x_{js})\sigma_{dt}^2(x_{it})\sigma_{ks}^2(x_{js})}{f_{dt}^2(x_{it})f_{ks}^2(x_{js})} \Bigg],
\end{aligned}$$

which gives us the second part of the variance. The central limit theorem follows directly from Vilar-Fernández and González-Manteiga (2004) or Dette and Neumeyer (2001).

A.2 Influence Functions

The influence functions for TT_a (for $p_{dt}(x) = Pr(D=d, T=t|x)$) can be written as

$$\begin{aligned}
\varphi_a(X) = &\tfrac{DT}{E[DT]} \left[m_{11}(X) - m_{10}(X) - \{m_{01}(X) - m_{00}(X)\} - TT_a \right] \\
&+ \tfrac{DT}{E[DT]} \{Y - m_{11}(X)\} - \tfrac{D(1-T)}{E[DT]} \tfrac{p_{11}(X)}{p_{10}(X)} \{Y - m_{10}(X)\} \\
&- \tfrac{(1-D)T}{E[DT]} \tfrac{p_{11}(X)}{p_{01}(X)} \{Y - m_{01}(X)\} \\
&+ \tfrac{(1-D)(1-T)}{E[DT]} \tfrac{p_{11}(X)}{p_{00}(X)} \{Y - m_{00}(X)\} + R_{h,n_{11}}(X),
\end{aligned}$$

where $R_{h,n_{11}}(X)$ is a remainder term due to the nonparametric estimates $\widehat{m}_{dt}(\cdot)$. Here we have used that

$$\begin{aligned} E[D(1-T)p_{11}(X)p_{10}^{-1}(X)] &= E[(1-D)Tp_{11}(X)p_{01}^{-1}(X)] \\ &= E[(1-D)(1-T)p_{11}(X)p_{00}^{-1}(X)] = E[DT] \ . \end{aligned}$$

Noting that $n_{11} = n\ E[DT]$, we immediately get the seemingly simpler (compared to the one given in Proposition 2.3.2) variance representation

$$\begin{aligned} Var(\widehat{TT}_a) = E\Bigg[\Big\{&\{m_{11}(X) - m_{10}(X) - m_{01}(X) + m_{00}(X) - TT_a\}^2 \\ +\sigma_{11}^2(X) + \tfrac{p_{11}(X)}{p_{10}(X)}\sigma_{10}^2(X) &+ \tfrac{p_{11}(X)}{p_{01}(X)}\sigma_{01}^2(X) + \tfrac{p_{11}(X)}{p_{00}(X)}\sigma_{00}^2(X)\Big\}\tfrac{p_{11}(X)}{E[DT]}\Bigg]\tfrac{1}{n_{11}}. \end{aligned}$$

It is not very hard to see how this changes when we consider TT_b. In that case it is helpful to define the propensity score $p(x) = Pr(D = 1|x)$. Then the influence function for TT_b can be written as

$$\begin{aligned} \varphi_b(X) = \tfrac{D}{E[D]}&\left[m_{11}(X) - m_{10}(X) - \{m_{01}(X) - m_{00}(X)\} - TT_b\right] \\ &+\tfrac{DT}{E[DT]}\{Y - m_{11}(X)\} - \tfrac{D(1-T)}{E[D(1-T)]}\{Y - m_{10}(X)\} \\ &-\tfrac{(1-D)T}{E[DT]}\tfrac{p(X)}{1-p(X)}\{Y - m_{01}(X)\} \\ &+\tfrac{(1-D)(1-T)}{E[D(1-T)]}\tfrac{p(X)}{1-p(X)}\{Y - m_{00}(X)\} + R_{h,n_1}(X) \qquad . \end{aligned}$$

Consequently, $n_1 = n_{11} + n_{10}$ replaces n_{11} and the variance expression becomes

$$\begin{aligned} Var(\widehat{TT}_b) = E\Bigg[&\tfrac{p(X)}{E^2[D]}\{m_{11}(X) - m_{10}(X) - m_{01}(X) + m_{00}(X) - TT_b\}^2 \\ +\tfrac{p_{11}(X)}{E^2[DT]}\sigma_{11}^2(X) &+ \tfrac{p_{10}(X)}{E^2[D(1-T)]}\sigma_{10}^2(X) + \tfrac{p_{01}(X)}{E^2[DT]}\tfrac{p^2(X)}{\{1-p(X)\}^2}\sigma_{01}^2(X) + \\ &\tfrac{p_{00}(X)}{E^2[D(1-T)]}\tfrac{p^2(X)}{\{1-p(X)\}^2}\sigma_{00}^2(X)\Bigg]\tfrac{1}{n}, \end{aligned}$$

where $n = n_{11} + n_{10} + n_{01} + n_{00}$. As $n_1 = n\ E[D]$, we see how the convergence rate of the variance changes from n_{11}^{-1} to $(n_{11} + n_{10})^{-1}$. Another difference is that the first term of the variance is more affected by changing from $\widehat{TT}_a$ to $\widehat{TT}_b$ than the other four terms. The reason is that we use essentially the same information for the prior steps, but the final average from which results the first term of the variance(s) is in case $\widehat{TT}_b$ taken over all members of the treatment group, but for $\widehat{TT}_a$ only over the treated observed in $t = 1$. This difference can be seen more easily when also for the cohorts we suppose $D \perp T|X$. In that case the first variance term of $\widehat{TT}_a$ differs from that of $Var(\widehat{TT}_b)$ by the factor $1/P(T = 1)$. If we have $n_{11} = n_{10}$, it means that this term is twice as big for $\widehat{TT}_a$; exactly what intuition would tell us.

It should be clear that the expressions simplify if $D \perp T|X$ which is unfortunately not guaranteed by the standard assumption $D \perp T$ if X is allowed to vary over time. If X does not change over time, then $X \perp T$ and $D \perp T|X$ follows from $D \perp T$. To see how much this simplifies for instance $Var(\widetilde{TT}_b)$, note that $p_{1t}(x) = p(x)\ Pr(T = t|D = 1, x)$ and $p_{0t}(x) = \{1 - p(x)\}Pr(T = t|D = 0, x)$, $E[DT] = E[D] \cdot E[T]$, etc.

Let us consider the special case of the simplified variance for balanced panels with all covariate values fixed to the observations in $t = 0$, cf. Corollary 2.3.2 also for notation. It is not hard to see that it can be written along the above expressions as

$$\frac{1}{n^1} E\Bigg[\frac{p(X)}{E[D]}\{m_{11}(X) - m_{10}(X) - m_{01}(X) + m_{00}(X) - \widetilde{TT}\}^2 \\ + \frac{p(X)}{E[D]}\sigma_1^2(X) + \frac{p^2(X)}{E[D]\{1-p(X)\}}\sigma_0^2(X)\Bigg],$$

which again coincides with the efficiency bounds of Sant'Anna and Zhao (2020).

B Additional Discussion and Final Thoughts

This appendix discusses the alternative of a parametric estimation based on fully saturated parametric models and how they relate to a nonparametric approach. We also append concluding remarks, including a discussion about the post-selection inference problem.

B.1 Nonparametric versus Parametric Saturated Models

In the economics literature, there does not appear to be a consistent definition of a saturated model. It is common to refer to it in order to justify the use of a parametric model, sometimes without specifying which definition is applied. A popular definition is that a model is saturated when the number of parameters is equal to the number of data points. Another popular alternative is to say a saturated model perfectly reproduces all of the variances, covariances and means of the observed variables. For the regression context, you may think of an interpolation where the curve or surface passes through each point, i.e., an exact fitting model. In a (generalized) linear regression model, 'parameters' refer to 'coefficients'. If the covariates can only take a limited number of values, thinking e.g., only of discrete variables with finite support, such a model can easily become overparametrized, and a re-definition is needed. We would then call any model saturated if it reproduces the same fit as the overparametrized one.

In the regression context this is easy to illustrate and understand: imagine a case in which you have a relatively small number of discrete covariates X that split the sample into few groups (cells) of identical information regarding X.[1] For regression, compute the respective response means of Y within each cell and weight them (or their differences when looking at deviations from the overall mean) with the proportion of each particular cell in the sample. This means transforming all covariates into complete sets of dummy variables, and taking all possible interactions of the highest-order between all dummies. Equivalently, instead of taking all the highest-order interaction terms, you take a set of the same number of terms out of the full set of dummies and interactions but fulfilling the full rank condition. It is not hard to see that you can calculate the coefficients of one model out the coefficients of such an alternative model. Clearly, this is only feasible if (a) all covariates are discrete, (b) having a finite support, and (c) each cell contains a reasonably large number of observations. This is actually equivalent to the use of nonparametric regression with $\lambda = 1$ (or, if using W for all covariates, when taking bounded kernels with h close to zero). In case you have at least one continuous covariate, this strategy cannot provide you a saturated model. However, even when all X are discrete with a finite support, in practice you may find several cells that are either empty or contain only a few observations. This problem increases dramatically with both, the number of covariates and/or their support(s). Even if the sample is sufficiently large such that this is a minor problem, you then reach computational limitations due to the size of the projection matrix. This was clearly an issue in the (parametric replication portion of our) DACA application.

Some people switch to what is sometimes also called a 'reasonably' saturated model, which is even less clearly defined. In practice, its choice is either subjective or random; in either case it risks approximation bias which to some extent corresponds to the smoothing bias in nonparametrics. The advantage in nonparametrics is threefold then: (1) this choice corresponds to the bandwidth choice and can easily be done in a data-driven way, (2) we understand the risk and know the smoothing bias so that we can deal with it, and (3) computationally it is essentially always feasible as we do not need to split the categorical variables into dummies.

The problem of no or few observations in a cell is not just a question of overparametrization for saturated models, it is thereby related to the question of optimal estimation (or prediction) in the sense of minimal mean-squared-errors. This is exactly how the nonparametric approach deals with it: looking for the optimal balance between approximation bias and overparametrization. Consequently, while asymptotically taking a saturated model (if possible, i.e., only discrete covariates with finite support are included) is equivalent to nonparametric regression, in finite samples, doing the latter will result in a smaller mean squared error which is the main objective we should have in this context.[2]

[1] For instance, if all information you have is sex assigned at birth (bi-variate) and one of four educational levels, the sample splits at most in eight cells.

[2] Remember that your estimate is just a realized random variable; unbiasedness only says that the average of a many those estimates converges to the true value, but the mean square error approach aims on

Commonly raised concerns against nonparametrics in this context are the slower rate of convergence and the curse of dimensionality. We have contested this criticism by emphasizing that both issues only concern (i) the conditional treatment effects if heterogeneity is explored over a continuous variable, i.e., if one conditions on a continuous x, and (ii) more generally, if one included more than three continuous covariates without applying bias reducing methods like higher-order polynomials. Without denying that criticism, nor weakening our replies, the above outlined considerations can give us further insight to these issues.

Regarding the convergence rate: unless your parametric model is correctly specified, a 'reasonably' saturated model requires you to increase, for increasing sample size, the cells generated by a continuous covariate (or by a discrete covariate with infinite support). The optimal rate at which their number increases corresponds to (the inverse of) the bandwidth rate such that the convergence rate of the estimator in a 'reasonably' saturated model equals the one of nonparametric estimation (as said, in the optimal case, else it converges slower than the nonparametric one). Even the argument that a parametric approximation would do equally well if one were only interested in the average is an illusion: suppose x is a univariate continuous covariate, and we are indeed only interested in the population or sample average of $\partial E[Y|x = x_i]/\partial x$. Thinking of $E[Y|x = x_i] = \beta_i x_i$, then the β of a linear model is the average of the β_i only if the latter are uncorrelated with x_i (which is a strong assumption), whereas we do not need anything similar for their nonparametric counterparts $\partial E[Y|x = x_i]/\partial x$.

Regarding the curse of dimensionality: for simplicity, suppose all potential covariates were discrete with each having a support of cardinality K. Then a saturated model with k covariates has K^k cells. For both parametric and nonparametric models, increasing k (or K) can become a problem. While it is true that in theory K and k are fixed while the sample size increases, in practice you face even more problems with parametric estimation (unless you significantly simplify your model), risking serious approximation biases whose size and direction you don't know. Note that for fixed k, K, none of the methods suffer asymptotically from decreasing rates. Unfortunately, this is only the case for asymptotic theory.

B.2 Concluding Remarks

We suggest a complete framework for causal analysis (with covariates) via model-free DiD estimation and testing. We show how to automatically select confounders and the scale of the outcome variable, estimate TTs, choose bandwidths and construct standard errors and confidence intervals. We also present model-free testing for significance and heterogeneity

minimizing the distance of your estimate to the true value in probability. Moreover, in the parametric world, 'unbiasedness' only means to have such convergence towards the projection of the real world on your model which can be biased or even meaningless; we know nothing about the distance to the 'truth'.

of treatment effects. Importantly, we also provide a bootstrap test for credibility of the identification assumptions. These results can be used in many common situations and result in robust analysis. We provide asymptotic theory for both cohorts and panels, for time-varying and for time constant covariates. The finite sample performance has been verified by simulation studies under rather complex designs.

We apply our techniques to study the impact of DACA on human capital decisions. We compare our results to Kuka *et al.* (2020). If their models were correctly specified, we would expect that we get similar results. As in their paper, we find a positive (but larger) impact of DACA on high school attendance and high school completion, but we also find that they were unable to identify the negative impact of DACA on school enrollment of college aged individuals. Our findings are closer to what intuition suggests. We also examined heterogeneity of our treatment effects. These results uncovered several interesting findings that were masked by looking at average effects. For example, we found that the effects were positive and significant for males, but insignificantly different from zero for females.

We proposed a selection of scale and covariates along (2.8), (2.9) and (2.10) in the spirit of the non-testable identifying Assumption I. If we want to address the post-selection inference problem, we suggested an equivalent to the sample splitting approach (Kuchibhotla *et al.*, 2022). Alternatively, to account for all variation of the entire statistical analysis, we could apply an outer bootstrap loop that runs over all steps of the analysis until the final estimate. In practice this would be extremely costly and may also give unreasonably large standard errors. In our context (i.e., given the objective of the first steps), it is questionable if the practitioner should be interested in such variance.

References

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

Abadie, A. and G. W. Imbens. (2008). “On the Failure of the Bootstrap for Matching Estimators”. *Econometrica*. 76(6): 1537–1557.

Amuedo-Dorantes, C. and F. Antman. (2017). “Schooling and Labor Market Effects of Temporary Authorization: Evidence from DACA”. *Journal of Population Economics*. 30: 339–373.

Ang, D. (2019). “Do 40-Year-Old Facts Still Matter? Long-Run Effects of Federal Oversight under the Voting Rights Act”. *American Economic Journal: Applied Economics*. 11(3): 1–53.

Athey, S. and G. W. Imbens. (2006). “Identification and Inference in Nonlinear Difference-in-Differences Models”. *Econometrica*. 74(2): 431–497.

Barbeito, I., R. Cao, and S. Sperlich. (2023). “Bandwidth Selection for Statistical Matching and Prediction”. *TEST*. 32: 418–446.

Benini, G. and S. Sperlich. (2022). “Modeling Heterogeneous Treatment Effects in the Presence of Endogeneity”. *Econometric Reviews*. 41(3): 359–372.

Bodory, H., L. Camponovo, M. Huber, and M. Lechner. (2020). “The Finite Sample Performance of Inference Methods for Propensity Score Matching and Weighting Estimators”. *Journal of Business & Economic Statistics*. 38(1): 183–200.

Cao-Abad, R. and W. González-Manteiga. (1993). “Bootstrap Methods in Regression Smoothing”. *Journal of Nonparametric Statistics*. 2(4): 379–388.

Card, D. and A. B. Krueger. (1992). “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania”. *American Economic Review*. 84(4): 772–793.

Chaisemartin, C. de and X. D’Haultfoeuille. (2020). “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects”. *American Economic Review*. 110(9): 2964–2996.

Chan, K. C. G., S. C. P. Yam, and Z. Zhang. (2016). “Globally Efficient Non-parametric Inference of Average Treatment Effects by Empirical Balancing Calibration Weighting”. *Journal of the Royal Statistical Society, Series B*. 78(3): 673–700.

Chu, C.-Y., D. J. Henderson, and C. F. Parmeter. (2015). “Plug-in Bandwidth Selection for Kernel Density Estimation with Discrete Data”. *Econometrics*. 3(2): 199–214.

Cornillon, P.-A., N. Hengartner, and E. Matzner-Løber. (2017). “Statistics and Causal Inference”. *Journal of Statistical Software*. 77(9).

Davidson, R. and E. Flachaire. (2008). “The wild bootstrap, tamed at last”. *Journal of Econometrics*. 146(1): 162–169.

Dette, H. and N. Neumeyer. (2001). “Nonparametric Analysis of Covariance”. *Annals of Statistics*. 29(5): 1361–1400.

Drake, C. (1993). “Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect”. *Biometrics*. 49(4): 1231–1236.

Faraway, J. J. (1990). “Bootstrap Selection of Bandwidth and Confidence Bands for Nonparametric Regression”. *Journal of Statistical Computation and Simulation*. 37(1-2): 37–44.

Frölich, M. (2005). “Matching Estimators and Optimal Bandwidth Choice”. *Statistics and Computing*. 156: 197–215.

Frölich, M. and S. Sperlich. (2019). *Impact Evaluation: Treatment Effects and Causal Analysis*. Cambridge University Press.

Galdo, J. C., J. Smith, and D. Black. (2008). “Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data”. *Annales d’Économie et de Statistique*. 91-92: 189–216.

Häggström, J. and X. Luna. (2014). “Targeted Smoothing Parameter Selection for Estimating Average Causal Effects”. *Computational Statistics*. 29: 1727–1748.

Hall, P. and J. Horowitz. (2013). “A Simple Bootstrap Method for Constructing Nonparametric Confidence Bands for Functions”. *Annals of Statistics*. 41(1): 1892–1921.

Hall, P., Q. Li, and J. S. Racine. (2007). “Nonparametric Estimation of Regression Functions in the Presence of Irrelevant Regressors”. *Review of Economics and Statistics*. 89(4): 784–789.

Härdle, W. and J. S. Marron. (1991). “Bootstrap Simultaneous Error Bars for Nonparametric Regression”. *Annals of Statistics*. 19(2): 778–796.

Härdle, W. and T. M. Stoker. (1989). “Investigating Smooth Multiple Regression by the Method of Average Derivatives”. *Journal of the American Statistical Association*. 84(408): 986–995.

Hayfield, T. and J. S. Racine. (2008). “Nonparametric Econometrics: The np Package”. *Journal of Statistical Software, Articles*. 27(5): 1–32. ISSN: 1548-7660.

Heckman, J. J., H. Ichimura, and P. E. Todd. (1997). “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme”. *Review of Economic Studies*. 64(4): 605–654.

Henderson, D. J. and C. F. Parmeter. (2015). *Applied Nonparametric Econometrics*. Cambridge University Press.

Holland, P. W. (1986). “Statistics and Causal Inference”. *Journal of the American Statistical Association*. 81(396): 945–960.

Hsin, A. and F. Ortega. (2018). “The Effects of Deferred Action for Childhood Arrivals on the Educational Outcomes of Undocumented Students”. *Demography*. 55: 1487–1506.

Jayachandran, S., A. Lleras-Muney, and K. V. Smith. (2010). “Modern Medicine and the Twentieth Century Decline in Mortality: Evidence on the Impact of Sulfa Drugs”. *American Economic Journal: Applied Economics*. 2(2): 118–46.

Kahn-Lang, A. and K. Lang. (2019). “The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications”. *Journal of Business and Economic Statistics*. 38(3): 613–620.

Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small. (2017). “Non-parametric Methods for Doubly Robust Estimation of Continuous Treatment Effects”. *Journal of the Royal Statistical Society, Series B*. 79(4): 1229–1245.

Köhler, M., A. Schindler, and S. Sperlich. (2014). “A Review and Comparison of Bandwidth Selection Methods for Kernel Regression”. *International Statistical Review*. 82: 243–274.

Kuchibhotla, A. K., J. E. Kolassa, and T. A. Kuffner. (2022). “Post-Selection Inference”. *Annual Review of Statistics and Its Application*. 9: 505–527.

Kuka, E., N. Shenhav, and K. Shih. (2020). “Do Human Capital Decisions Respond to the Returns to Education? Evidence from DACA”. *American Economic Journal: Economic Policy*. 12(1): 293–324.

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

Lee, B. K., J. Lessler, and E. A. Stuart. (2010). “Improving Propensity Score Weighting using Machine Learning”. *Statistics in Medicine*. 29(3): 337–346.

Li, Q., J. Racine, and J. Wooldridge. (2009). “Efficient Estimation of Average Treatment Effects with Mixed Categorical and Continuous Data”. *Journal of Business and Economic Statistics*. 27(2): 206–223.

Mammen, E. (1992). *When Does Bootstrap Work? Asymptotic Results and Simulations*. Springer-Verlag.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral. (2004). “Propensity score estimation with boosted regression for evaluating causal effects in observational studies”. *Psychological methods*. 9(4): 403–425.

McKenzie, D., C. Theoharides, and D. Yang. (2014). “Distortions in the International Migrant Labor Market: Evidence from Filipino Migration and Wage Responses to Destination Country Economic Shocks”. *American Economic Journal: Applied Economics*. 6(2): 49–75.

Meyer, B. D. (1995). “Natural and Quasi-Experiments in Economics”. *Journal of Business & Economic Statistics*. 13(2): 151–161.

Neumann, M. H. and J. Polzehl. (1998). “Simultaneous Bootstrap Confidence Bands in Nonparametric Regression”. *Journal of Nonparametric Statistics*. 9(4): 307–333.

Neumeyer, N. and S. Sperlich. (2006). “Comparison of Separable Components in Different Samples”. *Scandinavian Journal of Statistics*. 33: 477–501.

Ouyang, D., Q. Li, and J. S. Racine. (2009). “Nonparametric Estimation of Regression Functions with Discrete Regressors”. *Econometric Theory*. 25(1): 1–42.

Panhans, M. (2019). “Adverse Selection in ACA Exchange Markets: Evidence from Colorado”. *American Economic Journal: Applied Economics*. 11(2): 1–36.

Parmeter, C., Z. Zheng, and P. McCann. (2009). “Cross-Validated Bandwidths and Significance Testing”. *Advances in Econometrics*. 25: 71–98.

Politis, D. N. (2013). “Model-Free Model-Fitting and Predictive Distributions”. *TEST*. 22: 183–221.

Pope, N. G. (2016). “The Effects of DACAmentation: The Impact of Deferred Action for Childhood Arrivals on Unauthorized Immigrants”. *Journal of Public Economics*. 143: 98–114.

Qin, J. and B. Zhang. (2008). “Empirical-likelihood-based Difference-in-differences Estimators”. *Journal of the Royal Statistical Society, Series B*. 70(2): 329–349.

Racine, J. and Q. Li. (2004). “Nonparametric estimation of regression functions with both categorical and continuous data”. *Journal of Econometrics*. 119(1): 99–130.

Roca-Pardiñas, J. and S. Sperlich. (2010). “Feasible Estimation in Generalized Structured Models”. *Statistics and Computing*. 20: 367–379.

Rolling, C. A. and Y. Yang. (2014). “Model Selection for Treatment Effects”. *Journal of the Royal Statistical Society, Series B*. 76(4): 749–769.

Roth, J. (2022). “Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends”. *AER: Insights*. 4(3): 305–322.

Roth, J. and P. H. Sant’Anna. (2023). “When Is Parallel Trends Sensitive to Functional Form?” *Econometrica*. 91(2): 737–747.

Rubin, D. B. (1977). “Assignment to Treatment Group on the Basis of a Covariate”. *Journal of Educational Statistics*. 2(1): 1–26.

Ruggles, S., K. Genadek, R. Goeken, J. Grover, and M. Sobek. (2018). *Integrated Public Use Microdata Series: Version 7.0 [dataset]*.

Sant’Anna, P. H. and J. Zhao. (2020). “Doubly Robust Difference-in-differences Estimators”. *Journal of Econometrics*. 219(1): 101–122.

Sheather, S. J. and M. C. Jones. (1991). “A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation”. *Journal of the Royal Statistical Society. Series B*. 53(3): 683–690.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

Sperlich, S. (2013). “Comments on: Model-Free Model-Fitting and Predictive Distributions”. *TEST*. 22: 227–233.

Sperlich, S. (2014). “On the Choice of Regularization Parameters in Specification Testing: A Critical Discussion”. *Empirical Economics*. 47: 427–450.

Taylor, J. and R. J. Tibshirani. (2015). “Statistical Learning and Selective Inference”. *Proceedings of the National Academy of Sciences*. 112(25): 7629–7634.

Vilar, J. M. and J. A. Vilar. (2012). “A Bootstrap Test for the Equality of Nonparametric Regression Curves Under Dependence”. *Communications in Statistics - Theory and Methods*. 41(6): 1069–1088.

Vilar-Fernández, J. M. and W. González-Manteiga. (2004). “Nonparametric Comparison of Curves with Dependent Errors”. *Statistics.* 38(2): 81–99.

Xia, Y. (1998). “Bias-corrected Confidence Bands in Nonparametric Regression”. *Journal of the Royal Statistical Society, Series B.* 60(4): 797–811.